# STAR: MULTI-LEVEL ROBUSTNESS INJECTION FOR VISION TRANSFORMER DOWNSTREAM ADAPTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Pretrained Vision Transformers (ViTs) are increasingly adapted downstream by parameter-efficient fine-tuning (PEFT) under a small learnable-parameter budget $\rho$. Adapted models face transfer attacks from similar ViTs, and even white-box attacks if the learnable module leaks, so robust adaptation under such a $\rho$ is a practical need. Across PEFT methods and full fine-tuning, we find that standard adaptation is not robust and that adversarial training (AT) cannot balance clean accuracy and robustness once $\rho$ shrinks. We propose *Smoothness and Token Alignment Regularization* (`STAR`), a robust adaptation algorithm that spreads the robustness signal over multiple relaxed constraints that a small learnable subspace can satisfy: label-free class-token alignment, boundary smoothness, and down-weighted cross-entropies, at the representation, distribution, and decision levels, respectively. It applies unchanged to every adaptation method. Results from multiple datasets show that the advantage of `STAR` over AT is more pronounced under PEFT than under full fine-tuning. Analyses also show that robustness depends more on how learnable parameters reach the Encoder than on their number.

## 1 INTRODUCTION

Vision Transformers (ViTs) (Dosovitskiy et al., 2021) pretrained at scale serve as backbones of vision foundation models (Bommasani et al., 2021), whose checkpoints are reused across tasks. Rather than training from scratch, practitioners adapt a pretrained backbone to downstream tasks, increasingly by parameter-efficient fine-tuning (PEFT) (Jia et al., 2022; Han et al., 2024), which freezes the backbone and learns a small module. Linear probing (LP) (Yosinski et al., 2014), adapters (Adpt) (Houlsby et al., 2019), prefix tuning (PFT) (Li & Liang, 2021), prompt tuning (PT) (Lester et al., 2021), low-rank adaptation (LoRA) (Hu et al., 2022), and full fine-tuning (FFT) (Yosinski et al., 2014) span multiple orders of magnitude in the fraction $\rho$ of parameters they learn, from under 0.01% to 100%, which we call the *learnable-parameter budget*. Since adapted models share a public backbone, a black-box attacker can craft adversarial examples (AEs) on the same or similar ViTs and transfer them (Wei et al., 2022; Naseer et al., 2022). If the task module also leaks, the model is exposed to white-box attacks too, so its robustness matters as much as clean accuracy.

The adversarial robustness of ViTs has mostly been studied and improved for models trained from scratch or under FFT (Mahmood et al., 2021; Liu et al., 2023a; Jain & Dutta, 2024), where the whole model is trainable and no budget $\rho$ arises. For adapted models, prior work on CNNs and ViTs shows that a downstream task can partially inherit the robustness of a robust source model through adversarial FFT (Shafahi et al., 2020; Xu et al., 2024), adversarial distillation (Huang et al., 2025), or adversarial PEFT (Hua et al., 2024). However, this route requires an adversarially pretrained backbone, which is costly to produce (Salman et al., 2020). From a standard ViT backbone under a small $\rho$, adversarial training (AT) itself yields little robustness (Huang et al., 2026). Recent work either enlarges the learnable modules with heavier LoRA (Yuan et al., 2025) or dynamically selected critical weights (Li et al., 2026), essentially increasing $\rho$, or hardens PT alone (Eskandar et al., 2025). Throughout, the standard AT objective (Madry et al., 2018) based on projected gradient descent (PGD) is kept without systematic analysis across adaptation methods and their budgets $\rho$.

**Motivation.** PEFT is precisely adopted to deploy one small module per task on a shared frozen backbone (Houlsby et al., 2019; Hu et al., 2022), so the learnable-parameter budget $\rho$ is fixed by deployment before robustness is considered. AT requires model capacity far exceeding that of standard



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

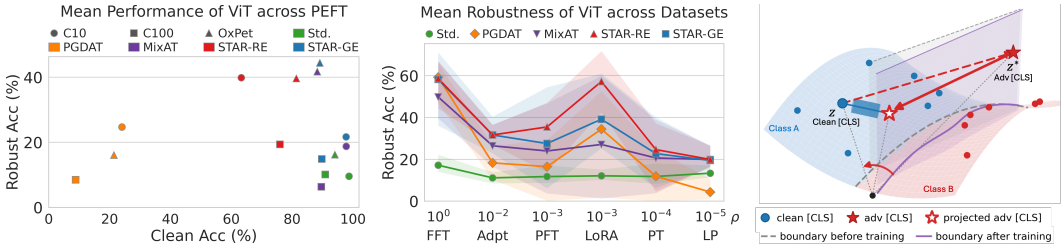


Figure 1: Overview of the STAR Study and Method. *Left:* Generalization against robustness of training mechanisms on each dataset, averaged over five PEFT adaptations. *Center:* Robustness of training mechanisms against learnable-parameter budget $\rho$, averaged over three datasets. *Right:* STAR aligns the adversarial class token $z^*$ with the clean $z$ and smooths the decision boundary.

training (Madry et al., 2018), which introduces an intrinsic conflict with PEFT, as PEFT restricts parameter updates to a small subspace (Hu et al., 2022). We therefore pose three research questions (RQs): ① *whether standard adaptation yields robustness at any budget* $\rho$; ② *whether existing AT survives a small* $\rho$; and ③ *if neither, what form of robust adaptation fits a small learnable subspace*.

**Our Study.** To answer ① and ②, we run a study that adapts a standard pretrained ViT to CIFAR-10, CIFAR-100, and Oxford-IIIT Pet with PEFT methods (i.e., Adpt, PFT, LoRA, PT, LP) and FFT under three training objectives: standard adaptation and the AT baselines PGDAT (Madry et al., 2018) and MixAT (Xie et al., 2020). The answer to ① is negative, as standard adaptation keeps above 93% clean accuracy on average at every $\rho$ while its black-box robust accuracy stays below 18% (Figure 1, center), and PGD drives it to near-zero robust accuracy at every $\rho$. The answer to ② is also negative. AT performs best only under FFT, and with a small $\rho$, clean accuracy of PGDAT under LP falls to the chance rate. MixAT keeps clean accuracy, but its best PEFT result (LoRA) keeps only half of its FFT black-box robust accuracy, and on complex CIFAR-100 even falls below standard adaptation (Figure 1, left and center). We trace degradation to a tension, as AT asks for a hard constraint that a small learnable subspace cannot meet together with accuracy.

**Our Method.** To answer ③, we propose a new robust adaptation algorithm, *Smoothness and Token Alignment Regularization* (STAR), which replaces the single hard robustness constraint with multiple relaxed ones, placed at three levels and computed from the class tokens of a clean input and its AE. ($i$) At the representation level, an alignment term rotates the adversarial class token toward the clean direction (Figure 1, right), a label-free constraint that a few Encoder parameters can meet. ($ii$) At the distribution level, a smoothness term asks the prediction at the interpolation of the two class tokens to match the clean distribution, which needs only a slight boundary shift rather than a correct label for the AE. ($iii$) At the decision level, the clean and adversarial cross-entropies anchor the true labels with reduced weight. Both new terms reuse the two class tokens and only add head evaluations, so STAR costs about as much as AT. Two modes of STAR are provided, with STAR-RE toward robustness capability and STAR-GE toward generalization. STAR never refers to where the learnable parameters sit, so it applies to every PEFT and FFT method.

**Experiments.** We evaluate all adapted ViT models under two threat models. A black-box attacker crafts AEs on two surrogates, PiT-B (Heo et al., 2021) and DeiT-B (Touvron et al., 2021), with three ViT-specific transfer attacks (Wei et al., 2022; Naseer et al., 2022; Zhang et al., 2023). A white-box attacker runs ViT-specific unbounded patch attacks, Patch-Fool (Fu et al., 2022) and Attention-Fool (Lovisotto et al., 2022), and $\ell_\infty$-bounded PGD and AutoAttack (Croce & Hein, 2020). We also trace how $\rho$ shapes robustness and training cost, and ablate the loss terms and weights.

In summary, we make the following contributions:

- STAR, a regularization that replaces the hard constraint of AT with label-free class-token alignment and boundary smoothness, which fit a small subspace and apply to any adaptation method.
- A systematic study across $\rho$, showing that standard adaptation lacks robustness, that AT severely degrades under a small $\rho$, and that its PGD robustness misses ViT-specific attacks.
- Experiments on ViTs over multiple datasets and PEFT methods under black-box and white-box attacks, where STAR balances clean accuracy and robustness better than AT under a small $\rho$.
- Analyses showing that robustness should be built inside the Encoder, that how parameters reach it matters more than how many there are, and that class-token alignment drives the gain of STAR.

## 2 PROBLEM SCOPE

**Vision Transformer.** A ViT (Dosovitskiy et al., 2021) treats an image $x \in \mathcal{X} = [0,1]^{3\times H\times W}$ as a sequence of $N = HW/P^2$ non-overlapping patches of size $P \times P$. Each patch $x_i^p$ is flattened and linearly projected to $D$ dimensions by $E_S \in \mathbb{R}^{(3P^2)\times D}$, a learnable *class token* $z_0^{cls}$ ([CLS]) is prepended and positional embeddings $E_{\text{pos}}$ are added, giving the input sequence $\mathbf{z}_0 = \left[z_0^{cls}; x_1^p E_S; \ldots; x_N^p E_S\right] + E_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$. The *Encoder* $h : \mathcal{X} \to \mathbb{R}^D$ applies $L$ Transformer blocks (Vaswani et al., 2017) to $\mathbf{z}_0$ and returns the [CLS] of the final block after its layer normalization, $z = h(x) = z_L^{cls}$, and the *Classification Head* $g(z) = Wz + b$ maps that token to logits over $C$ classes. With $\sigma$ the softmax, the ViT model $f_\theta = g \circ h$ produces the output distribution $p = \sigma(g(h(x)))$ and the predicted label $\hat{y} = \arg\max_c p_c$. Appendix A summarizes the notation.

**Downstream Adaptation.** In the parameter-update stage of transfer learning, a pretrained backbone is specialized to a target task under an adaptation budget. Standard downstream adaptation partitions the parameters $\theta$ of the pretrained model into a frozen and a trainable part, $\theta = (\theta_{\text{frz}}, \theta_{\text{ada}})$, and fits $\theta_{\text{ada}}$ to a target-domain dataset $\mathcal{D} = \{(x, y)\}$ with ground-truth labels $y \in \{1, \ldots, C\}$ by minimizing the cross-entropy loss $\mathcal{L}_{\text{CE}}(p, y) = -\log p_y$, as defined in Eq. 1.

$$\min_{\theta_{\text{ada}}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}_{\text{CE}}\big(\sigma(g(h(x))),\, y\big)\right]. \tag{1}$$

Many downstream adaptation methods instantiate Eq. 1, differing only in what $\theta_{\text{ada}}$ contains and where it intervenes in the forward pass. The Classification Head is trained in every case.

- Linear probing (LP) freezes the Encoder parameters and trains the Classification Head alone.
- Adapters (Adpt) keep every weight frozen and insert a small bottleneck after the multi-layer perceptron of each Transformer block, learning a residual correction to its output.
- Prefix tuning (PFT) prepends learnable key and value prefix vectors to the attention of every Transformer block, steering attention layer by layer without changing any weight.
- Prompt tuning (PT) intervenes only at the input, prepending learnable tokens to the input $\mathbf{z}_0$.
- LoRA freezes each $qkv$ projection matrix and learns a low-rank residual beside it, confining the update to a small subspace of that projection.
- Full fine-tuning (FFT) updates every parameter.

The key difference among various adaptation methods lies in the extent to which the pretrained backbone can be updated, which determines the capacity of learnable parameters available for the optimization objective. This capacity is referred to as the ***learnable-parameter budget***; see Eq. 2.

$$\rho = \frac{|\theta_{\text{ada}}|}{|\theta|} = \frac{|\theta_{\text{ada}}|}{|\theta_{\text{frz}}| + |\theta_{\text{ada}}|} \in (0, 1], \tag{2}$$

where $|\cdot|$ counts scalar parameters and the Classification Head is counted in both terms.

**Threat Model.** The attacker aims to induce untargeted misclassification of test inputs for a deployed model $f_\theta$ during inference, and cannot modify the model, training data, or adaptation. We consider two levels of attacker knowledge.

(*i*) A *transfer-based black-box* attacker knows only that the target is a ViT-family classifier for the task, has neither its weights nor query access, and optimizes the input perturbation $\delta$ on its own surrogate model $\tilde{f}$, relying on adversarial examples (AEs) transferring across architectures.

(*ii*) A *white-box* attacker knows the architecture, adaptation method, and full parameters of $f_\theta$, and computes gradients $\nabla_\delta$ through $f_\theta$ itself to optimize $\delta$. This strongest assumption upper-bounds attack strength, so the accuracy measured under it reflects the robustness lower bound.

The defender controls only the adaptation stage: it may choose the adaptation method and training objective, but does not retrain the backbone or alter the architecture, and it learns under a budget $\rho$.

**Problem Statement.** The robust downstream adaptation is the problem of finding $\theta_{\text{ada}}$, for a given backbone, dataset, and $\rho$, such that the adapted model is accurate on clean inputs and robust against evasion attacks. Our approach replaces the objective of Eq. 1 with a new one that adds a robustness term $\mathcal{R}$, regularizing the model based on both clean inputs $x$ and AEs $x^*$ generated during training.

$$\min_{\theta_{\text{ada}}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathcal{L}_{\text{CE}}\big(\sigma(g(h(x))),\, y\big) + \mathcal{R}\big(x,\, x^*,\, y;\, \theta\big)\Big]. \tag{3}$$

We investigate three **research questions** for ViTs: ① *Does standard adaptation yield robustness under various learnable-parameter budgets $\rho$?* ② *Do existing robustness terms, such as adversarial training, survive a small $\rho$?* ③ *If neither, what form of $\mathcal{R}$ can fit a small learnable subspace?*

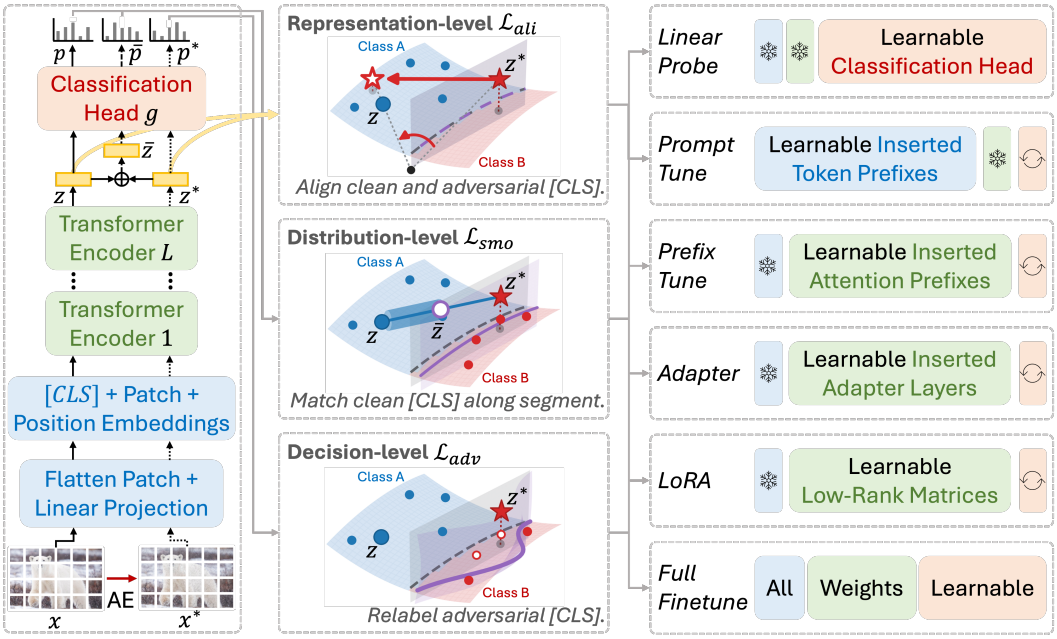


Figure 2: Framework of STAR. Left: A ViT maps clean images and their adversarial counterparts to their respective class tokens [CLS], $z$ and $z^*$, which the classification head $g$ turns into predictions, $p$ and $p^*$. Center: Impact of the loss terms on the model robustness. Blue and red sheets represent class tokens for two distinct classes. Gray dashed and purple solid lines denote the decision boundaries trained with clean and robust losses, respectively. Right: Various adaptation methods based on STAR, where snowflake icons mark frozen modules, and refresh icons mark learnable ones.

## 3 STAR: Smoothness and Token Alignment Regularization

We propose STAR, a robust training algorithm for the downstream adaptation of ViTs. Its design principle is to distribute robustness signals across several supervised and unsupervised weak constraints, which a small learnable-parameter subspace can satisfy, rather than relying on a single supervised constraint based on hard labels. It is realized by four loss functions at three model levels.

### 3.1 Adaptation Procedure

**Dual Path.** STAR operates inside the adaptation stage, with $\theta_{\text{frz}}$ frozen and $\theta_{\text{ada}}$ updated, and each training step propagates a batch along two paths through the same weights (see Figure 2, left). The *clean path* returns the class token [CLS] $z = h(x)$ and its output distribution $p = \sigma(g(z))$. The *adversarial path* first generates an adversarial example (AE) $x^*$ of $x$ subject to the training-time perturbation budget $\epsilon$ by solving the maximization problem, as defined in Eq. 4.

$$x^* = \mathcal{A}(x, y; \theta) \approx \underset{x' \in \mathcal{X},\ \|x'-x\|_\infty \le \epsilon}{\arg\max} \mathcal{L}_{\text{CE}}(\sigma(g(h(x'))), y), \quad z^* = h(x^*), \quad p^* = \sigma(g(z^*)), \tag{4}$$

Any attack algorithm returning an approximate maximizer can serve as the solver $\mathcal{A}$, which runs with $\theta$ fixed, and this work employs $\ell_\infty$ projected gradient descent.

**Class token mixup.** STAR also forms the midpoint $\bar{z}$ of the clean and adversarial class tokens and obtains its output distribution $\bar{p}$ through the classification head $g$ and softmax $\sigma$, as defined in Eq. 5.

$$\bar{z} = \tfrac{1}{2}(z + z^*), \qquad \bar{p} = \sigma(g(\bar{z})) \tag{5}$$

Since the head is an affine transformation, the logit $g(\bar{z})$ of the midpoint of the two class tokens equals the midpoint of their logits, i.e., $g(\bar{z}) = \frac{1}{2}(g(z) + g(z^*))$. Through the softmax, an arithmetic mean of the logits further becomes a geometric mean of the output distributions, as derived in Eq. 6, where $c \in \{1, \ldots, C\}$ indexes classes, $g(z)_c$ is the logit of class $c$, and $p_c = \sigma(g(z))_c$ is the derived probability. The proportionality sign $\propto$ hides the softmax normalizers, which do not depend on $c$.

$$\bar{p}_c = \sigma(g(\bar{z}))_c = \frac{e^{\frac{1}{2}(g(z)_c + g(z^*)_c)}}{\sum_{k=1}^{C} e^{\frac{1}{2}(g(z)_k + g(z^*)_k)}} \propto \sqrt{e^{g(z)_c}\, e^{g(z^*)_c}} \propto \sqrt{\sigma(g(z))_c\, \sigma(g(z^*))_c} = \sqrt{p_c\, p_c^*}. \tag{6}$$

Combining the classes yields $\bar{p} \propto \sqrt{p \odot p^*}$, where $\odot$ is the elementwise product, and the normalizing sum $\sum_c \sqrt{p_c\, p_c^*}$ measures the overlap between $p$ and $p^*$, equaling 1 only when $p = p^*$. Since $\sqrt{p_c\, p_c^*} \leq \frac{1}{2}(p_c + p_c^*)$ with equality only when $p_c = p_c^*$, $\bar{p}$ essentially discounts the probability values of every class based on the degree of divergence between the two paths regarding those classes, and shifts the corresponding probability mass to the classes on which the two paths are in agreement. Thus, the output distribution $\bar{p}$ of interpolated class token $\bar{z}$ is a compromise between $p$ and $p^*$ (Appendix D.1), sharp for classes on which the two paths result in a similar probability, and flattened for classes where one path assigns it high probability and the other low. Both $z$ and $z^*$ are computed once and reused, so each step runs the Encoder once per path, while the head runs three times.

### 3.2 Multi-level Robustness Regularization

We formulate three loss terms to inject adversarial robustness into the model at three levels: $\mathcal{L}_{\text{ali}}$ at the geometry of the representations, $\mathcal{L}_{\text{smo}}$ at the output distributions, and $\mathcal{L}_{\text{adv}}$ at the model decisions.

**Training Objective.** The three terms and the clean task loss are combined with non-negative weights $\boldsymbol{\lambda} = (\lambda_{\text{cle}}, \lambda_{\text{adv}}, \lambda_{\text{smo}}, \lambda_{\text{ali}})$, and their expectation is minimized over $\theta_{\text{ada}}$ alone, as defined in Eq. 7, so robustness is injected into the lightweight adaptation component, while the representation learner is primarily provided by the pretrained backbone. Algorithm 1 states one epoch of the procedure.

$$\mathcal{L}_{\text{STAR}}(\boldsymbol{\lambda}) = \lambda_{\text{cle}}\mathcal{L}_{\text{cle}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{smo}}\mathcal{L}_{\text{smo}} + \lambda_{\text{ali}}\mathcal{L}_{\text{ali}}, \qquad \min_{\theta_{\text{ada}}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\mathcal{L}_{\text{STAR}}\big]. \tag{7}$$

**Representation level.** Since the head reads only the class token [CLS], an AE can change the decision solely through the displacement of $z^*$ from $z$, so an Encoder that keeps the two close mitigates the attack before it reaches the head. To map $x^*$ close to $x$ in the representation space of class tokens, the alignment loss $\mathcal{L}_{\text{ali}}$ penalizes the angular distance between the clean and adversarial class tokens, as defined in Eq. 8 and shown in Figure 2 as the redirection of $z^*$ toward $z$.

$$\mathcal{L}_{\text{ali}} = 1 - \cos(z, z^*) = 1 - \frac{\langle z,\, z^* \rangle}{\|z\|_2\, \|z^*\|_2} \;\in\; [0, 2], \tag{8}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity and $\langle\cdot,\cdot\rangle$ the Euclidean inner product. Although $\mathcal{L}_{\text{ali}}$ involves neither the head nor the label, it still fosters robust decisions, because $\mathcal{L}_{\text{ali}} = 0$ holds exactly when $z^* = \alpha z$ (where $\alpha > 0$), in which case $g(z^*) = g(\alpha z) = \alpha g(z) + (1-\alpha)b$. The two logits differ only by a scaling and the bias term, therefore, when $b = 0$, the decisions for $x^*$ and $x$ are identical.

**Distribution level.** The smoothness loss $\mathcal{L}_{\text{smo}}$ requires the output distribution $\bar{p}$ of the interpolated [CLS] $\bar{z}$ to be consistent with the output distribution $p$ of the clean [CLS] $z$, as defined in Eq. 9.

$$\mathcal{L}_{\text{smo}} = D_{\text{KL}}\big(\bar{p} \,\big\|\, p\big) = \sum_{c=1}^{C} \bar{p}_c \log \frac{\bar{p}_c}{p_c}, \tag{9}$$

where $D_{\text{KL}}$ is the Kullback-Leibler (KL) divergence, which equals 0 only when $\bar{p} = p$. $\bar{p}$ is its first argument, and neither $\bar{p}$ nor $p$ is treated as a fixed target in our design, so $\bar{p}$ and $p$ are pulled toward each other rather than $\bar{p}$ alone toward a fixed $p$, driving simultaneous updates to adversarial representations, clean predictions, and boundaries. At the level of decisions, since the head $g$ is affine, writing the margin of the true class $y$ over any other class $c \neq y$ at a class token $u$ as $m_c(u) = g(u)_y - g(u)_c$, the midpoint identity $g(\bar{z}) = \frac{1}{2}(g(z) + g(z^*))$ gives $m_c(\bar{z}) = \frac{1}{2}\big(m_c(z) + m_c(z^*)\big)$. Keeping $\bar{z}$ on the clean side of each decision boundary indicates that no boundary crosses the first half of the segment $[z, z^*]$. Thus, $\mathcal{L}_{\text{smo}}$ can be viewed as a relaxation of the cross-entropy at $x^*$, requiring consistency along one segment rather than enforcing correct hard-label classification for the AE. In addition, pulling $p$ toward $\bar{p}$ can be viewed as the cross-entropy with a soft label, encouraging boundary smoothing and preventing the model from overfitting or becoming overconfident.

**Decision level.** Finally, the task loss $\mathcal{L}_{\text{cle}}$, which maintains downstream accuracy, is combined with the AT loss $\mathcal{L}_{\text{adv}}$, which requires the correct hard label on AEs, as defined in Eq. 10. Of the four terms, only these two carry the ground-truth label $y$ to the clean and adversarial paths, respectively.

$$\mathcal{L}_{\text{cle}} = \mathcal{L}_{\text{CE}}(p,\, y), \qquad \mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{CE}}(p^*,\, y). \tag{10}$$

**Algorithm 1** One epoch of robust downstream adaptation based on STAR

**Require:** dataset $\mathcal{D}$, model $\theta = (\theta_{\text{frz}}, \theta_{\text{ada}})$, attack solver $\mathcal{A}$, weights $\boldsymbol{\lambda}$, optimizer Opt over $\theta_{\text{ada}}$ only

**for** mini-batch $(x, y) \sim \mathcal{D}$ **do**
  $x^* \leftarrow \mathcal{A}(x, y; \theta)$ ▷ Eq. 4, $\theta$ fixed, $x^*$ detached
  $z \leftarrow h(x), \quad z^* \leftarrow h(x^*)$ ▷ class tokens, shared weights
  $\bar{z} \leftarrow \frac{1}{2}(z + z^*)$ ▷ Eq. 5
  $\mathcal{L}_{\text{ali}} \leftarrow \text{mean}(1 - \cos(z, z^*))$ ▷ Eq. 8, batch means throughout
  $\mathcal{L}_{\text{smo}} \leftarrow D_{\text{KL}}(\sigma(g(\bar{z})) \,\|\, \sigma(g(z)))$ ▷ Eq. 9
  $\mathcal{L}_{\text{cle}} \leftarrow \mathcal{L}_{\text{CE}}(\sigma(g(z)), y), \quad \mathcal{L}_{\text{adv}} \leftarrow \mathcal{L}_{\text{CE}}(\sigma(g(z^*)), y)$ ▷ Eq. 10
  $\mathcal{L}_{\text{STAR}} \leftarrow \lambda_{\text{cle}}\mathcal{L}_{\text{cle}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{smo}}\mathcal{L}_{\text{smo}} + \lambda_{\text{ali}}\mathcal{L}_{\text{ali}}$ ▷ Eq. 7
  $\theta_{\text{ada}} \leftarrow \text{Opt}(\theta_{\text{ada}}, \nabla_{\theta_{\text{ada}}}\mathcal{L}_{\text{STAR}})$ ▷ $\theta_{\text{frz}}$ unchanged
**end for**

### 3.3 Why STAR Fits a Small Learnable-Parameter Budget

**Robust representation.** With all weights trainable, AT ordinarily obtains robustness by relearning the whole Encoder, so that an AE moves the class token less, i.e., the displacement $z^* - z$ shrinks. Under PEFT, the pretrained weights stay frozen, and $h$ can change only through the few parameters of $\theta_{\text{ada}}$ in its forward pass. With $z^*$ decomposed into a radial component along $z$ and a tangential one orthogonal to $z$, $\mathcal{L}_{\text{ali}}$ depends only on their ratio, i.e., the angle between $z$ and $z^*$, and leaves the norm $\|z^*\|$ free, which does not affect the decision of the head when $b = 0$. $\mathcal{L}_{\text{ali}}$ thus asks only for a rotation of $z^*$ toward $z$, a weaker demand than shrinking the whole displacement, and its gradient reaches only the learnable parameters inside $h$ and leaves the head entirely untouched.

**Robust margin.** When $\theta_{\text{ada}}$ can change $h$ only slightly, the burden of $\mathcal{L}_{\text{adv}}$, classifying every $z^*$ as $y$, falls mostly on the head. As the head is affine and attacks push adversarial [CLS] in sample-specific directions, correcting them moves many hyperplanes, so wherever adversarial [CLS] of one class lie among the clean [CLS] of another, the clean ones are misclassified instead. Thus, $\mathcal{L}_{\text{adv}}$ alone costs much of the clean accuracy under a small $\rho$. $\mathcal{L}_{\text{smo}}$ asks for less, as it compares $\bar{p}$ with the clean distribution $p$ rather than with the hard label $y$, and the midpoint halves the logit gap between the two paths. Keeping the clean decision at $\bar{z}$ only requires that no boundary cross the first half of the segment $[z, z^*]$, not that $z^*$ itself be classified correctly. Once $\mathcal{L}_{\text{ali}}$ narrows the angle between $z$ and $z^*$, a small shift of the boundary suffices, sparing the nearby clean [CLS] of the other classes.

**Multiple levels.** Whatever $\rho$ is, the head remains one affine layer that can only shift and tilt its hyperplanes, while an adapted Encoder can move the class tokens themselves. Moving weight from $\mathcal{L}_{\text{adv}}$ to $\mathcal{L}_{\text{ali}}$ thus places the robustness signal where the model can absorb it, raising both clean and transfer robust accuracy, so both modes of STAR give $\lambda_{\text{ali}}$ the largest weight. Moreover, $\mathcal{L}_{\text{adv}}$ never reaches zero, whereas both new terms vanish once the Encoder maps $x^*$ to the same [CLS] as $x$, whatever the head predicts, so the adapted Encoder can reduce them without changing the head.

## 4 Experiments

We adapt pretrained ViTs to three datasets with six adaptation methods and five training conditions, and measure every model under black-box and white-box attacks. After the setup (§4.1), the black-box (§4.2) and white-box (§4.3) evaluations answer RQs and show where STAR stands. We then examine how the learnable-parameter budget $\rho$ shapes these results and what they cost (§4.4), and ablate the loss terms and weights (§4.5). Appendices C and E detail the setup and all results.

### 4.1 Setup

**Models and datasets.** We use ViT-B/16 (Dosovitskiy et al., 2021) pretrained on ImageNet-1k as the backbone, and the self-supervised DINOv2 (Oquab et al., 2024) in Appendix E.6. Neither pretraining sees an adversarial example (AE), so any robustness is produced by the adaptation. We adapt it to CIFAR-10, CIFAR-100 (Krizhevsky, 2009), each with 50k training images, and Oxford-IIIT Pet (Parkhi et al., 2012), a low-data task with 37 fine-grained classes and 3.7k training images.

**Adaptation baselines.** The adaptation methods span the learnable-parameter budget $\rho$ from the classification head alone to the whole backbone: Linear Probing (LP) (Yosinski et al., 2014), Adapters (Adpt) (Houlsby et al., 2019), Prefix Tuning (PFT) (Li & Liang, 2021), Prompt Tuning (PT) (Lester

Table 1: Robustness (%) of Downstream ViTs against Black-Box Attacks across Different Datasets.

| | | CIFAR-10 | | | | | | | Oxford-IIIT Pet-37 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | PiT-B Proxy | | | DeiT-B Proxy | | | | | PiT-B Proxy | | | DeiT-B Proxy | | |
| TL | Method | $\rho(\%)$ | Clean | PNA | SE | TGR | PNA | SE | TGR | $\rho(\%)$ | Clean | PNA | SE | TGR | PNA | SE | TGR |
| LP | Standard | 0.009 | **98.04** | 26.60 | 9.90 | 17.90 | 5.60 | 3.40 | 4.70 | 0.033 | **94.06** | 17.50 | 21.60 | 33.20 | 10.00 | 8.90 | 11.10 |
| | PGDAT | | 10.00 | 10.60 | 10.60 | 10.60 | 10.60 | 10.60 | 10.60 | | 2.73 | 1.50 | 1.50 | 1.50 | 1.50 | 1.50 | 1.50 |
| | MixAT | | 95.96 | 33.00 | 12.80 | 30.40 | 10.20 | 4.60 | 7.80 | | 91.20 | 31.50 | 28.80 | 45.20 | 20.10 | 13.40 | 20.50 |
| | STAR-RE | | 95.89 | **33.30** | **12.90** | 30.10 | 10.30 | 4.40 | 8.10 | | 90.24 | **31.80** | 28.10 | 44.30 | **21.00** | 13.30 | **20.60** |
| | STAR-GE | | 96.50 | 32.60 | 12.60 | **30.70** | 9.40 | 4.60 | 7.40 | | 92.59 | 31.00 | **29.10** | **46.90** | 18.00 | **13.70** | 19.20 |
| PT | Standard | 0.027 | **98.37** | 16.40 | 7.60 | 14.30 | 5.00 | 3.20 | 3.30 | 0.051 | **93.89** | 20.00 | 23.50 | 32.40 | 11.40 | 9.60 | 10.90 |
| | PGDAT | | 24.18 | 24.50 | 24.20 | 24.60 | 24.80 | 24.40 | 24.50 | | 2.73 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MixAT | | 97.42 | 19.60 | 19.10 | 19.00 | 19.50 | 20.00 | 18.10 | | 86.97 | 47.90 | 46.10 | **51.60** | 36.00 | 25.60 | 26.20 |
| | STAR-RE | | 37.56 | **38.30** | **36.60** | **37.00** | **38.60** | **34.50** | **37.90** | | 76.29 | 40.50 | 35.80 | 37.70 | 32.40 | 22.40 | 24.50 |
| | STAR-GE | | 96.65 | 21.90 | 22.70 | 21.80 | 23.10 | 19.20 | 22.30 | | 86.73 | **50.80** | **47.10** | 50.50 | **37.50** | **26.20** | **28.10** |
| LoRA | Standard | 0.181 | **99.13** | 24.90 | 10.40 | 18.60 | 4.80 | 3.40 | 3.60 | 0.205 | **94.36** | 14.90 | 20.20 | 32.80 | 7.10 | 7.40 | 8.30 |
| | PGDAT | | 48.33 | 49.40 | 48.70 | 48.60 | 49.50 | 48.90 | 49.20 | | 63.29 | 54.00 | 53.00 | 53.80 | 53.60 | 49.40 | 50.80 |
| | MixAT | | 99.04 | 20.00 | 20.30 | 20.40 | 20.00 | 20.10 | 20.00 | | 82.72 | 64.00 | 62.20 | 62.40 | 60.30 | 54.80 | 55.50 |
| | STAR-RE | | 75.50 | **72.80** | **71.50** | **71.30** | **72.70** | **70.70** | **71.50** | | 76.23 | 60.60 | 58.80 | 60.20 | 58.30 | 52.60 | 54.10 |
| | STAR-GE | | 98.87 | 25.60 | 25.20 | 25.30 | 25.40 | 24.90 | 25.30 | | 87.76 | **66.40** | **63.60** | **63.40** | **60.60** | **55.50** | **56.30** |
| PFT | Standard | 0.224 | **98.80** | 19.10 | 7.70 | 16.20 | 3.50 | 3.70 | 2.40 | 0.248 | **93.92** | 17.10 | 23.50 | 34.20 | 8.80 | 9.80 | 8.90 |
| | PGDAT | | 27.63 | 29.50 | 29.10 | 29.40 | 29.30 | 28.50 | 29.10 | | 5.21 | 0.30 | 0.30 | 0.20 | 0.10 | 0.20 | 0.20 |
| | MixAT | | 98.56 | 22.40 | 21.40 | 20.60 | 22.50 | 20.50 | 22.10 | | 90.68 | 54.80 | 55.80 | 59.30 | 42.30 | 32.50 | 35.50 |
| | STAR-RE | | 47.88 | **47.90** | **47.00** | **47.10** | **47.80** | **45.50** | **46.80** | | 77.79 | 51.50 | 48.40 | 50.40 | 45.30 | 37.60 | 40.20 |
| | STAR-GE | | 98.18 | 23.10 | 22.30 | 21.90 | 22.40 | 21.00 | 21.80 | | 87.68 | **62.20** | **61.50** | **61.70** | **53.50** | **42.20** | **45.40** |
| Adpt | Standard | 1.395 | **99.06** | 18.90 | 7.90 | 16.00 | 3.10 | 2.80 | 2.20 | 1.419 | **94.19** | 17.40 | 19.40 | 28.80 | 6.40 | 5.30 | 5.30 |
| | PGDAT | | 10.00 | 10.20 | 10.20 | 10.20 | 10.20 | 10.20 | 10.20 | | 33.17 | 27.20 | 26.20 | 26.40 | 26.70 | 25.20 | 26.50 |
| | MixAT | | 97.82 | 17.70 | 16.60 | 18.70 | 18.70 | 12.20 | 14.40 | | 89.75 | 40.20 | 35.90 | 38.80 | 41.60 | 28.50 | 34.10 |
| | STAR-RE | | 59.51 | **29.50** | **25.60** | **26.60** | **31.70** | **20.90** | **27.00** | | 86.35 | 40.30 | 34.40 | 36.90 | 39.90 | 30.80 | 36.20 |
| | STAR-GE | | 98.48 | 24.00 | 24.30 | 24.20 | 22.40 | 20.60 | 22.00 | | 90.79 | **44.00** | **39.20** | **43.20** | **45.50** | **31.80** | **38.30** |
| FFT | Standard | 100 | 98.30 | 39.80 | 38.70 | 33.10 | 9.80 | 6.60 | 5.10 | 100 | **92.53** | 23.90 | 21.30 | 27.10 | 9.40 | 4.50 | 5.60 |
| | PGDAT | | 65.54 | 65.10 | 64.90 | 64.20 | 65.20 | **63.90** | **65.40** | | 77.90 | 67.00 | 67.00 | 66.90 | 65.80 | 64.40 | 64.40 |
| | MixAT | | **98.84** | 57.50 | 59.00 | 52.50 | 35.00 | 27.40 | 27.10 | | 79.10 | 67.90 | 67.90 | 67.70 | 66.70 | 65.10 | 65.00 |
| | STAR-RE | | 98.62 | 57.40 | 58.40 | 54.60 | 34.60 | 30.90 | 30.80 | | 78.30 | 68.10 | 67.90 | 67.60 | 67.50 | 65.30 | **65.80** |
| | STAR-GE | | **98.84** | **79.70** | **80.90** | **75.90** | **66.50** | 60.50 | 63.00 | | 79.48 | **68.90** | **68.30** | **68.80** | **67.60** | **65.50** | 65.40 |

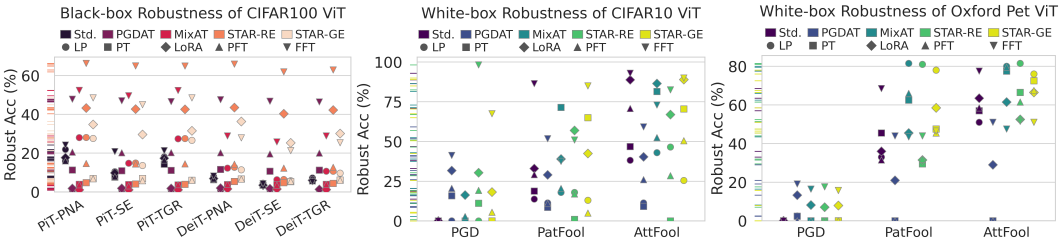


Figure 3: Robustness (%) of Downstream ViTs against Black-box and White-box Attacks.

et al., 2021), Low-rank adaptation (LoRA) (Hu et al., 2022), and full fine-tuning (FFT) (Yosinski et al., 2014). They are combined with five training conditions: standard adaptation, adversarial training (AT) baselines PGDAT (Madry et al., 2018) and MixAT (Xie et al., 2020), and `STAR` with two modes: `RE` (Robustness Enhance-oriented) and `GE` (Generalization Enhance-oriented).

**Evasion Attacks.** Every model is measured under both the black-box and white-box threats.

- Three transfer-based black-box attacks, all specifically proposed for ViTs: PNA (Pay No Attention with PatchOut) (Wei et al., 2022), SE (Self-Ensemble with Token Refinement) (Naseer et al., 2022), and TGR (Token Gradient Regularization) (Zhang et al., 2023). Each is mounted with MI-FGSM (Momentum Iterative Fast Gradient Sign Method) (Dong et al., 2018) at $\epsilon = 16/255$ and run from two surrogate models: PiT-B (Heo et al., 2021) and DeiT-B (Touvron et al., 2021).
- Three white-box attacks: The two ViT-specific ones, PatFool (Patch-Fool) (Fu et al., 2022) and AttFool (Attention-Fool) (Lovisotto et al., 2022), each optimizes an unbounded perturbation for 250 steps to a single patch that receives the most attention in block 5. Another is PGD (Madry et al., 2018) under an $\ell_\infty$ perturbation budget of $\epsilon = 8/255$ with step size $\epsilon_s = 2/255$ for 20 steps.

AutoAttack (AA) (Croce & Hein, 2020) is additionally run on every CIFAR-10 model and the loss-weight sweep. The attack $\mathcal{A}$ used during training (10 PGD steps at $\epsilon = 8/255$ with $\epsilon_s = 2/255$) differs from every evaluation attack, so none of the results stem from the seen attack during training.

**Metrics.** Clean accuracy, $CleAcc = \frac{1}{|\mathcal{D}_{\text{te}}|} \sum_{(x,y)\in\mathcal{D}_{\text{te}}} \mathrm{I}[\hat{y}(x) = y]$, is evaluated on the clean test set $\mathcal{D}_{\text{te}}$, where $\hat{y}(x)$ is the predicted label and $\mathrm{I}[\cdot]$ is the indicator function. Robust accuracy under attack $\mathcal{A}$ on the adversarial test set $\mathcal{D}^*_{\text{te}}$ is defined as $RobAcc(\mathcal{A}) = \frac{1}{|\mathcal{D}^*_{\text{te}}|} \sum_{(x,y)\in\mathcal{D}^*_{\text{te}}} \mathrm{I}[\hat{y}(\mathcal{A}(x,y)) = y]$.

### 4.2 Evaluation of Black-box Adversarial Robustness

We report the six transfer attacks on the three datasets in Table 1 and Figure 3. For RQ ①, standard adaptation fits every task but stays far from robust at every $\rho$, even under FFT, so robustness has to come from the training objective rather than from the choice of adaptation method. For RQ ②, AT is at its best under FFT but breaks down under a small $\rho$, where PGDAT gives up the task, with Clean Acc down to the chance rate under LP, and MixAT keeps the task but gives up much of its robustness, on CIFAR-100 even falling below standard adaptation. For RQ ③, `STAR` avoids both failures. With LoRA, `STAR-RE` is robust on nearly every input it classifies, as PGDAT is, but classifies far more of them, reaching about 72% Robust Acc on CIFAR-10 with 27 points more Clean Acc than PGDAT, while `STAR-GE` keeps the Clean Acc of MixAT and gains about 30 points of Robust Acc over it on CIFAR-100. The lead holds even under FFT, where `STAR-GE` exceeds MixAT by more than 20 points under every attack on CIFAR-10 at the same Clean Acc, and it is largest under the small $\rho$ where AT breaks down. Spreading the robustness signal over several weak constraints thus fits a small learnable subspace where the adversarial cross-entropy alone does not.

### 4.3 Evaluation of White-box Adversarial Robustness

We report the three white-box attacks in Figure 3. For RQ ①, standard adaptation has no white-box $\ell_\infty$ robustness, as PGD drives it to essentially 0 Robust Acc at every $\rho$ on all datasets. For RQ ②, AT under a small $\rho$ again cannot retain both the task and the robustness, since PGDAT holds PGD robustness only where its Clean Acc has collapsed, while MixAT, which keeps the task, holds little of it. Even under FFT, where AT works best, both PGDAT and MixAT fall far below standard adaptation under PatFool, implying that the PGD robustness gained through AT does not generalize to ViT-specific attacks. For RQ ③, `STAR` attains white-box robustness while still fitting the task. As PGD can overestimate robustness under gradient masking, e.g., on CIFAR-10 under FFT, where MixAT and `STAR` reach 67% to 99% under PGD but at most 2% under AA, we rely on AA to measure $\ell_\infty$ robustness. With LoRA on CIFAR-10, `STAR-RE` comes close to PGDAT under both PGD and AA while keeping 27 points more Clean Acc and nearly twice its PatFool Robust Acc. `STAR-GE` is the most robust under PatFool on all three datasets, and under FFT, it loses at most 22 points of PatFool Robust Acc to standard adaptation, against 24.5 to 48.9 points for PGDAT and MixAT. Thus, `STAR` alleviates the losses that come with AT under a small $\rho$ on white-box attacks.

### 4.4 Impact of the Learnable-Parameter Budget on Performance and Overhead

We trace Clean Acc and black-box Robust Acc across $\rho$ in Figure 4. $(i)$ Robustness has to be built inside the Encoder. Under LP, every defense that fits the task reaches nearly the same Robust Acc, since a frozen Encoder fixes both class tokens and leaves $\mathcal{L}_{\text{ali}}$ without gradient. $(ii)$ How the budget reaches the Encoder matters more than its size. Below FFT, `STAR` is most robust with LoRA, far above Adpt, which trains about seven times as many parameters, and with about 0.2%

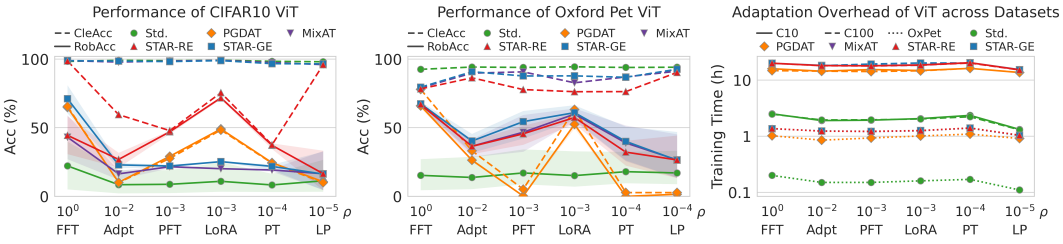


Figure 4: Effect of Learnable-Parameter Budget $\rho$. RobAcc is the mean over the six transfer attacks.

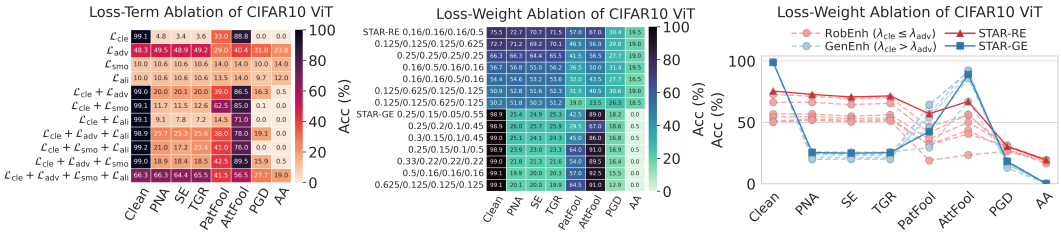


Figure 5: Loss Ablation with LoRA on CIFAR-10. Weights are listed as $\lambda_{\text{cle}}/\lambda_{\text{adv}}/\lambda_{\text{smo}}/\lambda_{\text{ali}}$.

of the parameters it comes close to its Robust Acc under FFT. (*iii*) A small $\rho$ does not make robust adaptation cheap. At every $\rho$, AT costs 5 to 12 times as much as standard adaptation (Figure 4, last panel), because the inner maximization back-propagates through the whole backbone, while STAR stays within 10% of MixAT and thus adds its robustness at no extra cost.

### 4.5 Ablation Studies

**Loss Terms.** We ablate the loss terms and weights with LoRA on CIFAR-10 (Figure 5). Partial objectives fall into the two failures of AT, as $\mathcal{L}_{\text{adv}}$ alone stays robust but loses about half of the Clean Acc, while adding one or two terms to $\mathcal{L}_{\text{cle}}$ keeps the task but not the robustness, just as MixAT does. Only all four terms together escape both failures, as even their untuned equal weighting beats $\mathcal{L}_{\text{adv}}$ alone in Clean Acc and in Robust Acc under both the transfer and the ViT-specific attacks.

**Loss Weights.** Among the four-term weightings, those with $\lambda_{\text{cle}} \leq \lambda_{\text{adv}}$ form a robust regime, the only one with $\ell_\infty$ robustness under AA, and the others an accurate regime near 99% Clean Acc. Within the robust regime, giving the largest weight to $\lambda_{\text{ali}}$ rather than to $\lambda_{\text{adv}}$ or $\lambda_{\text{smo}}$ raises both Clean Acc and transfer Robust Acc by 16 to 23 points, so the weak alignment constraint helps more than a stronger adversarial one. STAR-RE and STAR-GE make this choice in the two regimes, and STAR-RE is the most robust of all 14 weightings under the transfer attacks. Thus, for RQ ③, the robustness term that fits a small learnable subspace spreads the signal over weak constraints led by class-token alignment, with $\mathcal{L}_{\text{adv}}$ supplying the adversarial signal and $\mathcal{L}_{\text{cle}}$ anchoring the task.

## 5 Related Work

ViTs (Dosovitskiy et al., 2021) face token- and patch-level attacks on attention (Joshi et al., 2021; Mahmood et al., 2021; Fu et al., 2022; Lovisotto et al., 2022) and transfer attacks against them (Wei et al., 2022; Naseer et al., 2022; Zhang et al., 2023; Chen et al., 2023; Ming et al., 2024). Defenses train on AEs, as AT (Madry et al., 2018), logit pairing (Kannan et al., 2018), TRADES (Zhang et al., 2019), and MixAT (Xie et al., 2020) do, or exploit the architecture (Liu et al., 2023a; Lin et al., 2024; Jain & Dutta, 2024; Hu et al., 2026), all with a fully trainable model. TRADES approximates ablation with clean and smoothness terms but matches the outputs of the two paths. Under transfer learning, adversarially pretrained sources pass robustness downstream via fine-tuning (Shafahi et al., 2020; Salman et al., 2020), normalization statistics (Liu et al., 2023b), split objectives (Xu et al., 2024), linear probing (Hua et al., 2024), or distillation (Huang et al., 2025), yet are costly to pretrain. From a standard checkpoint, AT under PEFT (Houlsby et al., 2019; Li & Liang, 2021; Lester et al., 2021; Hu et al., 2022) yields little robustness (Hua et al., 2024; Huang et al., 2026), so recent work enlarges the learnable module (Yuan et al., 2025; Li et al., 2026; Lv et al., 2026), mostly raising $\rho$, or hardens PT alone (Eskandar et al., 2025). All of them keep the adversarial cross-entropy of AT, whereas STAR changes the objective itself with label-free alignment and smoothness terms and applies it unchanged to every adaptation method from LP to FFT and every parameter budget $\rho$.

## 6 Conclusion

We studied robust adaptation of ViTs under a learnable-parameter budget $\rho$, where standard adaptation is not robust, and AT cannot balance clean accuracy and robustness once $\rho$ shrinks. STAR spreads the robustness signal over weak constraints a small learnable subspace can satisfy, leading AT most under a small $\rho$ at a similar cost. Appendices B and E extend the related work and results.

## AI USE STATEMENT

Generative AI assisted with code preparation, table construction, figure creation, and manuscript polishing. We did not use generative AI to generate datasets, develop methods, formulate hypotheses, or design experiments. The authors reviewed and verified all AI-assisted content and take responsibility for the final paper.

## ETHICS STATEMENT

Work on adversarial robustness is dual-use, since the attacks that evaluate a defense could also be mounted against deployed models. We propose no new attack, and every attack we use is previously published and publicly implemented and is applied only to the models we train ourselves. Our experiments use the public CIFAR-10, CIFAR-100, and Oxford-IIIT Pet datasets and a publicly released ImageNet-1k checkpoint, none of which targets human subjects, and we report only aggregate accuracy. The work is intended to be defensive, as it shows practitioners who adapt a shared pretrained backbone that standard adaptation leaves the adapted model vulnerable to adversarial examples, and it offers a way to make the model robust.

## REPRODUCIBILITY STATEMENT

Section 3 specifies the `STAR` objective term by term, Algorithm 1 gives one full training epoch, and Appendix D extends the class-token mixup to a general coefficient and shows that the objective applies unchanged to every adaptation method. Section 4.1 and Appendix C give the full experimental setup, including the software versions and the seed (Appendix C.1), the backbone, surrogates, data splits, and preprocessing (Appendix C.2), the configuration and learnable-parameter budget of each adaptation method (Appendix C.3), the loss weights of both `STAR` modes, the inner maximization, and the learning rate of every run (Appendix C.4), and the settings of every attack (Appendix C.5). Appendix E lists the clean and robust accuracy of all trained models, three datasets by six adaptation methods by five training conditions, together with the training time of each run and all loss weightings of the ablation. All datasets and the pretrained ViTs are publicly available, and the code for training, attacks, and evaluation is available at `https://anonymous.4open.science/r/STAR-0D6B/`.

## REFERENCES


Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4489–4498, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607, 2020.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216, 2020.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

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.

Masih Eskandar, Tooba Imtiaz, Zifeng Wang, and Jennifer Dy. ADAPT to robustify prompt tuning vision transformers. *Transactions on Machine Learning Research*, 2025.

Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Celine Lin. Patch-Fool: Are vision transformers always robust against adversarial perturbations? In *International Conference on Learning Representations*, 2022.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*, 2024.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11936–11945, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jinyin Hu, Jiawei Zhou, Zhonghao Yang, Huadi Zheng, Daojing He, and Yu Li. AdaGuard: Location-aware adaptive defense against adversarial patch attacks in vision transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Andong Hua, Jindong Gu, Zhiyu Xue, Nicholas Carlini, Eric Wong, and Yao Qin. Initialization matters for adversarial transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Mengdie Huang, Yingjun Lin, Ninghui Li, Xiaofeng Chen, and Elisa Bertino. CARD: Robustness-preserving transfer learning for network intrusion detection via contrastive adversarial representation distillation. *IEEE Transactions on Dependable and Secure Computing*, 22(5):5134–5151, 2025.

Mengdie Huang, Yingjun Lin, Ninghui Li, and Elisa Bertino. TriXfer: Triple-cross transfer attack on vision transformer-adapted downstream classifiers. *IEEE Transactions on Dependable and Secure Computing*, 2026.

Samyak Jain and Tanima Dutta. Towards understanding and improving adversarial robustness of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24736–24745, 2024.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, 2022.

Ameya Joshi, Gauri Jagatap, and Chinmay Hegde. Adversarial token attacks on vision transformers. *arXiv preprint arXiv:2110.04337*, 2021.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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*, pp. 3045–3059, 2021.

Wenyun Li, Zheng Zhang, Dongmei Jiang, Yaowei Wang, and Xiangyuan Lan. Efficient adversarial training via criticality-aware fine-tuning. *arXiv preprint arXiv:2604.12780*, 2026.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 4582–4597, 2021.

Fudong Lin, Jiadong Lou, Xu Yuan, and Nian-Feng Tzeng. Towards robust vision transformer via masked adaptive ensemble. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1389–1399, 2024.

Liang Liu, Yanan Guo, Youtao Zhang, and Jun Yang. Understanding and defending patched-based adversarial attacks for vision transformer. In *International Conference on Machine Learning*, pp. 21631–21657, 2023a.

Ziquan Liu, Yi Xu, Xiangyang Ji, and Antoni B. Chan. TWINS: A fine-tuning framework for improved transferability of adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, and Jan Hendrik Metzen. Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15234–15243, 2022.

Kangtao Lv, Wenyan Fan, Huangsen Cao, Kainan Tu, Yihuai Xu, Zhimeng Zhang, Yang Li, Xin Ding, and Yongwei Wang. Hyper adversarial tuning for boosting adversarial robustness of pretrained large vision transformers. *Pattern Recognition*, 171:112158, 2026.

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.

Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7838–7847, 2021.

Di Ming, Peng Ren, Yunlong Wang, and Xin Feng. Boosting the transferability of adversarial attack on vision transformer with adaptive token tuning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 20887–20918, 2024.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. In *International Conference on Learning Representations*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust ImageNet models transfer better? In *Advances in Neural Information Processing Systems*, volume 33, pp. 3533–3545, 2020.

Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. In *International Conference on Learning Representations*, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2668–2676, 2022.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 819–828, 2020.

Xilie Xu, Jingfeng Zhang, and Mohan Kankanhalli. AutoLoRa: A parameter-free automated robust fine-tuning framework. In *International Conference on Learning Representations*, 2024.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Zheng Yuan, Jie Zhang, Shiguang Shan, and Xilin Chen. FullLoRA: Efficiently boosting the robustness of pretrained vision transformers. *IEEE Transactions on Image Processing*, 2025.

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*, pp. 7472–7482, 2019.

Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R. Lyu. Transferable adversarial attacks on vision transformers with token gradient regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16415–16424, 2023.

# A Overview and Notation

We outline the appendix in Table 2 and list the notation of the main paper in Table 3.

Table 2: Outline of the Appendix.


| Section | Title | Section | Title |
|---|---|---|---|
| A | Overview and Notation | D.1 | *Class-Token Mixup with a General Coefficient* |
| B | Related Work Extensions | D.2 | *Independence from the Adaptation Method* |
| C | Implementation and Evaluation Details | E | Experimental Result Extensions |
| C.1 | *Computing Platform* | E.1 | *Black-box Adversarial Robustness* |
| C.2 | *Models and Datasets* | E.2 | *White-box Adversarial Robustness* |
| C.3 | *Adaptation Methods* | E.3 | *Impact of the Learnable-Parameter Budget* |
| C.4 | *Training Configuration* | E.4 | *Training Overhead* |
| C.5 | *Attacks* | E.5 | *Ablation Studies* |
| D | Method Description Extension | E.6 | *Self-supervised Pretrained ViT Backbone* |


Table 3: Notation of the Main Paper.

| Group | Symbol | Meaning |
|---|---|---|
| Data | $x \in \mathcal{X} = [0,1]^{3\times H\times W}$ | input image of height $H$ and width $W$ |
| | $\mathcal{D} = \{(x, y)\},\ y \in \{1, \dots, C\}$ | target-domain dataset and ground-truth label |
| | $C,\ c$ | number of classes and class index |
| | $\mathcal{D}_{\text{te}},\ \mathcal{D}^*_{\text{te}}$ | clean and adversarial test sets |
| Model | $P,\ N = HW/P^2$ | patch size and number of non-overlapping patches |
| | $x_i^p,\ E_S \in \mathbb{R}^{(3P^2)\times D}$ | $i$-th flattened patch and its projection to $D$ dimensions |
| | $z_0^{cls},\ E_{\text{pos}}$ | learnable class token and positional embeddings |
| | $\mathbf{z}_0 \in \mathbb{R}^{(N+1)\times D}$ | input token sequence of the Encoder |
| | $L$ | number of Transformer blocks |
| | $h : \mathcal{X} \to \mathbb{R}^D$ | Encoder, returning the final normalized class token |
| | $g(z) = Wz + b$ | Classification Head with weight $W$ and bias $b$ |
| | $f_\theta = g \circ h$ | ViT model with parameters $\theta$ |
| | $z = h(x) = z_L^{cls}$ | class token [CLS] of a clean input |
| | $p = \sigma(g(z)),\ p_c$ | clean output distribution and its probability of class $c$ |
| | $\hat{y} = \arg\max_c p_c$ | predicted label |
| Adaptation | $\theta = (\theta_{\text{frz}}, \theta_{\text{ada}})$ | frozen and learnable parameters of the pretrained model |
| | $\rho = \lvert\theta_{\text{ada}}\rvert / \lvert\theta\rvert \in (0, 1]$ | learnable-parameter budget; $\lvert\cdot\rvert$ counts scalar parameters |
| | Opt | optimizer over $\theta_{\text{ada}}$ |
| Attack | $\delta,\ \tilde{f}$ | input perturbation and black-box surrogate model |
| | $\mathcal{A},\ \epsilon$ | training-time attack solver and its $\ell_\infty$ perturbation budget |
| | $\epsilon_s$ | step size of PGD |
| | $x^* = \mathcal{A}(x, y; \theta)$ | adversarial example (AE) of $x$ |
| | $z^* = h(x^*),\ p^* = \sigma(g(z^*))$ | class token and output distribution of the AE |
| Objective | $\bar{z} = \frac{1}{2}(z + z^*),\ \bar{p} = \sigma(g(\bar{z}))$ | interpolated class token and its output distribution |
| | $m_c(u) = g(u)_y - g(u)_c$ | margin of the true class $y$ over class $c$ at a class token $u$ |
| | $\mathcal{L}_{\text{CE}}(p, y) = -\log p_y$ | cross-entropy loss |
| | $\mathcal{R}(x, x^*, y; \theta)$ | robustness term added to the adaptation objective |
| | $\mathcal{L}_{\text{ali}} = 1 - \cos(z, z^*)$ | alignment loss at the representation level |
| | $\mathcal{L}_{\text{smo}} = D_{\text{KL}}(\bar{p} \,\|\, p)$ | smoothness loss at the distribution level |
| | $\mathcal{L}_{\text{cle}} = \mathcal{L}_{\text{CE}}(p, y)$ | clean cross-entropy, which maintains downstream accuracy |
| | $\mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{CE}}(p^*, y)$ | adversarial cross-entropy, requiring the true label on the AE |
| | $\boldsymbol{\lambda} = (\lambda_{\text{cle}}, \lambda_{\text{adv}}, \lambda_{\text{smo}}, \lambda_{\text{ali}})$ | non-negative weights of the four loss terms |
| | $\mathcal{L}_{\text{STAR}}(\boldsymbol{\lambda})$ | weighted sum of the four terms, minimized over $\theta_{\text{ada}}$ |
| Operator | $\sigma$ | softmax |
| | $\cos(\cdot,\cdot),\ \langle\cdot,\cdot\rangle$ | cosine similarity and Euclidean inner product |
| | $D_{\text{KL}}(\cdot \,\|\, \cdot)$ | Kullback-Leibler divergence |
| | $\odot,\ \propto$ | elementwise product and proportionality over classes $c$ |
| | $\mathrm{I}[\cdot]$ | indicator function |

# B Related Work Extensions

This section expands the related work of Section 5.

**Vision Transformers and downstream adaptation.** ViTs (Dosovitskiy et al., 2021) pretrained at scale serve as backbones of foundation models (Bommasani et al., 2021), and PEFT adapts them by training a small module on a frozen backbone, such as adapters (Houlsby et al., 2019), prefixes (Li & Liang, 2021), prompts (Lester et al., 2021; Jia et al., 2022), or low-rank updates (Hu et al., 2022). These methods differ mainly in how many parameters they train and where these parameters enter the forward pass, which the learnable-parameter budget $\rho$ summarizes.

**Adversarial attacks on ViTs.** Beyond the white-box PGD (Madry et al., 2018) and AutoAttack (Croce & Hein, 2020), ViT-specific attacks perturb individual tokens or patches (Joshi et al., 2021; Fu et al., 2022), draw the attention toward an adversarial patch (Lovisotto et al., 2022), or blend self-attention into the gradient (Mahmood et al., 2021). ViT-specific transfer attacks skip the attention gradients (Wei et al., 2022), ensemble the class tokens of intermediate blocks (Naseer et al., 2022), regularize token gradients (Zhang et al., 2023), or adapt surrogate ensembles and token gradients (Chen et al., 2023; Ming et al., 2024), and AEs crafted on a public backbone also transfer to its adapted classifiers (Huang et al., 2026), which motivates our black-box threat model.

**Robust training from scratch.** AT (Madry et al., 2018) trains on the AEs of an inner maximization, and later objectives pair the clean and adversarial logits (Kannan et al., 2018), add a KL divergence between the two outputs to trade accuracy for robustness (Zhang et al., 2019), or mix the clean and adversarial cross-entropies (Xie et al., 2020). These objectives regularize the output at the AE, whereas `STAR` aligns the directions of the clean and adversarial class tokens and matches the outputs at their midpoint. The alignment term resembles the cosine alignment of two views in contrastive learning (Chen et al., 2020), but it pairs an input with its AE and uses no negatives, and the smoothness term, like distillation (Hinton et al., 2015), pulls a prediction toward a distribution rather than a hard label. Architectural defenses instead regulate attention against patch attacks (Liu et al., 2023a; Hu et al., 2026), ensemble masked views (Lin et al., 2024), or revise the architecture (Jain & Dutta, 2024). These defenses change the model rather than the training objective and assume a fully trainable network, so they are complementary to `STAR`, which changes only the objective.

**Robust transfer from an adversarially pretrained backbone.** Shafahi et al. (2020) show that the robustness of an adversarially trained feature extractor carries over to new tasks when only its last layers are fine-tuned, and Salman et al. (2020) find that adversarially robust ImageNet models even transfer better. Robust fine-tuning keeps this robustness downstream by preserving the pretrained normalization statistics (Liu et al., 2023b), separating the clean and adversarial objectives through a LoRA branch (Xu et al., 2024), initializing the head by robust linear probing (Hua et al., 2024), or distilling adversarial representations (Huang et al., 2025). All of them start from an adversarially pretrained source, which is costly to obtain (Madry et al., 2018; Salman et al., 2020) and much rarer among public checkpoints than a standard one, whereas we start from a standard checkpoint.

**Robust adaptation from a standard backbone.** From a standard checkpoint, PEFT combined with AT yields little robustness (Hua et al., 2024; Huang et al., 2026). Recent work therefore enlarges the learnable module, with a heavier LoRA (Yuan et al., 2025), selected critical weights (Li et al., 2026), or hyper adversarial tuning (Lv et al., 2026), mostly raising $\rho$, or hardens PT alone (Eskandar et al., 2025). All of them keep the adversarial cross-entropy of AT and change the module around it. In contrast, `STAR` keeps every module unchanged and replaces the robustness term, and our study compares it with AT across six adaptation methods whose $\rho$ spans from under 0.01% to 100%.

# C Implementation and Evaluation Details

This section extends the experimental setup of Section 4.1 with the computing platform, models, datasets, adaptation methods, training configuration, and attacks behind every result of Section 4.

## C.1 Computing Platform

Every model is trained with seed 1, on a single NVIDIA A100 GPU (40 or 80 GB). The code is built on Python 3.10, PyTorch 2.3.0 with CUDA 12.1, torchvision 0.18.0, and timm 0.4.12.

### C.2 Models and Datasets

**Target backbone.** The backbone is the ViT-B/16 released in the timm library and pretrained on ImageNet-1k, with $L = 12$ Transformer blocks, a hidden size of $D = 768$, and about 86M parameters. It takes $224 \times 224$ inputs normalized with a mean and standard deviation of 0.5 per channel.

**Surrogate models.** The black-box attacker uses PiT-B (Heo et al., 2021) and DeiT-B (Touvron et al., 2021) classifiers of each downstream task. The PiT-B and DeiT-B surrogates are fine-tuned by us on CIFAR-10, CIFAR-100, and Oxford-IIIT Pet. Neither surrogate shares weights with the victims.

**Datasets.** CIFAR-10 and CIFAR-100 each have 10,000 test images, and Oxford-IIIT Pet has 3,680 training and 3,669 test images. A random 10% of each training split is held out for validation, which only monitors training, and the model after the last epoch is evaluated on the test set. Training images are resized to a shorter side of 256 pixels, randomly cropped and resized to $224 \times 224$, and randomly flipped horizontally, and test images are resized likewise and center-cropped to $224 \times 224$.

### C.3 Adaptation Methods

The downstream adaptation methods are configured as follows. Linear probing (LP) and full fine-tuning (FFT) (Yosinski et al., 2014) add no module. Prompt tuning (PT) (Lester et al., 2021; Jia et al., 2022) inserts 20 prompt tokens between [CLS] and the patch tokens, with zero positional embeddings. Low-rank adaptation (LoRA) (Hu et al., 2022) uses rank 4 and scaling 16, i.e., a residual multiplier of 4, with a zero-initialized up-projection. Prefix tuning (PFT) (Li & Liang, 2021) uses 10 key and 10 value prefix vectors of dimension $D$ per block, split across the 12 attention heads. Adapters (Adpt) (Houlsby et al., 2019) use a bottleneck of width 64 with a rectified linear unit, added residually to the output of every block. Without its original head, the pretrained backbone has 85.8M parameters, and each dataset attaches a new Classification Head of 7.7k to 76.9k parameters (Table 4). The learnable-parameter budgets therefore differ across datasets only through the head.

Table 4: Parameter Counts and Learnable-Parameter Budget $\rho$ (%).

| Backbone | Dataset | Head | LP | PT | LoRA | PFT | Adpt | FFT |
|---|---|---|---|---|---|---|---|---|
| ViT-B/16 (85.8M) | CIFAR-10 | 7.7k | 0.009 | 0.027 | 0.181 | 0.224 | 1.395 | 100 |
| | CIFAR-100 | 76.9k | 0.090 | 0.107 | 0.261 | 0.304 | 1.475 | 100 |
| | Oxford-IIIT Pet | 28.5k | 0.033 | 0.051 | 0.205 | 0.248 | 1.419 | 100 |

### C.4 Training Configuration

**Training conditions.** Standard adaptation minimizes the clean cross-entropy, PGD-based adversarial training (PGDAT) (Madry et al., 2018) minimizes the cross-entropy on AEs alone, and mixed adversarial training (MixAT) (Xie et al., 2020) averages the clean and adversarial cross-entropies. `STAR-RE` (robustness enhance-oriented) and `STAR-GE` (generalization enhance-oriented) set $(\lambda_{\text{cle}}, \lambda_{\text{adv}}, \lambda_{\text{smo}}, \lambda_{\text{ali}})$ in Eq. 7 to $(0.16, 0.16, 0.16, 0.50)$ and $(0.25, 0.15, 0.05, 0.55)$. We refer to PGDAT, MixAT, `STAR-RE`, and `STAR-GE` together as the four robust adaptation conditions. In both modes, the interpolation weight of $\bar{z}$ is fixed at $\frac{1}{2}$ (Section D treats a general weight), and the gradient of $\mathcal{L}_{\text{smo}}$ flows through both the clean and the interpolated class token.

**Inner maximization.** Every robust adaptation condition crafts its AEs with the training-time PGD of Section 4.1, implemented as Eq. 11. Starting from the clean input $x^{(0)} = x$, each of the $K$ iterations takes a signed-gradient step $\delta^{(k)}$ of size $\epsilon_s$ on the cross-entropy and projects the next iterate $x^{(k+1)}$ back into the $\epsilon$-ball around $x$, and the last iterate is the AE $x^*$.

$$\begin{aligned} x^{(0)} &= x, \qquad \delta^{(k)} = \epsilon_s \operatorname{sign}\Big(\nabla_{x'} \mathcal{L}_{\text{CE}}\big(\sigma(g(h(x'))), y\big)\Big|_{x'=x^{(k)}}\Big), \\ x^{(k+1)} &= \Pi_{\mathcal{X}}\Big(x + \Pi_{[-\epsilon,\epsilon]}\big(x^{(k)} + \delta^{(k)} - x\big)\Big), \qquad k = 0, \dots, K-1, \qquad x^* = x^{(K)}, \end{aligned} \tag{11}$$

where $\Pi_{\mathcal{S}}$ is the Euclidean projection onto a set $\mathcal{S}$, applied to the perturbation with $\mathcal{S} = [-\epsilon, \epsilon]$ and to the image with $\mathcal{S} = \mathcal{X}$, the valid pixel range. The iteration runs in pixel space before the input normalization, so $\epsilon$ and $\epsilon_s$ keep their usual meaning.

**Optimization.** Every model is trained for 20 epochs with AdamW (Loshchilov & Hutter, 2019), a weight decay of 0.05 on all learnable parameters, a mini-batch size of 64, and a cosine learning-rate

schedule that decays the initial rate of Table 5 to zero over training. Gradients are accumulated over two mini-batches, for an effective batch of 128, on CIFAR-100 and under the four robust adaptation conditions on CIFAR-10, while the other runs take one step per mini-batch. The AEs of every mini-batch are regenerated with the current parameters, and all computation runs in 32-bit precision.

Table 5: Initial Learning Rate (C10: CIFAR-10, C100: CIFAR-100, OxPet: Oxford-IIIT Pet).

| Method | Standard | PGDAT | MixAT | STAR-RE, STAR-GE |
|---|---|---|---|---|
| LP, PT, LoRA, PFT | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ |
| Adpt | $10^{-3}$ | $10^{-3}$ on C10 and OxPet; $10^{-4}$ on C100 | $10^{-3}$ on C10; $10^{-4}$ on C100 and OxPet | $10^{-4}$ |
| FFT | $10^{-4}$ | $10^{-4}$ | $10^{-4}$ | $10^{-4}$ |

### C.5 Attacks

*ViT-specific transfer attacks.* Pay No Attention (PNA) with PatchOut (Wei et al., 2022) skips attention gradients and perturbs random patches, Self-Ensemble (SE) with token refinement (Naseer et al., 2022) ensembles the per-block class tokens, and Token Gradient Regularization (TGR) (Zhang et al., 2023) damps extreme token gradients. All use the momentum iterative fast gradient sign method (MI-FGSM) (Dong et al., 2018) for 10 steps of size $\epsilon/10$ with momentum decay 1.0.

*ViT-specific white-box attacks.* Patch-Fool (PatFool) (Fu et al., 2022) and Attention-Fool (AttFool) (Lovisotto et al., 2022) perturb, without a norm bound, the single patch that receives the most attention in block 5 (counting from 0), outside the $\ell_\infty$ ball of the training attack. PatFool maximizes the cross-entropy plus 0.005 times the attention on the patch, and AttFool the pre-softmax attention logits toward it plus the cross-entropy, each with 250 Adam steps at a learning rate of 0.05.

*Conventional white-box attacks.* PGD (Madry et al., 2018) follows Section 4.1 and starts from a uniform random point of the $\epsilon$-ball, and the standard version of AutoAttack (AA) (Croce & Hein, 2020) runs at $\epsilon = 8/255$ on every trained CIFAR-10 model and on the loss-weight sweep.

## D Method Description Extension

This section supplements Section 3. It generalizes the class-token mixup, justifies the midpoint coefficient of STAR, and explains why the objective is independent of the adaptation method.

### D.1 Class-Token Mixup with a General Coefficient

The midpoint of Eq. 5 is the case $\beta = \frac{1}{2}$ of a convex combination $\bar{z}$ of the two class tokens with a mixing coefficient $\beta \in [0, 1]$, whose output distribution $\bar{p}$ is again obtained through the head $g$ and the softmax $\sigma$, as defined in Eq. 12.

$$\bar{z} = \beta\, z + (1-\beta)\, z^*, \qquad \bar{p} = \sigma(g(\bar{z})). \tag{12}$$

Since the head is affine, mixing the class tokens mixes the logits with the same coefficient, i.e., $g(\bar{z}) = \beta\, g(z) + (1-\beta)\, g(z^*)$. The softmax turns this linear relation between logits into a multiplicative one between output distributions, so a convex combination of the logits becomes a weighted geometric mean of the predictions, as derived in Eq. 13.

$$\begin{aligned} \bar{p}_c = \sigma(g(\bar{z}))_c = \frac{e^{\beta g(z)_c + (1-\beta) g(z^*)_c}}{\sum_{k=1}^{C} e^{\beta g(z)_k + (1-\beta) g(z^*)_k}} &\propto \left(e^{g(z)_c}\right)^{\beta} \left(e^{g(z^*)_c}\right)^{1-\beta} \\ \propto \left(\frac{e^{g(z)_c}}{\sum_{k=1}^{C} e^{g(z)_k}}\right)^{\beta} \left(\frac{e^{g(z^*)_c}}{\sum_{k=1}^{C} e^{g(z^*)_k}}\right)^{1-\beta} &= \sigma(g(z))_c^{\beta}\, \sigma(g(z^*))_c^{1-\beta} = p_c^{\beta}\, (p_c^*)^{1-\beta}, \end{aligned} \tag{13}$$

where $\propto$ hides the factors that do not depend on the class $c$. Normalizing over the classes yields Eq. 14, where $\odot$ is the elementwise product.

$$\bar{p} = \frac{p^{\beta} \odot (p^*)^{1-\beta}}{Z_\beta}, \qquad Z_\beta = \sum_{c=1}^{C} p_c^{\beta}\, (p_c^*)^{1-\beta}. \tag{14}$$

The weighted geometric mean never exceeds the weighted arithmetic mean, $p_c^{\beta}(p_c^*)^{1-\beta} \leq \beta p_c + (1-\beta)p_c^*$, so summing this inequality over the classes bounds $Z_\beta$ by 1, as derived in Eq. 15.

$$Z_\beta = \sum_{c=1}^{C} p_c^{\beta}\,(p_c^*)^{1-\beta} \;\leq\; \sum_{c=1}^{C}\big(\beta p_c + (1-\beta)p_c^*\big) = \beta + (1-\beta) = 1. \tag{15}$$

For $\beta \in (0,1)$, equality holds only when $p = p^*$, so $Z_\beta$ measures the overlap between the two output distributions. The mixed class token samples the segment between $g(z)$ and $g(z^*)$ at the logit level, but at the level of output distributions it follows the log-linear path $\log \bar{p} = \beta \log p + (1-\beta)\log p^* - \log Z_\beta$ rather than the arithmetic mixture $\beta p + (1-\beta)p^*$. By the same inequality, the distribution $\bar{p}$ discounts the confidence in every class on which the two paths disagree.

The coefficient $\beta$ sets how much weight each path carries, and STAR fixes it at $\beta = \frac{1}{2}$ as a structural choice rather than a tuned hyperparameter. First, the smoothness loss uses no ground-truth label, so neither prediction is a reliable target, as the clean one can be wrong on hard images and the adversarial one is pushed away from the label by construction. The midpoint is the only mixture that treats the two paths as interchangeable, so $\bar{p}$ is a consensus that keeps what they agree on and leaves the correction of each decision to the cross-entropies, which do see the label. Second, the midpoint passes equal shares of the gradient through $\bar{z}$ to the clean and adversarial class tokens, so the smoothness loss pulls the two toward each other, rather than one toward the other. A coefficient near either end would instead let one path dominate, fading the smoothness loss toward no regularization as $\beta$ approaches 1, or turning it into a consistency term at the AE as $\beta$ approaches 0.

### D.2 Independence from the Adaptation Method

Every term of Eq. 7 is computed from the forward pass and the label alone, including the two class tokens, their midpoint, and the head that reads them. No term refers to how $\theta_{\text{ada}}$ is parameterized (Figure 2, right), so the objective applies unchanged to each of the six adaptation methods of Section 2, from the Classification Head alone under LP to every weight under FFT. The objective and the adaptation method are therefore orthogonal design choices, the first fixing what the adaptation is asked to achieve and the second fixing the subspace in which it can be achieved, so the same objective applies under whatever learnable-parameter budget $\rho$ of Eq. 2 a deployment imposes.

## E Experimental Result Extensions

This section complements the experiments of Section 4. Wherever the main text shows a result only as a figure, it lists the underlying numbers in a table, and wherever the main text shows a result on only some of the datasets, it adds the remaining ones. Sections E.1 to E.5 follow Sections 4.2 to 4.5.

### E.1 Black-box Adversarial Robustness

This subsection extends Section 4.2. The CIFAR-100 results plotted in Figure 3 are listed in Table 6, and the CIFAR-10 and Oxford-IIIT Pet results listed in Table 1 are plotted in Figure 6. The Adpt rows are left out of Table 6, since all four robust adaptation conditions under Adpt on CIFAR-100 stay at the 1% chance rate, and this cell is excluded from the averages below.

Averaged over the 17 cells, STAR-RE gives the highest transfer robust accuracy (Robust Acc), 41.43% from PiT-B and 35.15% from DeiT-B, 8.90 and 10.86 points above the most robust baseline under each surrogate, and STAR-GE keeps the clean accuracy (Clean Acc) of MixAT, 91.95% against 91.61%, with 4.83 and 5.45 points more Robust Acc. The lead holds on every dataset, as the more robust mode of STAR exceeds the most robust baseline in Robust Acc averaged over the adaptation methods and the six transfer attacks by 9.25 points on CIFAR-10, 12.21 points on CIFAR-100, and 2.37 points on Oxford-IIIT Pet. It also holds for every attack, as STAR-GE exceeds MixAT by 4.59 to 5.73 points on average under each of the six transfer attacks, and STAR-RE exceeds PGDAT by 8.69 to 18.09 points while keeping 48.24 points more Clean Acc, 75.06% against 26.82%.

Beyond the two largest gains reported in Section 4.2, STAR-GE also exceeds MixAT by 7.72, 6.53, and 5.15 points under PFT on Oxford-IIIT Pet, Adpt on CIFAR-10, and LoRA on CIFAR-10. Under LoRA on CIFAR-100, it even has the highest Clean Acc of all five conditions, 93.23%, above the 92.76% of standard adaptation, so its robustness costs no accuracy there.

Table 6: Robustness (%) of Downstream ViTs against Black-Box Attacks on CIFAR-100.

| | | | CIFAR-100 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PiT-B Proxy | | | DeiT-B Proxy | | |
| TL | Method | $\rho$(%) | Clean | PNA | SE | TGR | PNA | SE | TGR |
| LP | Standard | 0.090 | **89.07** | 21.90 | 10.50 | 18.10 | 7.90 | 4.50 | 7.20 |
| | PGDAT | | 1.00 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 |
| | MixAT | | 84.25 | 28.00 | 14.80 | 27.40 | 12.30 | 6.30 | **10.70** |
| | STAR-RE | | 83.80 | **28.10** | **14.90** | **27.50** | **12.70** | **6.40** | **10.70** |
| | STAR-GE | | 85.96 | 27.60 | 13.60 | 26.60 | 11.40 | 5.80 | 9.70 |
| PT | Standard | 0.107 | **90.14** | **16.80** | 7.60 | **14.30** | 6.30 | 3.40 | 6.20 |
| | PGDAT | | 11.89 | 11.30 | **10.90** | 11.10 | **11.60** | **10.70** | **11.20** |
| | MixAT | | 89.68 | 3.90 | 3.80 | 3.60 | 3.80 | 3.70 | 3.90 |
| | STAR-RE | | 86.87 | 4.80 | 4.10 | 3.90 | 5.40 | 4.80 | 4.40 |
| | STAR-GE | | 88.39 | 6.90 | 6.90 | 6.00 | 6.70 | 5.30 | 6.10 |
| LoRA | Standard | 0.261 | 92.76 | 17.70 | 9.80 | 17.40 | 6.90 | 3.50 | 6.00 |
| | PGDAT | | 1.92 | 1.80 | 1.90 | 1.80 | 2.00 | 1.70 | 1.30 |
| | MixAT | | 92.61 | 1.10 | 1.10 | 1.10 | 1.60 | 1.50 | 1.20 |
| | STAR-RE | | 45.10 | **43.30** | **42.70** | **42.60** | **43.50** | **40.30** | **42.20** |
| | STAR-GE | | **93.23** | 34.80 | 29.60 | 31.50 | 36.30 | 25.30 | 30.10 |
| PFT | Standard | 0.304 | 91.59 | 16.00 | 7.90 | 16.50 | 6.80 | 3.60 | 5.70 |
| | PGDAT | | 20.52 | **20.50** | **20.20** | **20.30** | **20.20** | **19.60** | **20.00** |
| | MixAT | | **91.85** | 3.70 | 3.20 | 3.80 | 3.90 | 3.50 | 3.90 |
| | STAR-RE | | 88.30 | 14.70 | 14.40 | 14.50 | 14.30 | 12.30 | 12.70 |
| | STAR-GE | | 91.45 | 6.80 | 5.80 | 6.00 | 6.60 | 5.50 | 6.00 |
| FFT | Standard | 100 | 91.21 | 24.20 | 20.80 | 21.30 | 8.20 | 4.30 | 5.30 |
| | PGDAT | | 49.82 | 47.80 | 47.00 | 46.60 | 47.60 | 46.80 | 46.40 |
| | MixAT | | 90.93 | 52.40 | 49.60 | 52.30 | 28.90 | 25.80 | 28.90 |
| | STAR-RE | | 71.73 | **66.30** | **65.00** | **64.90** | **65.90** | **61.90** | **62.90** |
| | STAR-GE | | **91.60** | 48.50 | 45.00 | 48.50 | 27.90 | 21.30 | 25.40 |

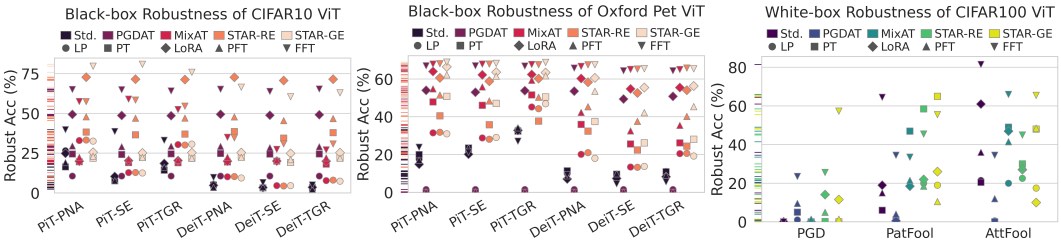


Figure 6: Robustness (%) of Downstream ViTs against Black-box Attacks on CIFAR-10 (Left) and Oxford-IIIT Pet (Middle) and against White-box Attacks on CIFAR-100 (Right).

### E.2 White-box Adversarial Robustness

This subsection extends Section 4.3. The white-box results plotted in Figure 3 for CIFAR-10 and Oxford-IIIT Pet are listed with CIFAR-100 in Table 7, and CIFAR-100 is plotted in Figure 6 (right).

Under the ViT-specific attacks, `STAR-GE` gives the highest PatFool Robust Acc averaged over the 17 cells, 43.88%, 4.26 points above MixAT and 8.50 points above standard adaptation, and it leads on every dataset, with 42.00%, 35.20%, and 53.00% on CIFAR-10, CIFAR-100, and Oxford-IIIT Pet against at most 38.08%, 27.80%, and 51.00% for the baselines. PGDAT, trained on the strongest $\ell_\infty$ attack, holds only 13.42% under PatFool and 17.09% under AttFool, so $\ell_\infty$ robustness alone does not carry over to patch-level attacks. The gap is largest under FFT, where AT is otherwise strongest, as `STAR-GE` loses only 1.40, 9.00, and 22.00 points of PatFool Robust Acc relative to standard adaptation on CIFAR-10, CIFAR-100, and Oxford-IIIT Pet, against 25.00 to 48.90 points for MixAT and 24.50 to 34.60 points for PGDAT. PGD is reported but not ranked, since three cells, MixAT under FFT on CIFAR-10 and CIFAR-100 and the collapsed PGDAT model under LoRA on CIFAR-100, give a PGD Robust Acc above their own Clean Acc, which no well-formed attack can produce and which indicates gradient masking. On CIFAR-10, AA confirms the $\ell_\infty$ robustness of `STAR-RE` beyond the LoRA cell of Section 4.3, with 11.00% under PFT at 20.25 points more Clean Acc than PGDAT, while MixAT and `STAR-GE`, which keep the Clean Acc of standard adaptation, retain at most 2.00% under AA.

Table 7: Robustness (%) of Downstream ViTs on White-Box Attacks (P-F: PatFool, A-F: AttFool).

| | | CIFAR-10 | | | | | CIFAR-100 | | | | | Oxford-IIIT Pet-37 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TL | Method | $\rho$(%) | Clean | PGD | AA | P-F | A-F | $\rho$(%) | Clean | PGD | P-F | A-F | $\rho$(%) | Clean | PGD | P-F | A-F |
| LP | Standard | 0.009 | **98.04** | 0.00 | 0.00 | 13.80 | 38.20 | 0.090 | **89.07** | 0.00 | 6.00 | 21.50 | 0.033 | **94.06** | 0.00 | 33.00 | 51.00 |
| | PGDAT | | 10.00 | 0.08 | 0.10 | 11.40 | 11.40 | | 1.00 | 1.00 | 0.50 | 0.50 | | 2.73 | 2.51 | 0.00 | 0.00 |
| | MixAT | | 95.96 | 0.00 | 0.00 | **18.00** | 43.00 | | 84.25 | 0.00 | 18.50 | 20.00 | | 91.20 | 0.00 | **81.50** | 80.00 |
| | STAR-RE | | 95.89 | 0.00 | 0.00 | **18.00** | **46.50** | | 83.80 | 0.01 | **20.00** | **22.50** | | 90.24 | 0.00 | 81.00 | **81.50** |
| | STAR-GE | | 96.50 | 0.00 | 0.00 | 13.00 | 25.50 | | 85.96 | 0.00 | 19.00 | 17.50 | | 92.59 | 0.00 | 78.00 | 76.00 |
| PT | Standard | 0.027 | **98.37** | 0.00 | 0.00 | 18.80 | 46.80 | 0.107 | **90.14** | 0.00 | 6.00 | 20.50 | 0.051 | **93.89** | 0.00 | 45.50 | 57.00 |
| | PGDAT | | 24.18 | 15.86 | 9.40 | 8.60 | 9.20 | | 11.89 | 5.01 | 1.00 | 0.00 | | 2.73 | 2.40 | 0.00 | 0.00 |
| | MixAT | | 97.42 | 0.00 | 0.00 | **71.50** | **81.50** | | 89.68 | 0.00 | 47.00 | **49.00** | | 86.97 | 0.00 | **62.50** | **77.50** |
| | STAR-RE | | 37.56 | 11.18 | 5.50 | 1.00 | 0.00 | | 86.87 | 0.17 | 58.50 | 30.00 | | 76.29 | 0.00 | 29.50 | 66.50 |
| | STAR-GE | | 96.65 | 0.00 | 0.00 | 65.00 | 70.50 | | 88.39 | 0.02 | **65.00** | 48.00 | | 86.73 | 0.00 | 47.50 | 72.50 |
| LoRA | Standard | 0.181 | **99.13** | 0.00 | 0.00 | 33.00 | 88.80 | 0.261 | 92.76 | 0.00 | 19.00 | **61.00** | 0.205 | **94.36** | 0.00 | 36.00 | 63.50 |
| | PGDAT | | 48.33 | 31.75 | **23.80** | 29.00 | 40.40 | | 1.92 | 2.71 | 0.50 | 3.00 | | 63.29 | 13.33 | 21.00 | 29.00 |
| | MixAT | | 99.04 | 16.31 | 0.50 | 39.00 | 86.50 | | 92.61 | 0.00 | 18.50 | 47.00 | | 82.72 | 8.20 | 45.50 | 61.50 |
| | STAR-RE | | 75.50 | 30.36 | 19.50 | **57.00** | 67.00 | | 45.10 | 14.09 | 22.00 | 27.00 | | 76.23 | 7.09 | 31.50 | 52.50 |
| | STAR-GE | | 98.87 | 18.22 | 0.00 | 42.50 | **89.00** | | **93.23** | 11.58 | **26.00** | 10.00 | | 87.76 | 7.96 | **58.50** | **66.50** |
| PFT | Standard | 0.224 | **98.80** | 0.00 | 0.00 | **29.20** | **70.80** | 0.304 | 91.59 | 0.00 | 15.00 | 36.00 | 0.248 | **93.92** | 0.00 | 31.50 | 58.50 |
| | PGDAT | | 27.63 | 20.55 | **19.20** | 11.00 | 26.00 | | 20.52 | 9.57 | 4.00 | 12.00 | | 5.21 | 0.05 | 0.00 | 0.00 |
| | MixAT | | 98.56 | 2.78 | 0.00 | 20.50 | 52.50 | | **91.85** | 1.37 | **21.50** | 41.50 | | 90.68 | 0.00 | **66.00** | **79.50** |
| | STAR-RE | | 47.88 | 19.29 | 11.00 | 17.00 | 28.50 | | 88.30 | 4.95 | 18.50 | 28.50 | | 77.79 | 0.00 | 32.00 | 61.50 |
| | STAR-GE | | 98.18 | 5.48 | 0.00 | 5.00 | 50.50 | | 91.45 | 1.37 | 10.50 | **48.50** | | 87.68 | 0.00 | 45.50 | 67.50 |
| Adpt | Standard | 1.395 | **99.06** | 0.00 | 0.00 | 28.20 | **90.80** | 1.475 | **92.54** | 0.00 | **23.00** | **65.00** | 1.419 | **94.19** | 0.00 | **67.00** | **81.50** |
| | PGDAT | | 10.00 | 10.00 | 9.40 | 9.40 | 9.40 | | 1.00 | 1.00 | 2.00 | 2.00 | | 33.17 | 7.17 | 1.50 | 5.00 |
| | MixAT | | 97.82 | 14.09 | 0.50 | **42.00** | 71.00 | | 1.00 | 1.00 | 2.00 | 2.00 | | 89.75 | 0.44 | 7.00 | 10.00 |
| | STAR-RE | | 59.51 | 13.65 | 0.50 | 6.50 | 12.50 | | 1.00 | 1.00 | 2.00 | 2.00 | | 86.35 | 0.00 | 43.00 | 59.00 |
| | STAR-GE | | 98.48 | 14.52 | 0.50 | 41.50 | 36.50 | | 1.00 | 1.00 | 2.00 | 2.00 | | 90.79 | 0.00 | 42.00 | 68.00 |
| FFT | Standard | 100 | 98.30 | 0.00 | 0.00 | **86.40** | **93.00** | 100 | 91.21 | 0.01 | **64.50** | **81.50** | 100 | **92.53** | 0.00 | **68.50** | **77.50** |
| | PGDAT | | 65.54 | 41.26 | **35.80** | 51.80 | 59.20 | | 49.82 | 23.50 | 34.50 | 34.50 | | 77.90 | 19.13 | 44.00 | 51.00 |
| | MixAT | | **98.84** | 99.22 | 0.00 | 37.50 | 73.00 | | 90.93 | 97.72 | 33.50 | 66.00 | | 79.10 | 16.43 | 43.50 | 47.50 |
| | STAR-RE | | 98.62 | 98.22 | 0.00 | 51.00 | 82.50 | | 71.73 | 25.56 | 45.50 | 45.00 | | 78.30 | 17.58 | 44.00 | 51.50 |
| | STAR-GE | | **98.84** | 67.53 | 2.00 | 85.00 | 90.00 | | **91.60** | 57.33 | 55.50 | 65.50 | | 79.48 | 15.73 | 46.50 | 51.00 |

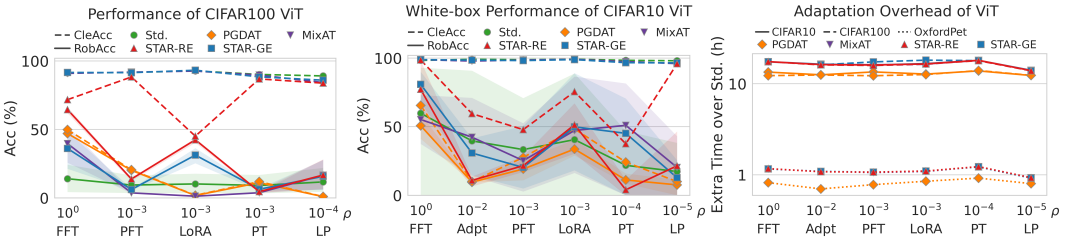


Figure 7: Effect of Learnable-Parameter Budget $\rho$ on Black-box Robustness on CIFAR-100 (Left), White-box Robustness on CIFAR-10 (Middle), and Extra Training Time (Right).

### E.3 Impact of the Learnable-Parameter Budget

This subsection extends Section 4.4. The Clean Acc and mean Robust Acc over six transfer attacks that Figure 4 plots against $\rho$ are listed for all three datasets in Table 8, and Figure 7 adds the CIFAR-100 (left), the plot for the white-box attacks on CIFAR-10 (middle), and training time (right).

On every dataset, MixAT and both modes of `STAR` end within 1 point of each other in RobAcc under LP, as a frozen Encoder leaves little to change, and below FFT `STAR` is most robust with LoRA. With LoRA, the stronger `STAR` mode reaches 71.75%, 42.43%, and 60.97% on CIFAR-10, CIFAR-100, and Oxford-IIIT Pet, against at most 26.88% and 40.33% under Adpt on CIFAR-10 and Oxford-IIIT Pet, and on CIFAR-10 `STAR-RE` exceeds the best FFT model, `STAR-GE` at 71.08%, with 0.181% of the parameters. On CIFAR-100, where MixAT falls below standard adaptation under PT, LoRA, and PFT, both modes reach 42.43% and 31.27% under LoRA against 1.27% for MixAT. The effect extends to white-box attacks (Figure 7, middle), averaged over PGD, PatFool, and AttFool without PGD above Clean Acc. Under LoRA on CIFAR-10, `STAR-RE` and `STAR-GE` lead with 51.45% and 49.91% against 47.27% for MixAT, the best baseline, with `STAR-GE` at 98.87% Clean Acc. Under FFT, their lead rests on PGD values not confirmed by AA (Table 7), as in Section 4.3.

Table 8: Performance (%) of Downstream ViTs across Learnable-Parameter Budgets.

| | | CIFAR-10 | | | CIFAR-100 | | | Oxford-IIIT Pet-37 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| TL | Method | $\rho$(%) | Clean | RobAcc | $\rho$(%) | Clean | RobAcc | $\rho$(%) | Clean | RobAcc |
| LP | Standard | 0.009 | **98.04** | 11.35 | 0.090 | **89.07** | 11.68 | 0.033 | **94.06** | 17.05 |
| | PGDAT | | 10.00 | 10.60 | | 1.00 | 0.90 | | 2.73 | 1.50 |
| | MixAT | | 95.96 | 16.47 | | 84.25 | 16.58 | | 91.20 | **26.58** |
| | STAR-RE | | 95.89 | **16.52** | | 83.80 | **16.72** | | 90.24 | 26.52 |
| | STAR-GE | | 96.50 | 16.22 | | 85.96 | 15.78 | | 92.59 | 26.32 |
| PT | Standard | 0.027 | **98.37** | 8.30 | 0.107 | **90.14** | 9.10 | 0.051 | **93.89** | 17.97 |
| | PGDAT | | 24.18 | 24.50 | | 11.89 | **11.13** | | 2.73 | 0.00 |
| | MixAT | | 97.42 | 19.22 | | 89.68 | 3.78 | | 86.97 | 38.90 |
| | STAR-RE | | 37.56 | **37.15** | | 86.87 | 4.57 | | 76.29 | 32.22 |
| | STAR-GE | | 96.65 | 21.83 | | 88.39 | 6.32 | | 86.73 | **40.03** |
| LoRA | Standard | 0.181 | **99.13** | 10.95 | 0.261 | 92.76 | 10.22 | 0.205 | **94.36** | 15.12 |
| | PGDAT | | 48.33 | 49.05 | | 1.92 | 1.75 | | 63.29 | 52.43 |
| | MixAT | | 99.04 | 20.13 | | 92.61 | 1.27 | | 82.72 | 59.87 |
| | STAR-RE | | 75.50 | **71.75** | | 45.10 | **42.43** | | 76.23 | 57.43 |
| | STAR-GE | | 98.87 | 25.28 | | **93.23** | 31.27 | | 87.76 | **60.97** |
| PFT | Standard | 0.224 | **98.80** | 8.77 | 0.304 | 91.59 | 9.42 | 0.248 | **93.92** | 17.05 |
| | PGDAT | | 27.63 | 29.15 | | 20.52 | **20.13** | | 5.21 | 0.22 |
| | MixAT | | 98.56 | 21.58 | | **91.85** | 3.67 | | 90.68 | 46.70 |
| | STAR-RE | | 47.88 | **47.02** | | 88.30 | 13.82 | | 77.79 | 45.57 |
| | STAR-GE | | 98.18 | 22.08 | | 91.45 | 6.12 | | 87.68 | **54.42** |
| Adpt | Standard | 1.395 | **99.06** | 8.48 | 1.475 | **92.54** | **8.73** | 1.419 | **94.19** | 13.77 |
| | PGDAT | | 10.00 | 10.20 | | 1.00 | 1.40 | | 33.17 | 26.37 |
| | MixAT | | 97.82 | 16.38 | | 1.00 | 1.40 | | 89.75 | 36.52 |
| | STAR-RE | | 59.51 | **26.88** | | 1.00 | 1.40 | | 86.35 | 36.42 |
| | STAR-GE | | 98.48 | 22.92 | | 1.00 | 1.40 | | 90.79 | **40.33** |
| FFT | Standard | 100 | 98.30 | 22.18 | 100 | 91.21 | 14.02 | 100 | **92.53** | 15.30 |
| | PGDAT | | 65.54 | 64.78 | | 49.82 | 47.03 | | 77.90 | 65.92 |
| | MixAT | | **98.84** | 43.08 | | 90.93 | 39.65 | | 79.10 | 66.72 |
| | STAR-RE | | 98.62 | 44.45 | | 71.73 | **64.48** | | 78.30 | 67.03 |
| | STAR-GE | | **98.84** | **71.08** | | **91.60** | 36.10 | | 79.48 | **67.42** |

Table 9: Training Time per Run (h) and Overhead Relative to Standard Adaptation.

| | | Training time (h) | | | | | Overhead vs. standard adaptation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Method | Standard | PGDAT | MixAT | STAR-RE | STAR-GE | PGDAT | MixAT | STAR-RE | STAR-GE |
| CIFAR-10 | LP | 1.32 | 13.65 | 15.37 | 15.26 | 15.37 | 10.3× | 11.6× | 11.6× | 11.6× |
| | PT | 2.35 | 16.18 | 20.37 | 20.44 | 20.43 | 6.9× | 8.7× | 8.7× | 8.7× |
| | LoRA | 2.06 | 14.84 | 18.53 | 18.59 | 18.71 | 7.2× | 9.0× | 9.0× | 9.1× |
| | PFT | 1.93 | 15.42 | 18.09 | 17.98 | 18.01 | 8.0× | 9.4× | 9.3× | 9.3× |
| | Adpt | 1.90 | 14.44 | 18.21 | 18.14 | 18.31 | 7.6× | 9.6× | 9.5× | 9.6× |
| | FFT | 2.51 | 15.94 | 19.92 | 20.01 | 19.97 | 6.4× | 7.9× | 8.0× | 8.0× |
| CIFAR-100 | LP | 1.30 | 13.68 | 15.28 | 15.18 | 15.16 | 10.5× | 11.8× | 11.7× | 11.7× |
| | PT | 2.22 | 16.19 | 20.31 | 20.15 | 20.17 | 7.3× | 9.1× | 9.1× | 9.1× |
| | LoRA | 2.03 | 14.65 | 18.41 | 18.50 | 20.19 | 7.2× | 9.1× | 9.1× | 9.9× |
| | PFT | 1.96 | 14.21 | 17.85 | 17.86 | 19.34 | 7.3× | 9.1× | 9.1× | 9.9× |
| | Adpt | 1.95 | 14.43 | 18.10 | 18.06 | 18.10 | 7.4× | 9.3× | 9.3× | 9.3× |
| | FFT | 2.47 | 14.75 | 19.96 | 19.96 | 19.93 | 6.0× | 8.1× | 8.1× | 8.1× |
| Oxford-IIIT Pet | LP | 0.11 | 0.91 | 1.04 | 1.04 | 1.03 | 8.3× | 9.5× | 9.5× | 9.4× |
| | PT | 0.17 | 1.09 | 1.40 | 1.40 | 1.39 | 6.4× | 8.2× | 8.2× | 8.2× |
| | LoRA | 0.16 | 1.01 | 1.25 | 1.26 | 1.26 | 6.3× | 7.8× | 7.9× | 7.9× |
| | PFT | 0.15 | 0.93 | 1.21 | 1.22 | 1.22 | 6.2× | 8.1× | 8.1× | 8.1× |
| | Adpt | 0.15 | 0.85 | 1.24 | 1.23 | 1.24 | 5.7× | 8.3× | 8.2× | 8.3× |
| | FFT | 0.20 | 1.02 | 1.36 | 1.37 | 1.36 | 5.1× | 6.8× | 6.9× | 6.8× |

### E.4 Training Overhead

This subsection extends the overhead analysis of Section 4.4 with the training time of every run (Table 9) and the overhead against $\rho$ (Figure 7, right). Standard adaptation takes 1.3 to 2.5 h on CIFAR-10 and CIFAR-100, whereas the robust adaptation conditions vary far less across adaptation methods, e.g., from 15.3 h under LP to 20.4 h under PT for MixAT and STAR on CIFAR-10, since the 10-step inner maximization of Eq. 11 costs about the same at every $\rho$ and dominates every step. The overhead is therefore largest under LP, 8.3 to 11.8 times, whose standard adaptation back-propagates

Table 10: Ablation on the Loss Terms of STAR, CIFAR-10 with LoRA (%).

| Term | Loss weight | | | | | Transfer (DeiT-B) | | | White-box | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Num | $\lambda_{\text{cle}}$ | $\lambda_{\text{adv}}$ | $\lambda_{\text{smo}}$ | $\lambda_{\text{ali}}$ | Clean | PNA | SE | TGR | PGD | AA | PatFool | AttFool |
| 1 | 1 | 0 | 0 | 0 | 99.13 | 4.80 | 3.40 | 3.60 | 0.00 | 0.00 | 33.00 | 88.80 |
| | 0 | 1 | 0 | 0 | 48.33 | 49.50 | 48.90 | 49.20 | **31.75** | **23.80** | 29.00 | 40.40 |
| | 0 | 0 | 1 | 0 | 10.00 | 10.60 | 10.60 | 10.60 | 10.00 | 14.00 | 14.00 | 14.00 |
| | 0 | 0 | 0 | 1 | 10.00 | 10.60 | 10.60 | 10.60 | 9.66 | 12.00 | 13.50 | 14.00 |
| 2 | 0.5 | 0.5 | 0 | 0 | 99.04 | 20.00 | 20.10 | 20.00 | 16.31 | 0.50 | 39.00 | 86.50 |
| | 0.5 | 0 | 0.5 | 0 | 99.12 | 11.70 | 11.50 | 12.60 | 0.10 | 0.00 | **62.50** | 85.00 |
| | 0.5 | 0 | 0 | 0.5 | 99.13 | 9.10 | 7.80 | 7.20 | 0.00 | 0.00 | 14.50 | 71.00 |
| 3 | 0.33 | 0.33 | 0 | 0.33 | 98.93 | 25.70 | 25.30 | 25.60 | 19.11 | 0.00 | 38.00 | 78.00 |
| | 0.33 | 0 | 0.33 | 0.33 | **99.16** | 21.00 | 17.20 | 23.40 | 0.04 | 0.00 | 41.00 | 76.00 |
| | 0.33 | 0.33 | 0.33 | 0 | 99.02 | 18.90 | 18.40 | 18.50 | 15.93 | 0.50 | 42.50 | **89.50** |
| 4 | 0.25 | 0.25 | 0.25 | 0.25 | 66.30 | **66.30** | **64.40** | **65.50** | 27.66 | 19.00 | 41.50 | 56.50 |

Table 11: Ablation on the Loss Weights of STAR with All Four Terms, CIFAR-10 with LoRA (%).

| | Loss weight | | | | | Transfer (DeiT-B) | | | White-box | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $\lambda_{\text{cle}}$ | $\lambda_{\text{adv}}$ | $\lambda_{\text{smo}}$ | $\lambda_{\text{ali}}$ | Clean | PNA | SE | TGR | PGD | AA | PatFool | AttFool |
| *Robustness Enhance-oriented,* $\lambda_{\text{cle}} \leq \lambda_{\text{adv}}$ | | | | | | | | | | | | |
| STAR-RE | 0.16 | 0.16 | 0.16 | 0.5 | **75.50** | **72.70** | **70.70** | **71.50** | 30.36 | **19.50** | **57.00** | **67.00** |
| | 0.125 | 0.125 | 0.125 | 0.625 | 72.74 | 71.20 | 69.20 | 70.10 | 29.80 | 19.00 | 46.50 | 56.00 |
| | 0.25 | 0.25 | 0.25 | 0.25 | 66.30 | 66.30 | 64.40 | 65.50 | 27.66 | 19.00 | 41.50 | 56.50 |
| | 0.16 | 0.5 | 0.16 | 0.16 | 56.74 | 56.80 | 55.00 | 56.20 | **31.37** | **19.50** | 36.50 | 50.00 |
| | 0.16 | 0.16 | 0.5 | 0.16 | 54.39 | 54.60 | 53.20 | 53.60 | 27.66 | 16.50 | 32.00 | 43.50 |
| | 0.125 | 0.625 | 0.125 | 0.125 | 50.88 | 52.80 | 51.60 | 52.30 | 30.63 | 19.00 | 31.50 | 40.50 |
| | 0.125 | 0.125 | 0.625 | 0.125 | 50.19 | 51.80 | 50.30 | 51.20 | 26.26 | 18.50 | 19.00 | 23.50 |
| *Generalization Enhance-oriented,* $\lambda_{\text{cle}} > \lambda_{\text{adv}}$ | | | | | | | | | | | | |
| STAR-GE | 0.25 | 0.15 | 0.05 | 0.55 | 98.87 | 25.40 | 24.90 | 25.30 | 18.22 | 0.00 | 42.50 | 89.00 |
| | 0.25 | 0.2 | 0.1 | 0.45 | 98.88 | **26.00** | **25.70** | **25.90** | **18.57** | 0.50 | 29.50 | 67.00 |
| | 0.3 | 0.15 | 0.1 | 0.45 | 99.02 | 25.10 | 24.10 | 24.30 | 16.80 | 0.50 | 45.00 | 86.00 |
| | 0.25 | 0.15 | 0.1 | 0.5 | 98.89 | 23.90 | 23.00 | 23.30 | 16.91 | 0.50 | 64.00 | 91.00 |
| | 0.33 | 0.22 | 0.22 | 0.22 | 99.01 | 21.80 | 21.30 | 21.60 | 16.40 | 0.00 | 54.00 | 89.50 |
| | 0.5 | 0.16 | 0.16 | 0.16 | **99.11** | 19.90 | 20.00 | 20.30 | 15.50 | 0.00 | 57.00 | **92.50** |
| | 0.625 | 0.125 | 0.125 | 0.125 | **99.11** | 20.10 | 20.00 | 19.90 | 12.88 | 0.00 | **64.50** | 91.00 |

through the head alone, and smallest under FFT, 5.1 to 8.1 times. PGDAT takes 69% to 90% of the time of MixAT, since its outer step runs the Encoder only on $x^*$, whereas MixAT and STAR run it on both $x$ and $x^*$. STAR adds only a head pass on $\bar{z}$ and two loss terms, both negligible next to an Encoder pass, so both modes take the time of MixAT to within 1%. The only larger gaps, 10% and 8% for STAR−GE with LoRA and PFT on CIFAR-100, we attribute to the assigned GPU nodes, as STAR−RE, which differs only in the loss weights, matches MixAT in both cells.

### E.5 Ablation Studies

This subsection extends Section 4.5, and Tables 10 and 11 list the 24 weightings of Eq. 7 drawn in Figure 5, the partial objectives in the first and all 14 four-term weightings in the second. The two label-free terms cannot fit the task alone, as $\mathcal{L}_{\text{smo}}$ or $\mathcal{L}_{\text{ali}}$ by itself stays at the 10% chance rate, but on top of the supervised terms they add complementary robustness, as $\mathcal{L}_{\text{ali}}$ added to $\mathcal{L}_{\text{cle}}$ and $\mathcal{L}_{\text{adv}}$ raises the transfer Robust Acc averaged over PNA, SE, and TGR from 20.03% to 25.53%, and $\mathcal{L}_{\text{smo}}$ added to $\mathcal{L}_{\text{cle}}$ raises the PatFool Robust Acc from 33.00% to 62.50%. Still, every partial objective that keeps about 99% Clean Acc stays at or below 0.50% under AA, whereas the equal four-term weighting reaches 19.00%, close to the 23.80% of $\mathcal{L}_{\text{adv}}$ alone at 17.97 points more Clean Acc. In the robust regime, beyond its transfer lead, STAR−RE also has the highest Clean Acc, PatFool, and AttFool Robust Acc, 2.76, 10.50, and 10.50 points above the next best weighting, whereas a larger $\lambda_{\text{adv}}$ of 0.5 or 0.625 gains at most 1.01 points under PGD and none under AA at a cost of at least 18.76 points of Clean Acc. A larger $\lambda_{\text{ali}}$ of 0.625 falls below STAR−RE on every metric, so the supervised terms still need their share, and $\lambda_{\text{ali}} = 0.5$ is the peak of the tested range, not its end. In the accurate regime, where Clean Acc varies by only 0.24 points, STAR−GE comes within 0.67 points of the most transfer-robust weighting with 13.00 and 22.00 points more under PatFool and AttFool, and against $\mathcal{L}_{\text{cle}}$ and $\mathcal{L}_{\text{adv}}$ alone at equal weights, it adds 5.17 points of transfer Robust Acc and 1.91, 3.50, and 2.50 points under PGD, PatFool, and AttFool at 0.17 points less Clean Acc.

### E.6 Self-Supervised Pretrained ViT Backbone

Every model of Section 4 adapts a ViT-B/16 pretrained on ImageNet-1k by supervised classification, whose features are already shaped by class labels. This subsection turns to a different pretraining paradigm, DINOv2 (Oquab et al., 2024), a ViT-B/14 pretrained by self-distillation on LVD-142M without any labels. Its 30 models cover the six adaptation methods and five training conditions of Section 4.1, and show whether the advantages of STAR carry over beyond supervised pretraining.

#### E.6.1 Black-box Adversarial Robustness

**Setup.** The six ViT-specific transfer attacks of Section C.5, PNA, SE, and TGR crafted on the PiT-B and DeiT-B proxies, reuse the AEs on which the ViT-B/16 models were scored, so both backbones face identical inputs (Table 12 and Figure 8, left).

**Results.** $(i)$ For RQ ①, standard adaptation fits the task at every $\rho$, with at least 98.24% Clean Acc, but its Robust Acc averaged over the six transfer attacks stays between 13.08% and 20.95% under the PEFT methods and reaches only 39.25% under FFT, so robustness must come from the training objective, not the adaptation method. $(ii)$ For RQ ②, AT is at its best under FFT, where MixAT reaches 81.38% averaged Robust Acc, but it cannot retain both Clean Acc and robustness once $\rho$ shrinks. PGDAT stays at the 10% chance rate in Clean Acc under LP, PT, LoRA, and PFT, and MixAT keeps the task but falls to between 10.63% and 26.28% averaged Robust Acc under the PEFT methods, even below standard adaptation under LoRA, PFT, and Adpt. $(iii)$ For RQ ③, STAR achieves both. Averaged over the six adaptation methods, STAR-GE gives the highest transfer Robust Acc from both proxies, 31.23% and 33.87%, 4.95 and 3.04 points above MixAT, the most robust baseline, at a higher Clean Acc, 94.48% against 92.10%. Under PFT, it exceeds MixAT under all six attacks, by 8.10 to 10.00 points, at nearly the same Clean Acc, 98.06% against 98.93%, and under PT it raises the averaged Robust Acc from 26.28% to 34.67%. Under FFT, STAR-GE is the most robust model, exceeding MixAT by 4.30 to 4.80 points under every attack, with 85.95% averaged Robust Acc at 98.23% Clean Acc, 12.40 points above the Clean Acc of MixAT.

Table 12: Robustness (%) of Downstream DINOv2 against Black-Box Attacks.

| | | | CIFAR-10 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PiT-B Proxy | | | DeiT-B Proxy | | |
| TL | Method | $\rho$(%) | Clean | PNA | SE | TGR | PNA | SE | TGR |
| LP | Standard | 0.009 | **98.24** | 14.10 | 6.60 | 12.40 | 23.10 | 19.00 | 21.90 |
| | PGDAT | | 10.00 | 10.20 | 10.20 | 10.20 | 10.20 | 10.20 | 10.20 |
| | MixAT | | 96.41 | **19.70** | 8.10 | 17.00 | 28.10 | 23.80 | **27.20** |
| | STAR-RE | | 96.44 | 19.40 | 8.10 | **17.10** | 27.70 | **24.10** | 27.10 |
| | STAR-GE | | 96.82 | 18.90 | 8.30 | 16.70 | **28.20** | 23.40 | 27.00 |
| PT | Standard | 0.027 | **98.76** | 12.10 | 5.50 | 11.40 | 19.90 | 19.80 | 19.30 |
| | PGDAT | | 10.00 | 8.60 | 8.60 | 8.60 | 8.60 | 8.60 | 8.60 |
| | MixAT | | 98.64 | 21.70 | 12.80 | 20.70 | **40.90** | 31.60 | 30.00 |
| | STAR-RE | | 20.63 | 19.00 | 17.40 | 17.70 | 18.70 | 18.40 | 18.40 |
| | STAR-GE | | 95.67 | **35.70** | **31.30** | **31.10** | 38.90 | **33.40** | **37.60** |
| LoRA | Standard | 0.181 | 99.28 | 10.20 | 5.00 | 9.60 | **18.60** | **18.00** | **17.10** |
| | PGDAT | | 10.48 | 15.20 | 15.70 | 18.00 | 16.50 | 15.80 | 16.80 |
| | MixAT | | **99.35** | 9.80 | **10.70** | **10.70** | 10.90 | 11.10 | 10.60 |
| | STAR-RE | | 44.38 | **10.70** | 10.30 | 10.30 | 12.30 | 11.50 | 11.00 |
| | STAR-GE | | 99.02 | 10.20 | 10.40 | 10.30 | 11.00 | 10.90 | 10.30 |
| PFT | Standard | 0.224 | **99.04** | 13.10 | 7.10 | 11.70 | 18.80 | 19.00 | 16.70 |
| | PGDAT | | 10.00 | 10.30 | 10.30 | 10.30 | 10.30 | 10.30 | 10.30 |
| | MixAT | | 98.93 | 11.20 | 12.40 | 11.60 | 10.80 | 11.30 | 12.10 |
| | STAR-RE | | 97.07 | 19.40 | 19.60 | **19.90** | 17.50 | 15.20 | 18.00 |
| | STAR-GE | | 98.06 | **21.00** | **21.20** | 19.70 | **20.10** | **20.00** | **22.10** |
| Adpt | Standard | 1.397 | **99.25** | 15.20 | 10.90 | 9.80 | **32.50** | **30.40** | **26.90** |
| | PGDAT | | 19.24 | 20.50 | 20.40 | 20.50 | 20.40 | 20.40 | 19.80 |
| | MixAT | | 73.43 | 20.90 | 20.80 | 20.50 | 20.90 | 21.20 | 20.60 |
| | STAR-RE | | 56.96 | **25.50** | **24.90** | **25.10** | 25.80 | 25.80 | 25.40 |
| | STAR-GE | | 79.07 | 23.60 | 22.80 | 22.80 | 23.50 | 22.80 | 22.80 |
| FFT | Standard | 100 | **98.68** | 34.10 | 28.60 | 27.80 | 50.70 | 45.70 | 48.60 |
| | PGDAT | | 66.63 | 65.30 | 65.40 | 64.90 | 64.80 | 64.70 | 64.50 |
| | MixAT | | 85.83 | 82.00 | 81.70 | 80.80 | 82.10 | 80.80 | 80.90 |
| | STAR-RE | | 88.53 | 85.70 | 85.00 | 84.20 | 85.80 | 84.50 | 85.10 |
| | STAR-GE | | 98.23 | **86.60** | **86.40** | **85.10** | **86.50** | **85.60** | **85.50** |

Table 13: Robustness (%) of Downstream DINOv2 against White-Box Attacks.

| TL | Method | $\rho$(%) | CIFAR-10 Clean | CIFAR-10 PGD | CIFAR-10 PatFool | CIFAR-10 AttFool |
|---|---|---|---|---|---|---|
| LP | Standard | 0.009 | **98.24** | 0.00 | 3.00 | 45.00 |
| | PGDAT | | 10.00 | 0.00 | 3.00 | 9.00 |
| | MixAT | | 96.41 | 0.00 | 4.50 | 40.00 |
| | STAR-RE | | 96.44 | 0.00 | 3.50 | **45.50** |
| | STAR-GE | | 96.82 | 0.00 | 5.50 | 30.50 |
| PT | Standard | 0.027 | **98.76** | 0.00 | 13.50 | **62.50** |
| | PGDAT | | 10.00 | 10.00 | 9.00 | 9.00 |
| | MixAT | | 98.64 | 0.29 | **16.50** | 14.00 |
| | STAR-RE | | 20.63 | 0.47 | 0.00 | 14.00 |
| | STAR-GE | | 95.67 | 0.00 | 1.00 | 30.00 |
| LoRA | Standard | 0.181 | 99.28 | 0.00 | 8.00 | **55.50** |
| | PGDAT | | 10.48 | 6.77 | 12.00 | 8.50 |
| | MixAT | | **99.35** | 0.70 | 6.50 | 14.50 |
| | STAR-RE | | 44.38 | 9.98 | 3.00 | 9.50 |
| | STAR-GE | | 99.02 | 9.60 | **18.00** | 11.50 |
| PFT | Standard | 0.224 | **99.04** | 0.00 | 7.00 | **55.50** |
| | PGDAT | | 10.00 | 5.64 | 0.00 | 9.00 |
| | MixAT | | 98.93 | 0.00 | 10.00 | 41.00 |
| | STAR-RE | | 97.07 | 1.60 | 3.00 | 14.50 |
| | STAR-GE | | 98.06 | 0.80 | 0.00 | 43.50 |
| Adpt | Standard | 1.397 | **99.25** | 0.00 | **42.50** | **42.50** |
| | PGDAT | | 19.24 | **14.65** | 24.00 | 27.00 |
| | MixAT | | 73.43 | 13.45 | 4.50 | 22.50 |
| | STAR-RE | | 56.96 | 13.77 | 11.50 | 29.00 |
| | STAR-GE | | 79.07 | 10.88 | 11.50 | 26.50 |
| FFT | Standard | 100 | **98.68** | 0.00 | 80.50 | **95.00** |
| | PGDAT | | 66.63 | 43.78 | 46.00 | 62.50 |
| | MixAT | | 85.83 | 44.89 | 68.50 | 74.00 |
| | STAR-RE | | 88.53 | 48.73 | 69.50 | 76.00 |
| | STAR-GE | | 98.23 | **51.46** | **86.00** | 85.50 |

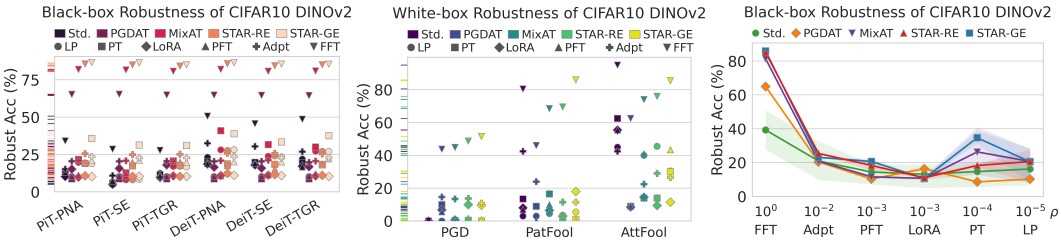


Figure 8: Robustness (%) of Downstream DINOv2 against Black-box (Left) and White-box (Middle) Attacks, and across Learnable-Parameter Budgets $\rho$ (Right). In the right panel, RobAcc is the mean over the six transfer attacks.

### E.6.2 White-box Adversarial Robustness

**Setup.** The white-box attacks of Section C.5 are PGD with 20 steps at $\epsilon = 8/255$ and the ViT-specific PatFool and AttFool on the most-attended patch of block 5 (Table 13 and Figure 8, middle).

**Results.** $(i)$ For RQ ①, standard adaptation has no white-box $\ell_\infty$ robustness, as PGD drives it to 0.00% Robust Acc under every adaptation method. $(ii)$ For RQ ②, AT under a small $\rho$ again cannot retain both, as PGDAT reaches at most 14.65% PGD Robust Acc under the PEFT methods only at a collapsed Clean Acc of at most 19.24%, while MixAT keeps the task but holds at most 0.70% under PGD with LP, PT, LoRA, and PFT. Even under FFT, PGDAT and MixAT fall below standard adaptation under PatFool, 46.00% and 68.50% against 80.50%, so their PGD robustness does not carry over to ViT-specific attacks. $(iii)$ For RQ ③, `STAR` attains white-box robustness while still fitting the task. Under FFT, `STAR-GE` has the highest PGD and PatFool Robust Acc of all five conditions, 51.46% and 86.00%, and exceeds PGDAT and MixAT under AttFool, 85.50% against at most 74.00%, at 98.23% Clean Acc, while `STAR-RE` also exceeds both AT baselines under all three attacks at a higher Clean Acc. Under LoRA, `STAR-GE` keeps the Clean Acc of MixAT, 99.02% against 99.35%, and raises PGD Robust Acc from 0.70% to 9.60% and PatFool Robust Acc from 6.50% to 18.00%, the highest under LoRA. Under Adpt, `STAR-RE` nearly matches the PGD Robust Acc of PGDAT, 13.77% against 14.65%, with 37.72 points more Clean Acc.

Table 14: Performance (%) of Downstream DINOv2 across Learnable-Parameter Budgets.

| TL | Method | CIFAR-10 $\rho$(%) | CIFAR-10 Clean | CIFAR-10 RobAcc |
|---|---|---|---|---|
| LP | Standard | 0.009 | **98.24** | 16.18 |
| | PGDAT | | 10.00 | 10.20 |
| | MixAT | | 96.41 | **20.65** |
| | STAR-RE | | 96.44 | 20.58 |
| | STAR-GE | | 96.82 | 20.42 |
| PT | Standard | 0.027 | **98.76** | 14.67 |
| | PGDAT | | 10.00 | 8.60 |
| | MixAT | | 98.64 | 26.28 |
| | STAR-RE | | 20.63 | 18.27 |
| | STAR-GE | | 95.67 | **34.67** |
| LoRA | Standard | 0.181 | 99.28 | **13.08** |
| | PGDAT | | 10.48 | 16.33 |
| | MixAT | | **99.35** | 10.63 |
| | STAR-RE | | 44.38 | 11.02 |
| | STAR-GE | | 99.02 | 10.52 |
| PFT | Standard | 0.224 | **99.04** | 14.40 |
| | PGDAT | | 10.00 | 10.30 |
| | MixAT | | 98.93 | 11.57 |
| | STAR-RE | | 97.07 | 18.27 |
| | STAR-GE | | 98.06 | **20.68** |
| Adpt | Standard | 1.397 | **99.25** | 20.95 |
| | PGDAT | | 19.24 | 20.33 |
| | MixAT | | 73.43 | 20.82 |
| | STAR-RE | | 56.96 | **25.42** |
| | STAR-GE | | 79.07 | 23.05 |
| FFT | Standard | 100 | **98.68** | 39.25 |
| | PGDAT | | 66.63 | 64.93 |
| | MixAT | | 85.83 | 81.38 |
| | STAR-RE | | 88.53 | 85.05 |
| | STAR-GE | | 98.23 | **85.95** |

Table 15: Training Time per Run (h) and Overhead Relative to Standard Adaptation on DINOv2.

| Dataset | Method | Training time (h) | | | | | Overhead vs. standard adaptation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Standard | PGDAT | MixAT | STAR-RE | STAR-GE | PGDAT | MixAT | STAR-RE | STAR-GE |
| CIFAR-10 | LP | 1.76 | 19.51 | 20.25 | 20.47 | 20.49 | 11.1× | 11.5× | 11.6× | 11.6× |
| | PT | 2.94 | 21.96 | 23.93 | 23.95 | 23.97 | 7.5× | 8.1× | 8.1× | 8.2× |
| | LoRA | 2.86 | 21.44 | 22.89 | 23.27 | 26.35 | 7.5× | 8.0× | 8.1× | 9.2× |
| | PFT | 3.02 | 20.76 | 22.55 | 22.20 | 22.15 | 6.9× | 7.5× | 7.4× | 7.3× |
| | Adpt | 2.72 | 20.76 | 22.44 | 22.30 | 22.88 | 7.6× | 8.2× | 8.2× | 8.4× |
| | FFT | 3.45 | 20.76 | 23.45 | 23.44 | 23.51 | 6.0× | 6.8× | 6.8× | 6.8× |

### E.6.3 Impact of the Learnable-Parameter Budget

**Setup.** RobAcc, the Robust Acc averaged over the six transfer attacks, is listed with Clean Acc for each adaptation method in order of its learnable-parameter budget $\rho$ (Table 14 and Figure 8, right).

**Results.** $(i)$ Once $\rho$ shrinks, AT cannot retain both Clean Acc and robustness, as PGDAT stays at the 10% chance rate in Clean Acc under LP, PT, LoRA, and PFT, and MixAT, at 81.38% RobAcc under FFT, loses 55.10 to 70.75 points of it under the PEFT methods, whereas `STAR-GE` keeps at least 95.67% Clean Acc under all four. $(ii)$ `STAR-GE` never falls more than 2.97 points below the Clean Acc of MixAT and exceeds it under LP, Adpt, and FFT, while the more robust mode of `STAR` exceeds MixAT in RobAcc under five of the six adaptation methods, by up to 9.12 points under PFT, and trails it by only 0.07 points under LP. $(iii)$ Robustness does not follow $\rho$, as `STAR-GE` under PT reaches 34.67% RobAcc with 0.027% of the parameters, above every model under Adpt.

### E.6.4 Training Overhead

Standard adaptation takes 1.76 to 3.45 h and every robust adaptation condition 19.51 to 26.35 h (Table 15), an overhead of 6.0× under FFT to 11.6× under LP, as on ViT-B/16 (Section E.4), since the 10-step inner maximization of Eq. 11 dominates every step. `STAR` costs about as much as MixAT, as both modes stay within 2% of its time except `STAR-GE` with LoRA at 15.1% more. We treat that run as an outlier, since the two modes compute the same four losses and differ only in their weights, and `STAR-RE` with LoRA stays within 1.7%. PGDAT takes 88.5% to 96.3% of the time of MixAT, since its outer step runs the Encoder only on the AE $x^*$, not on $x$.