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

**Zhen Chen**[1] **Yi Zhang**[2] **Fu Wang**[3] **Xingyu Zhao**[2] **Xiaowei Huang**[1] **Wenjie Ruan**[4*]
[1]University of Liverpool [2]University of Warwick [3]University of Exeter [4]USTC
{cz97, xiaowei.huang}@liverpool.ac.uk
{yi.zhang.16, xingyu.zhao}@warwick.ac.uk
fw377@exeter.ac.uk rwjie@ustc.edu.cn

## Abstract

Adversarial robustness and privacy of deep learning (DL) models are two widely studied topics in AI security. Adversarial training (AT) is an effective approach to improve the robustness of DL models against adversarial attacks. However, while models with AT demonstrate enhanced robustness, they become more susceptible to membership inference attacks (MIAs), thus increasing the risk of privacy leakage. This indicates a negative trade-off between adversarial robustness and privacy in general deep learning models. Visual prompting is a novel model reprogramming (MR) technique used for fine-tuning pre-trained models, achieving good performance in vision tasks, especially when combined with the label mapping technique. However, the performance of label-mapping-based visual prompting (LM-VP) under adversarial attacks and MIAs lacks evaluation. In this work, we regard the MR of LM-VP as a unified entity, referred to as the LM-VP model, and take a step toward jointly evaluating the adversarial robustness and privacy of LM-VP models. Experimental results show that the choice of pre-trained models significantly affects the white-box adversarial robustness of LM-VP, and standard AT even substantially degrades its performance. In contrast, transfer AT-trained LM-VP achieves a good trade-off between transferred adversarial robustness and privacy, a finding that has been consistently validated across various pre-trained models.

## 1 Introduction

Deep learning models have gained great success, yet concerns regarding their security continue to grow, as they are susceptible to various attacks [30, 13, 23, 29, 35]. In addition to the deep learning models, training samples are also the key to this success. Consequently, attacks targeting the relationship between training samples and models have emerged, such as adversarial attacks [23, 14, 43, 37] and membership inference attacks (MIAs) [25, 28, 30, 32]. Adversarial attacks are gradient-based methods that introduce imperceptible perturbations on inputs, generating adversarial examples (AEs) that cause the target models to give incorrect predictions. Adversarial training (AT) [23], proposed by Szegedy *et al.* has been recognized as one of the most effective defenses against such attacks. The basic idea of AT is to incorporate AEs into the training process, resulting in significantly improved performance under adversarial attacks compared to standard training (ST).

MIAs, on the other hand, aim to determine whether a specific sample was part of the model's training data. Various ideas exist for performing MIAs, *e.g.*, shadow model-based attacks [30] and prediction

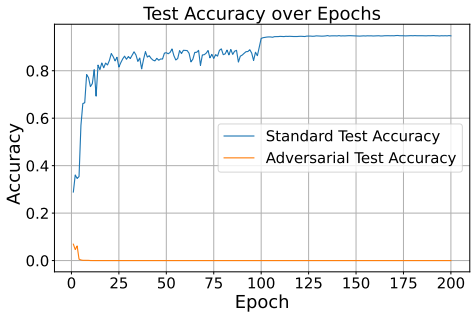
*Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

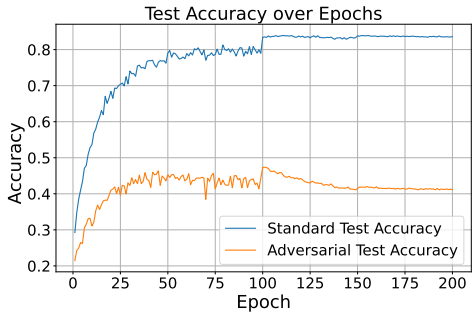
 https://doi.org/10.52202/079017-0217

confidence-based attacks [39, 28, 32]. The success of these attacks largely depends on the model's generalization error during training and testing [31]. Therefore, models with lower generalization errors are inherently more resistant to such attacks.

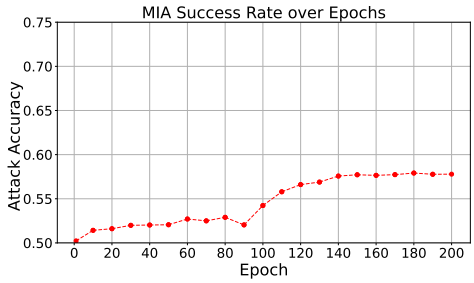
AT-trained models exhibit more severe privacy risks compared to ST: (1) larger generalization error, manifested in both natural and adversarial examples [32]; (2) higher sensitivity on training data compared with ST [32]; (3) robust overfitting [27], where the model's adversarial robustness declines despite the natural accuracy continuing to increase at a certain training stage. These issues result in a negative trade-off between adversarial robustness and privacy. As illustrated in Fig.1, while AT significantly enhances adversarial robustness, it is more susceptible to MIAs, especially after robust overfitting, *i.e.*, between 100-150 epochs in Fig.1d, where the MIA success rate increases significantly. **Notably, the above conclusions are only for general deep-learning models.**



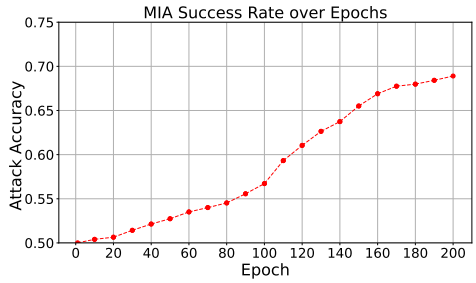

(a) Test Accuracy of Standard Training




(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.

Visual prompting (VP) [3] is a model reprogramming (MR) [10, 7] technique for pre-trained models, used for downstream image classification tasks. Initially, VP involves adding a single, input-agnostic prompt to input images to enhance a pre-trained model's generalization ability. Label mapping (LP) further improves VP's performance by mapping source labels to target labels, denoted as LM-VP, which exhibits strong performance in downstream tasks [3]. In this paper, we regard a general pre-trained model after LM-VP as a new model, and its security remains under-explored, including its susceptibility against AEs and MIAs, and compatibility with AT. We consider two forms of AT: the standard AT for white-box adversarial robustness and transfer AT for black-box transferred adversarial robustness, *i.e.*, generating adversarial examples through another threat model. We empirically demonstrate that the intuitions and relationships between adversarial robustness and privacy observed in general models do not always hold for the LM-VP model.

In summary, our contributions lie in:

- From a novel perspective of considering LM-VP as a distinct model, we conduct the first evaluation of its security, *i.e.*, (transferred) adversarial robustness and privacy;
- Based on the concept of transfer attacks, we implement transferred adversarial training for the LM-VP model to enhance its transferred adversarial robustness;
- We empirically demonstrate that intuitions regarding privacy in general models do not necessarily apply to LM-VP models. Furthermore, we show that standard AT is invalid

https://doi.org/10.52202/079017-0217

for LM-VP, while transfer AT on LM-VP exhibits a superior trade-off between transferred adversarial robustness and privacy across various pre-trained models.

## 2 Related Work and Background

### 2.1 Visual Prompting

The prompt technique [18, 20, 15] was originally employed in NLP tasks. In essence, it modifies the original input text to enhance specific task performance without altering the parameters of the pre-trained model. For instance, prompts indicating sentiment can be added for text classification [24], while prompts indicating the target language can be used for translation tasks [42]. Bahang *et al.*[3] first transfers this idea to computer vision tasks, *i.e.*, VP. Compared to traditional transfer learning methods like fine-tuning [33] and linear probes [1], VP does not modify the parameters of the pre-trained model. Instead, it alters the original image by adding prompts, *i.e.*, introducing additional pixels, enabling task-specific and input-agnostic adjustments. During VP training, only the prompts and output transformation are updated. VP exhibits strong performance across various datasets and significantly reduces the training parameters compared to traditional transfer learning methods. Output transformation, or label mapping (LM), is another key technique contributing to VP's performance. Chen *et al.*[6] proposes the iterative label mapping (ILM) method to replace the random mapping in vanilla VP. Arif *et al.*[2] and Li *et al.*[19] apply a trainable fully-connected layer for label mapping, achieving promising results and improved efficiency. Li *et al.*[19] also explore the VP in training differential private models using the PATE framework [26], while Chen *et al.*[5] investigate VP in test-time adversarial robustness by implementing adversarial prompts during testing.

### 2.2 Adversarial Robustness Evaluation

Adversarial robustness is an important metric for evaluating model robustness, referring to a model's performance under adversarial attacks. Adversarial attacks target image gradients, iteratively introducing imperceptible perturbations to images to generate AEs. These AEs often cause standard-trained deep learning models to misclassify them with high confidence. Commonly used adversarial attacks include PGD [23], FGSM [14], and CW [4] attacks. AT is an approach to improve the adversarial robustness of deep learning models. It can be formulated as a min-max problem, where the inner maximization searches for perturbations that maximize the loss, while the outer minimization optimizes the model, *i.e.*,

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{Z},y)\sim\mathcal{D}} \left[ \max_{\|\boldsymbol{\delta}\|\leq\epsilon} L\left(f_{\boldsymbol{\theta}}(\boldsymbol{X}+\boldsymbol{\delta}), y\right) \right]. \tag{1}$$

Various strategies exist to solve the inner maximization in AT, including PGD-AT [23], TRADES [43], and MART [37], etc. LOAT[40] boosts AT via a Fisher-Rao norm-based regularization, SEAT[36] extends AT to medical segmentation and FAAL[44] ensure both robustness and fairness during AT, Chen *et al.*[8] proposes NRAT to enhance adversarial robustness under noisy labels. While robustness evaluation mainly focuses on discriminative models, Zhang *et al.*[45] introduces a robustness notion of text-to-image (T2I) generative models and proposes the ProTIP framework for evaluation.

### 2.3 Membership Inference Attacks

MIAs refer to determining whether a given data point was part of the training data for a trained model, raising significant data privacy concerns. Shokri *et al.*[30] introduce the first MIA on classification models, *i.e.*, shadow training attack, which involves creating several shadow models that simulate the target model, with these shadow models trained on data records similar to those used for training the target model. An attack model is then trained to recognize the relationship between the members of the shadow models' training data and the shadow models' outputs, which turns out to be a binary classification. This attack model can subsequently infer the membership of the target model's training dataset. There are two findings in this work [30]: (1) the higher the degree of the model's overfitting, the higher the attack's success rate, and (2) the more complex the training dataset, the higher the MIA success rate. Intuitively, increasing the number of shadow models improves attack performance by providing more samples for training the attack model, but this also requires more computational resources. Yeom *et al.*[39] propose threshold-based MIAs, which compare the target model's prediction confidence for the true label against a certain threshold. This method achieves


https://doi.org/10.52202/079017-0217

performance similar to the shadow model training method with significantly reduced computational resource consumption. Song *et al.*[31] further proposes the class-dependent thresholds for a more powerful attack and implements the MIA based on prediction entropy.

## 3 Robustness and Privacy Evaluation of Visual Prompting Models

In this section, we first provide an overall design of LM-VP models, including the prompt designs, label mapping techniques, LM-VP model training, as well as the tricks we adopt. Then, we analyze the white-box adversarial robustness of LM-VP, demonstrating that pre-trained models largely influence its white-box adversarial robustness. We further propose transfer AT to enhance black-box transferred adversarial robustness. Finally, we analyze the intuition between LM-VP models and privacy.

### 3.1 Label Mapping Visual Prompting

LM-VP model aims to keep the parameters of the pre-trained models fixed while performing input transformation and output transformations, *i.e.*, label mapping. Therefore, we divide the LM-VP model into three parts: (1) prompt generation; (2) label mapping; and (3) model training.

#### 3.1.1 Prompt Generation

For the prompt generation, we first rescale the images from the target domain under a certain rescale ratio to a size smaller than that of the source domain. Then, trainable noise $w_1$, *i.e.*, prompt is added around the image to ensure the final image size matches that of the source domain. Therefore, the length (height and width) of each pixel patch $p$ is

$$p = 1/2[(H_1 - H_2) + (W_1 - W_2)], \tag{2}$$

where $H_1$ and $W_1$ represent the height and width of the source domain, $H_2$ and $W_2$ represent the height and width of the rescaled target domain, then the final shape of prompts $P$ is

$$P = C \times [H_1/p + W_1/p - 4] \times p^2, \tag{3}$$

where $C$ is the image channels, $[H_1/p + W_1/p - 4]$ represents the amount of pixel patches in each channel. In vanilla VP [3], they rescale the target images to the size of the source domain and replace the edges with random noise, *i.e.*, prompts. In comparison, although the final prompts $P$ are the same, we preserve the edge information of the target domain images, which is more intuitively reasonable as we retain all the information of target images. Preserving the edge information might help increase the correlation between prompts and images during training. Fig. 2 shows the difference between these two ways of prompt generation.

#### 3.1.2 Label Mapping

For output transformation, vanilla VP uses random mapping, randomly selecting some labels from the source domain to match those of the target domains and discarding the remaining unused labels. However, this random mapping often leads to a performance drop in VP as it may ignore some important information in the unused labels [19]. Therefore, we consider label mapping as a trainable component: using a fully connected layer to train the mapping from source labels to target labels, which is similar to [19] and [2]. Consequently, in LM-VP, we introduce two trainable parameters: the prompt noise parameters $w_1$ and the parameters $w_2$ in the label mapping layer $f_\ell(w_2; \theta_2)$.

#### 3.1.3 LM-VP Model Training

For LM-VP model training, our objective is to modify the prompts by updating the noise parameters $w_1$, and the parameters $w_2$ of FC layers of the label mapping. During the testing phase, the same optimized prompts are applied to all test data, fed into the frozen pre-trained model, and finally output the label mapping results. LM-VP relies on the pre-trained model and label mapping, indicating that it does not require a large target dataset. Therefore, we can train LM-VP models using a subset of the target dataset without a significant performance drop compared with training on the entire target dataset, which significantly improves training efficiency. In contrast, general models tend to underfitting when trained on small subsets, leading to poor generalization performance. Although the LM-VP model does not heavily rely on training data, insufficient data can still affect its generalization.

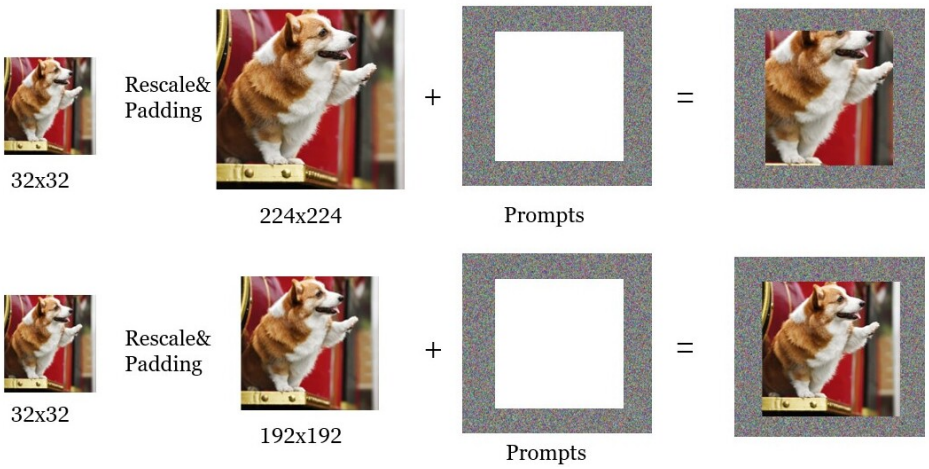



Figure 2: Two ways to add prompts: (1) Top: rescale a target image to the source domain size and replace the edge of the image with prompts; (2) Bottom: rescale a target image to a size smaller than the source domain and add prompts to make it the same size as source domain.

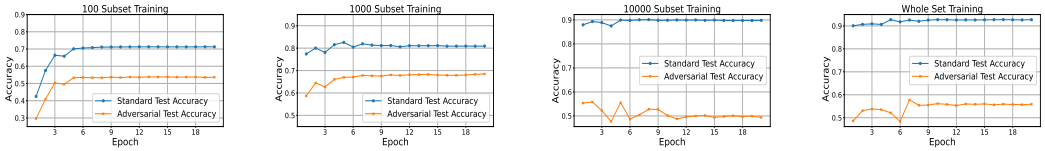



Figure 3: LM-VP model (pre-trained on Swin Transformer[21]) performance on the whole test set using standard training with different numbers of training subsets (random 100, 1000, 10000 subsets and whole training set) on CIFAR-10, transferred adversarial robustness is evaluated on $\ell_\infty$ threat model using ResNet18.

Therefore, we use the SAM [12] version of SGD optimizer to update the weights of $w_1$ to improve the LM-VP model's generalization, *i.e.*,

$$w_{t+1} = w_t - \eta \left( \nabla L \left( w_t + \epsilon_t \right) + \lambda w \right), \tag{4}$$

where $\eta$ is the learning rate, $\epsilon$ is the parameter to maximize the loss function $L$, and $\lambda$ is the weight decay. Given the prediction on source domain $\hat{y}_S$, the final prediction label on target domain $\hat{y}_T$ is obtained by

$$\hat{y}_T = \text{softmax} \left( f_\ell \left( \omega_2; \hat{y}_S \right) \right). \tag{5}$$

Fig. 3 illustrates the impact of data volume on VP performance during training, demonstrating that insufficient training data may not always hurt the performance of the LM-VP model, *e.g.*, the LM-VP model trained with a subset of 1000 samples (second figure) achieves the best transferred adversarial robustness; the LM-VP model trained with a subset of 10000 samples (third figure) achieves a similar standard test accuracy compared with the LM-VP model trained with the whole training set. Additionally, subset training significantly reduces the running time to 0.12x, 0.25x, and 0.45x on subsets of 100, 1,000, and 10,000 samples respectively.

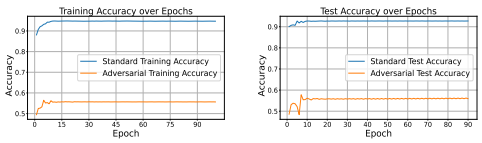



Figure 4: Training and testing performance of LM-VP (pre-trained on Swin Transformer) using standard training on CIFAR-10, transferred adversarial robustness is evaluated on $\ell_\infty$ threat model using ResNet18.

The LM-VP model also exhibits the property of rapid convergence, *i.e.*, it can quickly achieve a near-optimal performance and then remain steady with continued training, for both the natural samples and adversarial samples, as shown in Fig. 4, it takes about 10 epochs to convergence for both


https://doi.org/10.52202/079017-0217

training and testing. Therefore we can set a small number of epochs to improve training efficiency on LM-VP models.

### 3.2 White-box Adversarial Robustness of LM-VP Models

A crucial distinction between the LM-VP model and a general model lies in the presence of a pre-trained model that does not participate in training[3]. Using white-box adversarial robustness metrics to evaluate LM-VP models can be heavily influenced by the choice of pre-trained models, as shown in Table. 1 (Standard Training), there is no clear pattern in their best adversarial robustness across different pre-trained models, thus VP may play a limited role in defending against the white-box adversarial attack. From Fig. 5, for standard-trained LM-VP models, the best (highest) adversarial robustness was only observed in the early stages of training, as training progresses, for all pre-trained models, the adversarial robustness continues to decline until it reaches a stabilized status, thus the best adversarial robustness may largely reflect the pre-trained models' inherent adversarial robustness when transferred to the target dataset.

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** | |
|---|---|---|---|---|
| | Natural$_{te}$ | PGD-20 | 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 |

We also notice that the standard AT is invalid in LM-VP, with really poor results in both natural and adversarial performance, see Table. 1 (Adversarial Training). Since the LM-VP model is trained on the target dataset, but the generation of adversarial examples depends on a fixed pre-trained model from the source dataset domain, the domain shift may lead to unsatisfactory results of AT on the target dataset.

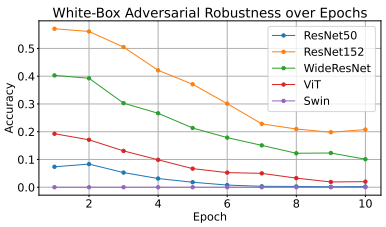



Figure 5: Epoch-wise white-box adversarial robustness of LM-VP using standard training on CIFAR-10.

Regarding how pre-trained models affect downstream adversarial robustness, [38]and [34] provide more insights, *e.g.*, Yamada *et al.*[38] conclude that network architecture is a strong source of robustness in transfer learning. In this sense, different pre-trained models may lead to different boundary relationships between adversarial robustness and privacy, evaluating LM-VP using white-box adversarial attacks may make it difficult to reach consistent conclusions.

### 3.3 LM-VP models with Transferred Adversarial Training

For the evaluation of the transferred adversarial robustness of the LM-VP model, we use another general model as the threat model to produce adversarial examples and train the LM-VP model to defend against them. In this scenario, the intensity of the transfer attack remains constant once the threat model is selected. This consistency holds true regardless of the chosen pre-trained model. This inherent consistency is thus helpful for exploring and establishing a sensible relationship between transferred adversarial robustness and privacy within LM-VP models. Compared to white-box adversarial robustness which is heavily influenced by the pre-trained models, utilizing transferred adversarial robustness serves as a more reliable and insightful evaluation method for LM-VP models.

Within the framework of transfer AT, the LM-VP model which comprises VP, a pre-trained model, and LM, is treated as a unified black-box system, A fixed-parameter threat model, excluded from the training process, is employed to generate AEs $x'$, and then train LM-VP models using adversarial loss, this transfer AT consistently optimizes in the same direction since the attack remains constant.

https://doi.org/10.52202/079017-0217

### 3.4 Relationship between LM-VP models and Training Dataset Privacy

In our work, we use the resistance of a model against MIAs to reflect its privacy. According to [32] and [31], AT-trained general models exhibit larger generalization errors and higher sensitivity to training data, wherein generalization errors and sensitivity are two key factors that influence the success of MIAs on general models, thus indicating a contradiction between adversarial robustness and privacy for general models. However, in the LM-VP model, the architecture of pre-trained models appears to be the more significant factor affecting its (white-box) standard adversarial robustness (Table. 1). As a result, we are not able to establish a reliable boundary relationship between its standard adversarial robustness and privacy, which has prompted us to focus primarily on the relationship between its transferred adversarial robustness and privacy.

In the training of LM-VP models, the parameters (weights and biases) of the pre-trained models are fixed, and the trainable parameters are noise parameters and label mapping parameters. This significantly reduces the influence of the input (both natural and adversarial samples) on the LM-VP model compared to general models, which also enables effective training of LM-VP models with a small subset of data, indicating lower sensitivity of LM-VP models to training data, the results in Fig. 3 also support this statement. Additionally, the generalization ability of LM-VP models mainly relies on the pre-trained models, which have been trained on large-scale datasets and learned rich universal features. With fixed model parameters, the risk of overfitting is reduced.

As shown in Fig. 4, during LM-VP standard training and transfer AT, both training accuracy and test accuracy on natural examples and adversarial examples are very close. Based on this empirical evidence, LM-VP models exhibit minimal generalization error and low sensitivity to training data, which should intuitively enhance their resistance to MIAs and better protect the privacy of training data. However, apart from generalization errors and sensitivity, there might be other factors influencing the LM-VP model's resistance to MIAs, such as the prior knowledge embedded in different pre-trained models. In our experiments, we empirically demonstrate that MIA analyses applied to general models do not always hold for LM-VP models, and transferred adversarial robustness and privacy can be improved simultaneously using transfer AT.

## 4 Experiments

In this section, we conduct comprehensive experiments to evaluate the performance of LM-VP models under transferred adversarial attacks and a threshold-based MIA. Regarding the trade-off among standard accuracy, transferred adversarial robustness, and MIA success rate, we comprehensively compare different pre-trained models in LM-VP models. We conduct main experiments[2] on CIFAR-10 and additional experiments on Tiny-ImageNet to show the good generalization performance of transfer AT. We implement all experiments on a server with an RTX3090 GPU.

### 4.1 Experimental Setup

For adversarial attacks, all experiments follow the standard settings: $\ell_\infty$ threat models for all methods, the perturbation limit $\epsilon = 8/255$, and step size 2/255, we mainly choose ResNet18 as the threat model. For LM-VP models, we use the source models pre-trained on 224x224 ImageNet. For training, we follow the settings in [19], *i.e.*, SGD with SAM technique and a momentum of 0.9 to optimize the LM-VP defense models, the total training epoch is 20. We choose ResNet50 [17], ResNet152 [17], WideResNet-50-2 [41], VIT [9], Swin Transformer [21], ConvNext [22], and EVA [11] models to show the effect of different pre-trained source models in LM-VP model training.

For evaluation, (1) adversarial attacks: we choose PGD-20 and CW-20; (2) MIA: we implement the threshold-based attack on both the natural examples and adversarial examples.

### 4.2 Classification Evaluation of Standard Trained LM-VP Models

To evaluate the performance of LM-VP models using standard training, *i.e.*, its training loss is given by

$$\ell^{ST}\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}\right) = \mathrm{CE}\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_i + P\right), y_i\right). \tag{6}$$

[2]Code is available at `https://github.com/TrustAI/TARP-VP`


https://doi.org/10.52202/079017-0217

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 | Natural$_{tr}$ | Natural$_{te}$ | 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 | Natural$_{tr}$ | Natural$_{te}$ | 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 | - |

$P$ is the prompt. We report its natural accuracy and transferred adversarial robustness in Table. 2, where the "**best performance**" refers to the performance under the epoch of the best (standard or transferred) adversarial robustness.

Based on the parameter capacity of the pre-trained models, we regard them as small, medium, and large models. Specifically, ResNet50, ResNet152, and WRN-50-2 are small models, ViT and Swin Transformer are medium models, and ConvNext and EVA are large models. We report only the natural accuracy during training, while different attacks are employed during testing. The results in Table. 2 show that: (1) There is a clear hierarchy in natural accuracy and transferred adversarial robustness based on the size of the pre-trained models, *i.e.*, small models have a natural accuracy below 90%, medium models around 92%, and large models around 98%; for transferred adversarial robustness, small models are below 45%, medium models range from 50% to 60%, and large models range from 85% to 90%; (2) In LM-VP models, the transferred adversarial robustness is not significantly affected by the size of the threat model, *i.e.*, a larger threat model (WRN-34-10) may not be more challenge compared with ResNet18. For instance, small and medium models sometimes have higher transferred adversarial robustness achieved under WRN-34-10; (3) In LM-VP models, even when more attack steps are used, the transferred adversarial robustness of the test data often remains higher than that of the training data set, although more attack steps being considered more powerful (comparing PGD-10$_{tr}$ and PGD-20).

### 4.3 Classification Evaluation of Transfered AT-based LM-VP Models

We implement the transfer AT of LM-VP models proposed in Section 3.3 and report the performance in Table. 3. Specifically, during training, we use PGD attack with 10 steps to generate AEs:

$$x^0 = x + \sigma, \text{ where } \sigma \sim \mathcal{N}(0, 1), \tag{7}$$

$$x^{t+1} = \Pi_{x+\mathcal{S}}(x^t + \alpha sign(\nabla_x \mathcal{L}(\theta, x^t, y)), \tag{8}$$

$x^0$ is obtained by perturbing $x$ with random noise $\sigma$ sampled from the normal distribution $\mathcal{N}(0, 1)$, $t$ denotes the current attack step, $\alpha$ is the step size, $\Pi$ denotes the projection function, $\mathcal{S} \subseteq \mathbb{R}^d$ denotes the perturbation set of AEs, we train LM-VP models with the following training loss:

$$\ell^{AT}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) = \mathrm{CE}\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_i' + P\right), y_i\right), \tag{9}$$

where $\mathbf{x}_i'$ denotes the AE after PGD. To ensure consistency with the ST in Table. 2, we report the results under the same metrics. Consistent with our findings in Section 4.2, AEs generated by different

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 |

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 |

threat models have minimal impact on both training and testing. Therefore, we exclude the results for WRN-34-10 as a threat model.

The results in Table. 3 indicate that: (1) Transferred AT significantly enhances the transferred adversarial robustness at the cost of reduced natural accuracy. Specifically, the transferred adversarial robustness improvement is around 20%-35% for small models, 20%-25% for medium models, and 3%-6% for large models; (2) Compared to the standard-trained LM-VP models, the transferred adversarial robustness of the training set (PGD-10$_{tr}$) is usually higher than that of the test set (PGD-20) in transferred AT-trained LM-VP models.

### 4.4 Privacy Evaluation of LM-VP Models

In this section, we evaluate the privacy performance of LM-VP models under the threshold-based MIA. Our attack implementation is based on [16], the MIA success rate with a threshold $\eta$ is given by:

$$MIA(\eta) = \frac{1}{2} \times \left( \frac{\sum_{(x,y)\in D_{\text{train}}} \mathbf{1}\left[f_\theta(x)_y \geq \eta\right]}{|D_{\text{train}}|} + \frac{\sum_{(x,y)\in D_{\text{test}}} \mathbf{1}\left[f_\theta(x)_y < \eta\right]}{|D_{\text{test}}|} \right), \tag{10}$$

where the $\eta_{optim}$ is obtained by computing all possible $\eta$ that maximizes the MIA success rate, *i.e.*,

$$\eta_{\text{optim}} = \arg\max_{\eta} MIA(\eta). \tag{11}$$

This attack is mainly based on the model generalization error, *i.e.*, models with higher generalization error are more susceptible to the attack. Conversely, the success rate should approach 50% for a model with little generalization error. However, this principle does not always consistently apply to LM-VP models. From Table. 4, for standard-trained LM-VP models, only the VIT and Swin Transformer can effectively resist the attack with an MIA success rate near 50%. Another observation is the MIA success rate on AEs is lower for the standard-trained LM-VP models.

For transferred-AT trained LM-VP models, the MIA success rate for most cases is around 50%, markedly reducing the risk of training data privacy leakage. This indicates that transferred adversarial


https://doi.org/10.52202/079017-0217

Table 5: Best performance(%) on Tiny-ImageNet with different pre-trained models in Standard-Trained LM-VP models and Transfered AT-Trained LM-VP models under threat model ResNet18.

| **Pre-trained models** | **Standard Training** | | | **Transfer Adversarial Training** | | |
|---|---|---|---|---|---|---|
| | Natural$_{te}$ | PGD-20 | MIA Nat | Natural$_{te}$ | PGD-20 | MIA Nat |
| **ResNet50** | 62.74 | 10.26 | 57.46 | 50.42 | 34.60 | 50.90 |
| **ResNet152** | 65.00 | 20.53 | 62.14 | 57.36 | 38.81 | 50.85 |
| **WRN-50-2** | 70.12 | 16.59 | 53.50 | 50.50 | 30.59 | 50.89 |
| **VIT** | 80.97 | 37.77 | 54.00 | 72.02 | 50.22 | 51.45 |
| **Swin** | 79.93 | 41.81 | 56.95 | 75.08 | 55.81 | 51.35 |
| **ConvNext** | **89.01** | 73.47 | 58.47 | 87.60 | **76.61** | 52.04 |

robustness and privacy can be simultaneously achieved in LM-VP models. One plausible explanation is that during transfer AT, the original training examples are perturbed before feeding into the model, thus these data are not exposed to the trained model, this may help mitigate the MIA issue since LM-VP models also do not suffer from large generalization error (Table. 2 and Table. 3) and increased training data sensitivity (Fig. 3) and transfer AT do not train the original training examples.

### 4.5 Results on Tiny-ImageNet

To demonstrate the efficiency of transferred AT on LM-VP models, we provide the results on Tiny-ImageNet which has a resolution of 64x64 and contains 200 classes, results shown in Table. 5, LM-VP models with transfer AT improve transfer adversarial robustness by 3%-24% and mitigate the MIA success rate by 3%-12% compared to LM-VP models with standard training.

## 5 Conclusion

In this paper, we regard the models trained using the LM-VP technique as a novel model type and analyze its adversarial robustness and privacy. We observe that the choice of pre-trained models significantly influences the white-box adversarial robustness of LM-VP, making it hard to draw consistent conclusions. Therefore, we focus more on its transferred adversarial robustness and its interaction with MIA-based privacy. To address both concerns, we propose the transfer AT method for LM-VP models to enhance performance on both fronts. Experiments across various pre-trained models demonstrate that: *(i) both standard-trained and transfer AT-trained LM-VP models show a positive correlation between transferred adversarial robustness and pre-trained model size*, and *(ii) transfer AT significantly boosts the transferred adversarial robustness of LM-VP models while also enhancing its training data privacy*. These findings indicate the advantage of LM-VP models trained with transfer AT in AI security. In future work, we will integrate relevant theories from related domains to delve deeper into the security implications of LM-VP models.

## Acknowledgments

ZC's contribution is supported by the University of Liverpool and China Scholarship Council (CSC). XH's contribution is supported by the UK EPSRC through End-to-End Conceptual Guarding of Neural Architectures [EP/T026995/1]. XZ's contribution is supported by the UK EPSRC New Investigator Award through Harnessing Synthetic Data Fidelity for Assured Perception of Autonomous Vehicles.

## References


[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[2] Huzaifa Arif, Alex Gittens, and Pin-Yu Chen. Reprogrammable-fl: Improving utility-privacy tradeoff in federated learning via model reprogramming. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 197–209. IEEE, 2023.

[3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[5] Aochuan Chen, Peter Lorenz, Yuguang Yao, Pin-Yu Chen, and Sijia Liu. Visual prompting for adversarial robustness. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[6] Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, and Sijia Liu. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19133–19143, 2023.

[7] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22584–22591, 2024.

[8] Zhen Chen, Fu Wang, Ronghui Mu, Peipei Xu, Xiaowei Huang, and Wenjie Ruan. Nrat: towards adversarial training with inherent label noise. *Machine Learning*, 113(6):3589–3610, 2024.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[10] Gamaleldin F. Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. In *International Conference on Learning Representations*, 2019.

[11] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[12] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[13] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[15] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192, 2022.

[16] Fengxiang He, Shaopeng Fu, Bohan Wang, and Dacheng Tao. Robustness, privacy, and generalization of adversarial training. *arXiv preprint arXiv:2012.13573*, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.


 https://doi.org/10.52202/079017-0217

[18] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

[19] Yizhe Li, Yu-Lin Tsai, Chia-Mu Yu, Pin-Yu Chen, and Xuebin Ren. Exploring the benefits of visual prompting in differential privacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5158–5167, 2023.

[20] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[24] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753, 2023.

[25] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 634–646, 2018.

[26] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations*, 2018.

[27] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International conference on machine learning*, pages 8093–8104. PMLR, 2020.

[28] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*, 2018.

[29] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

[30] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[31] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632, 2021.

[32] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 241–257, 2019.

[33] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[34] Pratik Vaishnavi, Kevin Eykholt, and Amir Rahmati. A study of the effects of transfer learning on adversarial robustness. *Transactions on Machine Learning Research*.

[35] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019.

[36] Fu Wang, Zeyu Fu, Yanghao Zhang, and Wenjie Ruan. Self-adaptive adversarial training for robust medical segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 725–735. Springer, 2023.

[37] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019.

[38] Yutaro Yamada and Mayu Otani. Does robustness on imagenet transfer to downstream tasks? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9215–9224, 2022.

[39] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[40] Xiangyu Yin and Wenjie Ruan. Boosting adversarial training via fisher-rao norm-based regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24544–24553, 2024.

[41] Sergey Zagoruyko. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[42] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR, 2023.

[43] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[44] Yanghao Zhang, Tianle Zhang, Ronghui Mu, Xiaowei Huang, and Wenjie Ruan. Towards fairness-aware adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24746–24755, 2024.

[45] Yi Zhang, Yun Tang, Wenjie Ruan, Xiaowei Huang, Siddartha Khastgir, Paul Jennings, and Xingyu Zhao. ProTIP: Probabilistic robustness verification on text-to-image diffusion models against stochastic perturbation. In *The 18th European Conference on Computer Vision (ECCV'24)*, 2024.

 https://doi.org/10.52202/079017-0217

# A Appendix / supplemental material

## A.1 Ablation Studies on Rescale Factor

In [19], they conclude the impact of the rescale ratio on the training of LM-VP models, *i.e.*, a larger rescale ratio yields better performance. However, an excessively large rescale ratio can also lead to overfitting to the target domain. In this section, we compare two rescale ratios, corresponding to the two cases shown in Fig. 2. In terms of performance, there is only a minimal gap between the two ways.

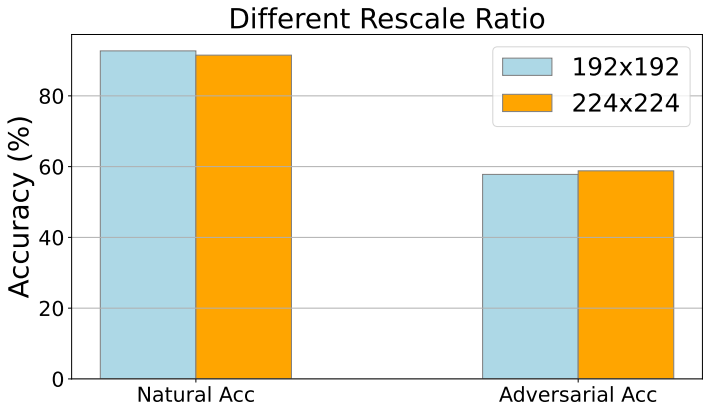



Figure 6: Comparison on different rescale ratios. The pre-trained model is Swin Transformer.

## A.2 Ablation Studies on Perturbation Limit

We set $\epsilon = 4/255$ and evaluate the LM-VP model based on the ResNet50 and ConvNext pre-trained models. The result is consistent with our main experiment where $\epsilon = 8/255$, *i.e.*, transfer AT achieves better transferred adversarial robustness and privacy trade-offs at the cost of natural accuracy.

Table 6: Best performance(%) on CIFAR-10 with two pre-trained models in Standard-Trained LM-VP models and Transfered AT-Trained LM-VP models under $\epsilon = 4/255$ of threat model ResNet18.

| **Pre-trained models** | **Standard Training** | | | **Transfer Adversarial Training** | | |
|---|---|---|---|---|---|---|
| | Natural$_{te}$ | PGD-20 | MIA Nat | Natural$_{te}$ | PGD-20 | MIA Nat |
| **ResNet50** | 84.90 | 33.19 | 73.99 | 67.96 | 65.10 | 51.91 |
| **ConvNext** | 97.72 | 86.86 | 79.84 | 97.68 | 89.77 | 51.20 |

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] ,

   Justification: The abstract and introduction show the main findings of this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and potential future work in the conclusion.


https://doi.org/10.52202/079017-0217

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: Some findings and claims in this paper are based on empirical findings, which may not be consistent with previous common intuition.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the setup and will make the code public. The empirical results can support our claims in the abstract, introduction, and conclusion.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Our experiments are based on public datasets, we have not yet released our code but we will create a GitHub repo for the code soon.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).


https://doi.org/10.52202/079017-0217

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have a setup section to show the above information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments can support the main claims in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computer resources in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: There is no ethics issues in this paper.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to our conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: There is no such issues in this paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.


https://doi.org/10.52202/079017-0217

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper of the original code we follow.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.


https://doi.org/10.52202/079017-0217