FAST ADVERSARIAL TRAINING NEEDS DEBIASING AND CORRECTION TO ESCAPE THE TRAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Fast adversarial training (FAT) reduces the cost of robust training through single-step perturbation generation but remains vulnerable to catastrophic overfitting (CO), an abrupt collapse in adversarial robustness. By systematically dissecting CO from two complementary perspectives, we uncover two key failure mechanisms. First, we uncover an implicit class-specific backdoor-like effect: the outer minimization repeatedly associates biased perturbations with their ground-truth labels, causing persistent class-specific patterns to evolve into predictive triggers. Second, we uncover gradient drift along the often-overlooked post-initialization update trajectory, where the initialization gradient becomes severely misaligned with subsequent local loss-ascent directions. To address these failures, we propose DC-FGSM, which combines class-adaptive spatial debiasing to disrupt the accumulation of class-specific trigger patterns with drift-aware gradient correction to compensate for directional deviations. Experiments across multiple datasets and perturbation budgets demonstrate that DC-FGSM achieves state-of-the-art robustness among single-step adversarial training methods. Furthermore, our gradient correction module consistently improves a broad range of existing FAT baselines, demonstrating its plug-and-play generality.

1 INTRODUCTION

Despite its success across a wide range of tasks, deep learning remains vulnerable to adversarial attacks (Goodfellow et al., 2015; Szegedy et al., 2014). Adversarial training improves robustness by training models on adversarial examples, but generating these examples with multi-step attacks such as PGD (Madry et al., 2018; Zhang et al., 2019) is computationally expensive. Fast adversarial training (FAT) reduces this cost by using a single-step attack (Wong et al., 2020; de Jorge Aranda et al., 2022). However, FAT is prone to catastrophic overfitting (CO): robustness against multi-step attacks abruptly collapses, while accuracy against single-step attacks remains high (de Jorge Aranda et al., 2022). We discuss prior work on FAT in Appendix A.

Many methods mitigate CO by carefully designing the starting point of the inner maximization to improve the effectiveness of single-step perturbations (Wong et al., 2020; de Jorge Aranda et al., 2022; Pan et al., 2025a; Kim et al., 2021). Although these initialization schemes can alleviate CO, it remains unclear why they work and when they become insufficient. This raises two questions:

(Q1) Why is initialization effective, and (Q2) why does its effectiveness decline as training progresses or the perturbation budget increases?

To answer the first question, we analyze layer-wise feature responses (Saxe et al., 2019; Goldfeld et al., 2019; Kumar & Agarwal, 2026) and observe a backdoor-like phenomenon in CO. CO models exhibit a separation between clean and FGSM-perturbed inputs in their internal responses, resembling the clean-trigger separation in backdoored models (Gu et al., 2017; Kolouri et al., 2020; Min et al., 2025; Chen et al., 2017; Yang et al., 2026). These observations suggest that outer minimization reinforces class-specific perturbation–label associations, causing them to evolve into predictive triggers. By varying the starting perturbation, initialization can disrupt the accumulation of these patterns, providing an explanation for its effectiveness. To answer the second question, we examine how the gradient direction evolves after initialization (Andriushchenko & Flammarion, 2020). We find that the gradient at the initialization point can deviate substantially from local loss-ascent directions

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.

along the subsequent update. This gradient drift becomes more pronounced as training progresses or the perturbation budget increases. Changing the starting point alone therefore does not ensure a reliable direction for the single-step update. Our analysis indicates that effective FAT must address both the choice of perturbation initialization and the reliability of the subsequent update direction.

Motivated by this insight, we propose DC-FGSM, which combines class-adaptive spatial debiasing to suppress the accumulation of class-specific trigger patterns with drift-aware gradient correction to improve the reliability of single-step updates. Our main contributions are summarized as follows:

- We uncover two phenomena associated with catastrophic overfitting (CO) in FAT: the formation of class-specific trigger patterns and post-initialization gradient drift.
- We introduce a class-adaptive spatial debiasing strategy that disrupts persistent class-specific perturbation patterns while preserving adversarial effectiveness.
- We develop a lightweight drift-aware correction scheme that uses the gradient difference between the initialization point and a probing point to refine the single-step update direction.
- Extensive experiments across multiple datasets and network architectures show that DC-FGSM achieves leading robustness among FAT methods and remains stable under larger perturbation budgets. Furthermore, the proposed gradient correction module consistently improves a broad range of existing FAT methods.

2 METHODOLOGY

2.1 PRELIMINARIES

Notations and Definitions. Let $(\boldsymbol{x}, y)$ denote a sample from a K-class dataset $\mathcal{D}$, where $\boldsymbol{x} \in [0,1]^d$ is an input image and $y \in \{1, \dots, K\}$ is its ground-truth label. Let $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^K$ be a DNN parameterized by θ that outputs logits, written as f for brevity, and let $\mathcal{L}(f(\boldsymbol{x}), y)$ denote the classification loss. We consider the ℓ_∞ threat model with perturbation budget ϵ, where the feasible set is $\mathcal{B}_\epsilon(\boldsymbol{x}) = \{\boldsymbol{x}' : \|\boldsymbol{x}' - \boldsymbol{x}\|_\infty \le \epsilon\}$, and denote by $\Pi_\epsilon(\cdot)$ the projection onto $\mathcal{B}_\epsilon(\boldsymbol{x})$.

Adversarial Training (AT). Given a clean input $\boldsymbol{x}$, an adversarial example (AE) takes the form $\boldsymbol{x}_{adv} = \boldsymbol{x} + \boldsymbol{\delta}$, where the perturbation $\boldsymbol{\delta}$ satisfies $\|\boldsymbol{\delta}\|_\infty \le \epsilon$. A model f is ϵ-robust on $(\boldsymbol{x}, y)$ if $\arg\max f(\boldsymbol{x}_{adv}) = y$ for all such $\boldsymbol{\delta}$. AT seeks a robust model by solving the min-max problem

$$\min_\theta \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} \mathcal{L}\big(f_\theta(\boldsymbol{x}+\boldsymbol{\delta}), y\big) \right], \tag{1}$$

where the inner maximization crafts loss-maximizing perturbations, and the outer minimization updates θ on the resulting adversarial examples. Standard AT approximates the inner maximization with multi-step projected gradient descent (PGD), which incurs substantial computational overhead.

Fast Adversarial Training (FAT). FAT reduces the cost of AT by approximating the inner maximization with a single gradient step. The canonical choice is FGSM, which generates AE as

$$\boldsymbol{x}_{adv} = \Pi_\epsilon\big[\boldsymbol{x} + \epsilon \cdot \operatorname{sign}\big(\nabla_{\boldsymbol{x}} \mathcal{L}(f(\boldsymbol{x}), y)\big)\big]. \tag{2}$$

Despite its efficiency, FGSM-based FAT is prone to catastrophic overfitting (CO): robustness against multi-step attacks abruptly collapses, trapping the model in a largely irreversible degenerate solution.

2.2 DISSECTING THE TRAP IN FAST ADVERSARIAL TRAINING

We systematically dissect the CO trap in FAT from two complementary perspectives: *"Trigger-like Patterns"* and *"Gradient Drift"*. The first examines how class-specific, trigger-like patterns emerge during FAT and identifies the mechanism that induces them. The second moves beyond prior analyses that focus solely on the initialization point, revealing the directional drift of single-step perturbations along their update trajectory.

Trigger-Like Patterns. As shown in Fig. 1, FGSM perturbations generated after CO exhibit class-specific, trigger-like behavior. In the left panel, clean and FGSM-perturbed samples induce clearly separated layer-wise responses in the CO model, similar to the representation separation typically

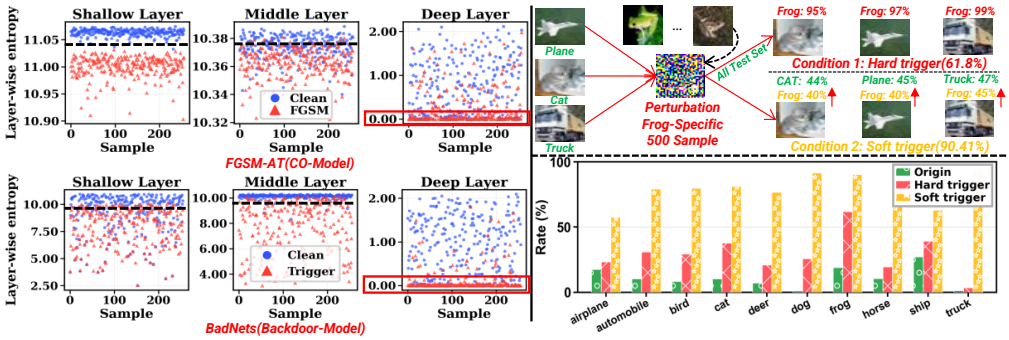

Figure 1: Class-specific trigger-like responses of FGSM perturbations after CO. Left: response patterns in catastrophically overfitted and backdoored models. Right: class-specific hard- and soft-trigger responses.

observed between clean inputs and trigger-embedded inputs in backdoored models. The right panel further reveals two complementary triggering effects. Let $\boldsymbol{x}_{cat}$ be a sample from the *cat* class, and $\boldsymbol{\delta}_{frog}$ a perturbation associated with the *frog* class. (i) *"Hard triggering"* occurs when adding $\boldsymbol{\delta}_{frog}$ flips the prediction to *frog i.e.*, $p_{frog}(\boldsymbol{x}_{cat} + \boldsymbol{\delta}_{frog}) \gg p_{cat}(\boldsymbol{x}_{cat})$. (ii) *"Soft triggering"* occurs when the perturbation raises the *frog* probability without necessarily changing the prediction, *i.e.*, $p_{frog}(\boldsymbol{x}_{cat} + \boldsymbol{\delta}_{frog}) > p_{frog}(\boldsymbol{x}_{cat})$. Together, these effects indicate that FGSM perturbations encode class-specific predictive patterns after CO (Shafahi et al., 2018; Gu et al., 2017; Turner et al., 2019; Min et al., 2025). Despite this resemblance, the two phenomena arise from fundamentally different mechanisms. In a backdoor attack, the adversary deliberately binds a trigger τ to a target label y_t by minimizing $\mathcal{L}(f(\mathcal{A}(\boldsymbol{x}, \tau)), y_t)$ on poisoned samples, where $\mathcal{A}(\cdot, \cdot)$ applies the trigger to the input. In FAT, the outer minimization likewise minimizes $\mathcal{L}(f(\boldsymbol{x} + \boldsymbol{\delta}), y)$, with $\boldsymbol{\delta}$ (generated by inner maximization) playing the role of the trigger and the ground-truth label y that of the target label. Since same-class perturbations share the same supervision, they tend to develop common patterns, which the model can exploit as class-discriminative shortcuts.

Gradient Drift. Initialization-based designs (Wong et al., 2020; Jia et al., 2022; de Jorge Aranda et al., 2022; Pan et al., 2025a) mitigate CO primarily by modifying the starting point of the inner maximization, which can be viewed as disrupting the class-specific patterns induced by its interplay with the outer minimization. However, these designs implicitly assume that the gradient at the initialization point remains reliable throughout the subsequent update, leaving the post-initialization perturbation dynamics largely unexplored. This raises a key question: *Does the gradient at the initialization point remain a reliable ascent direction along the subsequent update trajectory?*

To answer this question, we introduce *gradient drift* to quantify the deviation between the initialization gradient and the local ascent direction along the subsequent update. Given input $\boldsymbol{x}_0$ with the initialization perturbation $\boldsymbol{\delta}$, let $\boldsymbol{g}_0 = \nabla_{\boldsymbol{\delta}} \mathcal{L}\big(f(\boldsymbol{x}_0 + \boldsymbol{\delta}), y\big)$ denote the loss gradient at the initialization point. We probe the update trajectory at $\boldsymbol{\delta}_r = \boldsymbol{\delta}_0 + \epsilon_r \cdot \mathrm{sign}(\boldsymbol{g}_0)$ with $\epsilon_r = \epsilon/2$, and let $\boldsymbol{g}_r$ denote the loss gradient at this probing point. Gradient drift is measured as the discrepancy between $\boldsymbol{g}_0$ and $\boldsymbol{g}_r$, where a larger discrepancy indicates greater local gradient variation and a less reliable initialization gradient for the subsequent update. As shown in Fig. 2, gradient drift is negatively correlated with robust accuracy across the evaluated FAT methods and initialization schemes. The left panels track epoch-level trends, while the middle and right panels compare batch-level behavior before and after CO. Before CO, both metrics remain stable; after CO, drift becomes highly volatile, with sharp spikes coinciding with abrupt drops in robust accuracy despite persistently high FGSM accuracy. Additional analyses of gradient drift across other FAT methods are provided in Appendix B.2.

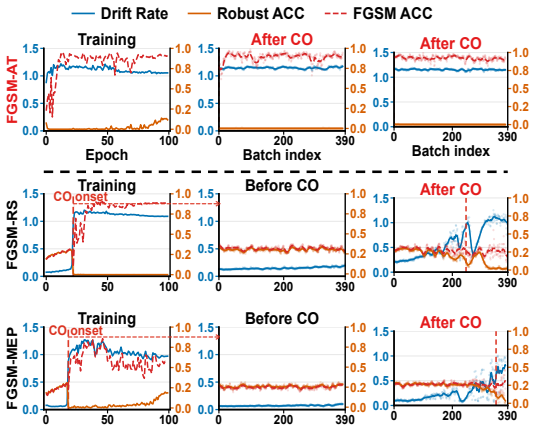

Figure 2: Gradient drift and robust accuracy under different FAT initialization schemes.

2.3 DEBIASED INITIALIZATION

Existing initialization-based methods use random noise or historical perturbations to mitigate CO. Random noise can reduce pattern accumulation but may introduce stochastic directional bias, whereas reusing historical perturbations can preserve and accumulate class-specific trigger-like structures. To tackle this issue, we propose a class-adaptive debiased initialization strategy that suppresses the accumulation of shared spatial trigger patterns while maintaining effective adversarial directions.

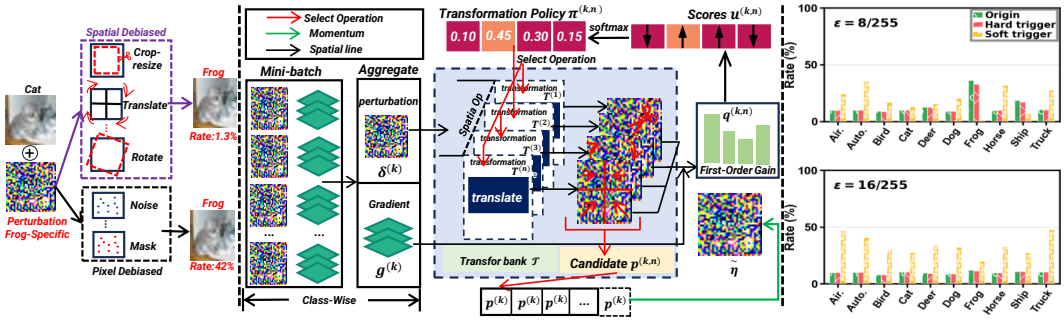

Figure 3: Overview of the class-adaptive spatial debiasing in DC-FGSM. For each class, perturbation priors are transformed by candidate spatial operations and scored to select a class-adaptive initialization. Left: pixel-level and spatial disruption strategies. right: class-wise triggering statistics under different perturbation budgets, demonstrating reduced trigger patterns with our method.

Class-Adaptive Spatial Trigger Disruption. As shown in the left panel of Fig. 3, trigger-like perturbations exhibit class-specific spatial structures. We therefore cast trigger disruption as a *"class-adaptive transformation selection"* problem. Let $k \in \{1, \cdots, K\}$ index the classes and $\mathcal{T} = \{T^{(1)}, \cdots, T^{(N)}\}$ denote a bank of N candidate spatial transformations. For each class k, we aggregate the perturbations and input gradients of its samples into an empirical perturbation prototype $\boldsymbol{\delta}^{(k)}$ and an ascent direction $\boldsymbol{g}^{(k)}$, respectively. We sample a transformation index $n \sim \boldsymbol{\pi}^{(k)}$ from a class-specific policy $\boldsymbol{\pi}^{(k)} = (\pi^{(k,1)}, \ldots, \pi^{(k,N)})$ over $\mathcal{T}$, where $\sum_{n=1}^{N} \pi^{(k,n)} = 1$, and apply it to the prototype to obtain the candidate perturbation

$$\boldsymbol{p}^{(k,n)} = \mathcal{N}_\epsilon\big(T^{(n)}(\boldsymbol{\delta}^{(k)})\big), \tag{3}$$

where $\mathcal{N}_\epsilon$ rescales the transformed perturbation to match the mean absolute magnitude of $\boldsymbol{\delta}^{(k)}$ and projects it onto $\mathcal{S}_\epsilon = \{\boldsymbol{p} : \|\boldsymbol{p}\|_\infty \leq \epsilon\}$. For class k, we further maintain a buffer $\mathcal{P}^{(k)}$ of historical perturbations, whose mean $\bar{\boldsymbol{p}}^{(k)}$ is set to zero when the buffer is empty. Ideally, for k-th class, to preserve adversarial effectiveness after transformation, the policy should select the transformation that maximizes the loss gain over this historical prior:

$$n_k^* = \underset{n \in \{1, \cdots, N\}}{\arg\max} \left[\ell^{(k)}\left(\boldsymbol{p}^{(k,n)}\right) - \ell^{(k)}\left(\bar{\boldsymbol{p}}^{(k)}\right) \right], \tag{4}$$

where $\ell^{(k)}(\cdot)$ denotes the average classification loss over class-k samples. Exact optimization of equation 4 requires evaluating all N candidates, while hard selection based on instantaneous gains is sensitive to minibatch noise and may collapse onto a single transformation. We therefore update $\boldsymbol{\pi}^{(k)}$ online using history-aware first-order feedback to balance adversarial effectiveness and transformation diversity.

First-Order Alignment Score. Specifically, we reuse the class-specific gradient $\boldsymbol{g}^{(k)}$ computed during adversarial example generation and define the normalized first-order alignment score as

$$q^{(k,n)} = \frac{\left\langle \boldsymbol{g}^{(k)}, \boldsymbol{p}^{(k,n)} - \bar{\boldsymbol{p}}^{(k)} \right\rangle}{\left\| \boldsymbol{g}^{(k)} \right\|_1}, \tag{5}$$

which measures the candidate's alignment with the current class-specific ascent direction, with larger values indicating more effective transformations while reducing sensitivity to gradient scale.

Since instantaneous alignment scores vary across minibatches, we standardize them within each class as $\tilde{q}^{(k,n)} = (q^{(k,n)} - \mu_q^{(k)})/\sigma_q^{(k)}$, where $\mu_q^{(k)}$ and $\sigma_q^{(k)}$ are the mean and standard deviation of the N

candidate scores, respectively. Rather than greedily committing to a noisy instantaneous estimate, we treat $\widetilde{q}^{(k,n)}$ as online feedback and accumulate it into a class-specific utility $u^{(k,n)}$, initialized to zero, which in turn parameterizes a stochastic policy:

$$u^{(k,n)} \leftarrow (1-\rho)\cdot u^{(k,n)} + \rho \cdot \widetilde{q}^{(k,n)}, \qquad \pi^{(k,n)} = (1-\xi)\cdot \frac{\exp\left(u^{(k,n)}/\omega\right)}{\sum_{j=1}^{N}\exp\left(u^{(k,j)}/\omega\right)} + \frac{\xi}{N}, \quad (6)$$

where ρ denotes the adaptation rate, ω the policy temperature, $\xi \in [0,1]$ the exploration coefficient, and $N = |\mathcal{T}|$ the number of candidate transformations. *The smoothed utilities reduce sensitivity to transient minibatch noise, so the policy favors consistently effective transformations. Meanwhile, the uniform mixture guarantees $\pi^{(k,n)} \geq \xi/N$, preserving exploration and preventing probability collapse*. At each iteration, we sample $n \sim \boldsymbol{\pi}^{(k)}$ and add $\boldsymbol{p}^{(k,n)}$ to $\mathcal{P}^{(k)}$.

Debiased Prior Momentum. Unlike prior-guided initialization methods that directly reuse historical perturbations (Jia et al., 2022; Pan et al., 2024), our class-specific buffers store perturbation prototypes selected through class-adaptive spatial transformations. Each buffer $\mathcal{P}^{(k)}$ follows a first-in-first-out (FIFO) scheme and retains only the most recent entries, so that its mean $\bar{\boldsymbol{p}}^{(k)}$ tracks the evolving model while suppressing transient variations across minibatches. Building on this, we maintain a class-wise prior $\boldsymbol{m}^{(k)}$, initialized to zero and updated via exponential momentum smoothing:

$$\boldsymbol{m}^{(k)} = \mu \cdot \boldsymbol{m}^{(k)} + (1-\mu)\cdot \operatorname{sign}\left(\bar{\boldsymbol{p}}^{(k)}\right), \quad (7)$$

For a training sample $(\boldsymbol{x}, y)$, the debiased initialization is then given by

$$\tilde{\boldsymbol{x}} = \Pi_\epsilon\left(\boldsymbol{x} + \tilde{\boldsymbol{\eta}}^{(y)}\right), \quad \tilde{\boldsymbol{\eta}}^{(y)} = \epsilon \cdot \operatorname{sign}(\boldsymbol{m}^{(y)}), \quad (8)$$

where $\tilde{\boldsymbol{\eta}}^{(y)}$ denotes the class-wise debiased perturbation for class y. *Since $\boldsymbol{m}^{(y)}$ accumulates transformed prototypes selected by the alignment score rather than raw historical perturbations, $\tilde{\boldsymbol{\eta}}^{(y)}$ remains aligned with the class-specific ascent direction while breaking the fixed spatial layouts that would otherwise evolve into triggers.* As shown in the right panel of Fig. 3, the debiased initialization consistently reduces hard- and soft-trigger responses across classes and perturbation budgets, while preserving adversarial effectiveness.

2.4 Gradient Correction

Building on the gradient-drift phenomenon identified in Section 2.2, we quantify the error from reusing the initialization gradient during the single-step update. Let $\boldsymbol{g}_0$ and $\boldsymbol{g}_r$ denote the loss gradients at the initialization perturbation $\boldsymbol{\delta}_0$ and the probing perturbation $\boldsymbol{\delta}_r$, respectively. We define the gradient-drift residual as $\bar{R} = \|\boldsymbol{g}_r - \boldsymbol{g}_0\|_2$, which measures the local deviation from the initialization gradient along the update trajectory.

Restoring Reliable Single-Step Updates. To improve the reliability of the single-step update, we probe the local loss landscape around the debiased initialization and use the resulting gradient variation to correct post-initialization drift. Starting from the debiased initialization $\tilde{\boldsymbol{x}}$, we compute its input gradient as $\tilde{\boldsymbol{g}} = \nabla_{\tilde{\boldsymbol{x}}}\mathcal{L}(f(\tilde{\boldsymbol{x}}), y)$, and take a probing step along the initial ascent direction:

$$\boldsymbol{x}_r = \Pi_\epsilon\left(\tilde{\boldsymbol{x}} + \epsilon_r \operatorname{sign}\left(\tilde{\boldsymbol{g}}\right)\right), \quad (9)$$

where ϵ_r denotes the probing step size. Let $\boldsymbol{h} = \boldsymbol{x}_r - \tilde{\boldsymbol{x}}$ be the resulting local displacement and $\boldsymbol{g}_r = \nabla_{\boldsymbol{x}_r}\mathcal{L}\big(f(\boldsymbol{x}_r), y\big)$ the gradient at the probing point. A first-order Taylor expansion of the gradient around $\tilde{\boldsymbol{x}}$ gives $\boldsymbol{g}_r = \tilde{\boldsymbol{g}} + \boldsymbol{H}_{\tilde{\boldsymbol{x}}}\boldsymbol{h} + o(\|\boldsymbol{h}\|)$, where $\boldsymbol{H}_{\tilde{\boldsymbol{x}}}$ is the input Hessian of $\mathcal{L}$ at $\tilde{\boldsymbol{x}}$. The gradient difference $\boldsymbol{g}_r - \tilde{\boldsymbol{g}} \approx \boldsymbol{H}_{\tilde{\boldsymbol{x}}}\boldsymbol{h}$ thus captures the local curvature along the update, which is exactly the drift measured by $\bar{R}$. Based on this approximation, we define the corrected gradient as

$$\widehat{\boldsymbol{g}} = \tilde{\boldsymbol{g}} + \lambda \cdot (\boldsymbol{g}_r - \tilde{\boldsymbol{g}}) \approx \nabla_{\tilde{\boldsymbol{x}}+\lambda\boldsymbol{h}}\mathcal{L}\left(f\left(\tilde{\boldsymbol{x}} + \lambda\boldsymbol{h}\right), y\right), \quad (10)$$

where λ is the correction coefficient. In other words, $\widehat{\boldsymbol{g}}$ approximates the gradient at a point along the update trajectory, integrating the initialization gradient with its local variation without explicit second-order computation. Detailed theoretical analysis and proofs are provided in the Appendix B.2. Finally, given the *debiased initialization* $\tilde{\boldsymbol{x}}$ and the *corrected gradient* $\widehat{\boldsymbol{g}}$, we generate the final adversarial example using a single projected ascent step:

$$\tilde{\boldsymbol{x}}_{adv} = \Pi_\epsilon\left(\tilde{\boldsymbol{x}} + \epsilon \cdot \operatorname{sign}\left(\widehat{\boldsymbol{g}}\right)\right). \quad (11)$$

Together, the two components disrupt class-specific spatial triggers and realign the single-step update with the local ascent direction.

Overall Loss. Inspired by prior work on consistency regularization (Min et al., 2025; Andriushchenko & Flammarion, 2020; Zhang et al., 2019; Pan et al., 2025a), we incorporate a Jensen-Shannon divergence (JSD) regularizer into the outer minimization to enforce prediction consistency along the inner update trajectory, *i.e.*, between the debiased initialization $\tilde{\boldsymbol{x}}$ and the final adversarial example $\tilde{\boldsymbol{x}}_{\text{adv}}$, which helps stabilize the training gradients. Accordingly, the model is updated by jointly minimizing this consistency term and the classification loss on $\tilde{\boldsymbol{x}}_{\text{adv}}$, yielding the overall training objective:

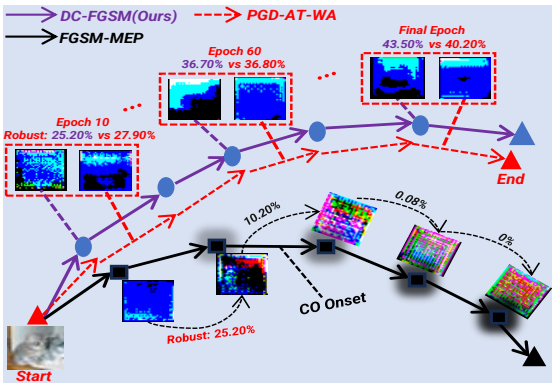

Figure 4: Training trajectories of DC-FGSM, FGSM-MEP, and multi-step AT under the perturbation budget of $\epsilon = 16/255$.

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\mathcal{L}\big(f(\tilde{\boldsymbol{x}}_{adv}), y\big) + \beta\mathcal{L}_{\text{JSD}}\big(f(\tilde{\boldsymbol{x}}), f(\tilde{\boldsymbol{x}}_{adv})\big)\right], \tag{12}$$

where β controls the strength of the inner-trajectory consistency regularizer. As shown in Fig. 4, DC-FGSM follows a more stable optimization trajectory than methods that directly reuse historical momentum. The complete training procedure is provided in Algorithm 1 of Appendix C.

3 Experiments

3.1 Experimental Setup

Attack Evaluation. We evaluate all methods using standard adversarial attacks, including PGD with 10, 20, and 50 steps (Madry et al., 2018), CW (Carlini & Wagner, 2017), AutoAttack (Croce & Hein, 2020), and the recent white-box attacks CR (Pan et al., 2025b) and PMA (Xie et al., 2025). During training, we use PGD-10 accuracy to monitor robustness and select checkpoints. In the figures, we refer to this metric as Robust ACC. We use ℓ_∞ perturbation budgets of $\epsilon \in \{8/255, 16/255\}$ for CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), and $\epsilon \in \{8/255, 12/255\}$ for Tiny-ImageNet (de Jorge Aranda et al., 2022; Pan et al., 2025a; Deng et al., 2009).

Baseline Methods. We compare DC-FGSM with representative FAT methods, including FGSM-RS, N-FGSM (de Jorge Aranda et al., 2022), FGSM-MEP (Jia et al., 2022), GAT (Sriramanan et al., 2020), GradAlign (Andriushchenko & Flammarion, 2020), NU-AT (Sriramanan et al., 2021), and LIET (Pan et al., 2025a). Because clean accuracy alone can be misleading when CO occurs, we summarize the clean–robust trade-off using the average of clean accuracy and robust accuracies across our full suite of adversarial attacks. Representative attacks in the main text include PGD-50, CW, AA, CR, and PMA. further attack results are available in the Appendix G. We also include PGD-AT-WA*** as a strong multi-step reference, where WA denotes weight averaging. We denote our variant with debiased initialization alone as D-FGSM and the full method with both debiased initialization and gradient correction as DC-FGSM.

Training Details. All attacks are evaluated in the white-box setting: adversarial examples are generated directly on the model under evaluation, with full access to its parameters and gradients. We train the target model for 100 epochs using SGD with momentum 0.9, weight decay 5×10^{-4}, a one-cycle learning-rate schedule, and a batch size of 128. Based on our ablation study, we set the probing distance and gradient-correction coefficient to $(r, \lambda) = (0.6\epsilon, 0.8)$ for $\epsilon = 8/255$ and $(r, \lambda) = (0.4\epsilon, 0.5)$ for $\epsilon = 16/255$. We implement all experiments in PyTorch 1.12.1 and run them on NVIDIA A10 GPUs. Unless otherwise stated, we report mean results over three random seeds with all other experimental settings held fixed. Additional implementation details are provided in Appendix G.1.

Datasets and Backbones. We evaluate DC-FGSM on three standard benchmarks for adversarial robustness: CIFAR-10, CIFAR-100, and Tiny-ImageNet. For CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), we use ResNet-18 (He et al., 2016), ResNet-34, and WideResNet-28-10 (Zagoruyko & Komodakis, 2016) as backbone networks. For Tiny-ImageNet (Deng et al., 2009), we use ResNet-18.

Table 1: Clean and robust accuracy (%) on CIFAR-10 and CIFAR-100 under $\epsilon = 8/255$ and $\epsilon = 16/255$. Avg. denotes the average of clean and robust accuracy. Best single-step results are shown in **bold red**.

Method	CIFAR-10							CIFAR-100						
	Clean	PGD-50	CW	AA	CR	PMA	Avg.	Clean	PGD-50	CW	AA	CR	PMA	Avg.
$\epsilon = 8/255$, Best checkpoint														
PGD-AT-WA***	82.15	54.36	52.34	50.49	50.54	50.71	56.76	58.23	31.38	28.33	26.48	26.48	26.74	32.94
FGSM-RS	79.94	52.74	49.56	47.29	47.36	47.48	54.06	46.02	22.24	19.37	17.51	17.48	17.72	23.39
N-FGSM	79.88	53.02	50.13	48.00	48.09	48.19	54.55	54.48	30.66	27.51	25.40	25.44	25.66	31.53
FGSM-MEP	79.77	55.72	51.47	50.00	49.98	50.18	56.19	56.75	31.71	27.92	26.24	26.26	26.50	32.56
GAT	85.07	53.36	50.35	48.70	48.66	48.87	55.84	65.73	26.32	23.81	21.91	21.84	22.09	30.28
GRAD-ALIGN	80.11	52.75	50.30	48.19	48.24	48.39	54.66	55.09	30.85	27.47	25.30	25.29	25.58	31.60
NU-AT	80.91	54.74	51.81	50.38	50.36	50.49	56.45	57.92	20.77	19.26	13.28	16.72	16.96	24.15
LIET	80.22	56.22	51.44	49.97	49.93	50.13	56.32	51.38	32.49	27.74	26.67	26.68	26.93	31.98
DC-FGSM(Ours)	80.96	**57.27**	**52.06**	**50.61**	**50.62**	**50.79**	**57.05**	53.49	**33.12**	**28.57**	**27.27**	**27.27**	**27.52**	**32.87**
$\epsilon = 8/255$, Last checkpoint														
PGD-AT-WA***	85.17	50.55	49.45	47.44	47.60	47.62	54.64	59.17	27.24	24.38	22.70	22.67	22.94	29.85
FGSM-RS	90.84	12.80	9.27	0.00	10.74	18.36	23.67	61.72	0.15	0.10	0.00	1.50	0.74	10.70
N-FGSM	83.62	47.32	46.41	43.86	43.87	44.13	51.53	57.37	26.72	23.87	22.16	22.15	22.32	29.10
FGSM-MEP	83.14	53.14	50.42	48.51	48.50	48.74	55.41	58.14	27.73	24.17	22.59	22.61	22.85	29.68
GAT	85.33	51.98	49.66	47.62	47.54	47.81	54.99	64.80	26.13	23.55	21.73	21.68	21.96	29.97
GRAD-ALIGN	85.10	46.55	45.99	43.47	43.62	43.80	51.42	57.73	23.46	20.72	18.56	18.55	18.99	26.34
NU-AT	81.96	53.91	51.22	49.56	49.55	49.69	55.98	62.98	18.44	16.96	11.24	16.20	15.75	23.59
LIET	82.71	54.74	50.77	49.08	49.09	49.31	55.95	52.58	32.18	27.50	26.35	26.35	26.57	31.92
DC-FGSM(Ours)	82.71	**56.13**	**51.55**	**49.85**	**49.84**	**50.12**	**56.63**	54.08	**32.87**	**28.29**	**26.93**	**26.93**	**27.27**	**32.73**
$\epsilon = 16/255$, Best checkpoint														
PGD-AT-WA***	72.45	30.03	28.32	24.10	24.70	25.02	34.10	46.38	17.12	14.32	12.28	12.50	12.73	19.22
FGSM-RS	49.75	23.36	19.08	16.12	16.28	16.63	23.54	26.43	10.72	8.12	6.98	7.17	7.25	11.11
N-FGSM	64.03	31.14	26.48	22.13	22.69	23.12	31.60	41.42	15.52	13.05	10.94	11.17	11.32	17.24
FGSM-MEP	43.60	24.51	20.52	19.09	19.09	19.50	24.39	29.09	12.00	9.16	8.02	8.05	8.27	12.43
GAT	70.07	18.68	14.75	12.96	12.79	13.16	23.73	50.03	2.91	2.69	1.40	1.48	1.64	10.02
GRAD-ALIGN	62.54	28.93	24.10	19.68	20.17	20.75	29.36	31.75	11.64	9.05	7.52	7.66	7.87	12.58
NU-AT	91.14	6.23	8.32	0.11	3.33	3.71	18.81	27.61	13.30	9.53	8.24	8.34	8.60	12.60
LIET	60.40	33.14	26.56	22.97	23.19	23.92	31.70	33.19	15.87	12.55	11.11	11.23	11.62	15.93
DC-FGSM(Ours)	72.54	**34.00**	**27.46**	**23.46**	**23.49**	**24.60**	**34.25**	47.38	**18.41**	**15.00**	**11.95**	**12.41**	**13.05**	**19.70**
$\epsilon = 16/255$, Last checkpoint														
PGD-AT-WA***	76.10	25.01	26.07	19.57	20.31	22.55	31.60	50.44	13.42	12.27	9.55	10.42	10.69	17.80
FGSM-RS	69.57	0.36	0.54	0.00	1.84	3.17	12.58	34.20	0.00	0.00	0.00	0.31	0.09	5.77
N-FGSM	70.69	23.00	23.00	15.80	16.18	18.86	27.92	46.75	11.46	10.57	7.85	8.78	8.96	15.73
FGSM-MEP	88.48	6.43	3.61	0.13	16.07	11.69	21.07	64.55	0.55	0.32	0.04	6.90	1.74	12.35
GAT	91.67	3.08	3.04	0.99	2.21	2.03	17.17	70.77	0.58	0.48	0.06	1.48	0.60	12.33
GRAD-ALIGN	72.44	19.07	19.28	12.22	12.84	15.37	25.20	52.77	0.00	0.01	0.00	0.70	0.12	8.93
NU-AT	91.27	5.64	5.57	0.11	3.02	2.50	18.02	62.86	3.45	3.11	1.45	3.38	2.87	12.85
LIET	76.97	12.72	13.58	7.38	14.05	12.33	22.84	53.19	6.25	5.25	4.00	5.81	5.66	13.36
DC-FGSM(Ours)	77.51	**29.09**	**24.61**	**19.41**	**19.74**	**21.33**	**31.95**	50.58	**16.67**	**13.61**	**10.16**	**11.27**	**11.94**	**19.04**

Table 2: Clean and robust accuracy (%) on Tiny-ImageNet under $\epsilon = 8/255$ and $\epsilon = 12/255$. Avg. denotes the average of clean and robust accuracy. Best single-step results are shown in **bold red**.

Method	$\epsilon = 8/255$							$\epsilon = 12/255$						
	Clean	PGD-50	CW	AA	CR	PMA	Avg.	Clean	PGD-50	CW	AA	CR	PMA	Avg.
Best checkpoint														
PGD-AT-WA***	49.18	26.24	21.87	19.89	19.91	20.27	26.23	41.55	18.53	14.38	12.28	12.26	12.74	18.62
FGSM-RS	49.35	25.11	20.66	18.40	18.35	18.69	25.09	24.76	9.89	6.88	5.67	5.64	5.88	9.79
N-FGSM	46.12	25.11	20.58	18.66	18.62	18.95	24.67	40.19	16.73	13.08	11.00	10.93	11.30	17.21
FGSM-MEP	49.27	26.11	21.27	19.40	19.38	19.73	25.86	23.81	10.58	7.34	6.13	6.06	6.45	10.06
GAT	57.43	16.88	12.98	11.20	11.13	11.36	20.16	43.46	0.49	0.40	0.12	0.13	0.18	7.46
GRAD-ALIGN	47.38	24.73	20.52	18.50	18.46	18.88	24.75	35.65	13.93	10.49	8.52	8.46	8.86	14.32
NU-AT	27.91	15.21	11.55	10.40	10.41	10.69	14.36	41.28	7.64	5.56	3.50	4.02	4.48	11.08
LIET	45.02	26.71	21.35	19.84	19.83	20.10	25.47	33.93	17.34	13.23	11.29	11.12	11.83	16.46
DC-FGSM(Ours)	46.53	**27.29**	**21.91**	**20.27**	**20.23**	**20.57**	**26.13**	40.15	**18.14**	**13.45**	**11.37**	**11.29**	**11.85**	**17.71**
Last checkpoint														
PGD-AT-WA***	51.39	23.60	19.70	17.59	17.61	17.84	24.62	44.51	16.65	13.34	11.07	11.01	11.49	18.01
FGSM-RS	47.65	10.50	8.34	7.06	7.14	7.21	14.65	18.38	0.01	0.00	0.00	0.00	0.00	3.06
N-FGSM	49.37	21.98	17.59	15.68	15.65	15.95	22.70	42.73	14.76	11.30	9.29	9.24	9.58	16.15
FGSM-MEP	52.76	12.11	9.27	8.18	11.23	8.66	17.04	40.59	0.00	0.00	0.00	3.04	0.00	7.27
GAT	57.39	17.02	13.09	11.34	11.28	11.54	20.28	63.96	0.12	0.06	0.00	0.07	0.06	10.71
GRAD-ALIGN	49.35	11.03	8.91	7.85	7.84	8.00	15.50	39.67	8.00	5.85	4.63	4.60	4.86	11.27
NU-AT	0.50	0.50	0.50	0.50	0.50	0.50	0.50	56.02	4.40	2.77	1.30	1.94	2.53	11.49
LIET	45.89	25.92	20.90	19.26	19.24	19.56	25.13	47.61	5.31	3.46	1.62	3.36	3.54	10.82
DC-FGSM(Ours)	47.77	**26.86**	**21.62**	**20.00**	**19.98**	**20.33**	**26.09**	42.49	**17.46**	**13.15**	**10.97**	**10.83**	**11.41**	**17.72**

3.2 Main Results

Results on CIFAR-10 and CIFAR-100. DC-FGSM achieves the highest clean–robust average accuracy among single-step methods on both datasets at both perturbation budgets and checkpoints (Table 1). At the final checkpoint under $\epsilon = 16/255$, it retains 19.41% AutoAttack accuracy on CIFAR-10, exceeding N-FGSM and GradAlign by 3.61 and 7.19 percentage points, respectively. Standard deviations across random seeds are reported in Appendix G.2.

Results on Tiny-ImageNet. DC-FGSM also achieves the highest clean–robust average accuracy among single-step methods on Tiny-ImageNet at both perturbation budgets and checkpoints. At the final checkpoint, it reaches 26.09% under $\epsilon = 8/255$ and 17.72% under $\epsilon = 12/255$, exceeding the strongest single-step baselines by 0.96 and 1.57 percentage points, respectively. Additional results are provided in Appendix G.2.

3.3 MECHANISTIC ANALYSIS OF DC-FGSM

We examine the applicability of gradient correction across FAT methods and analyze the training dynamics of DC-FGSM.

Compatibility of Gradient Correction. We integrate gradient correction (GC) into seven FAT baselines. As shown in Table 3, GC generally improves robustness across both datasets and the evaluated attacks. These results support its applicability across FAT methods and suggest that correcting post-initialization gradient mismatch can benefit methods with different strategies.

Table 3: Compatibility of gradient correction with existing FAT methods. ★ marks corrected variants. Improvements over the corresponding original method are highlighted in **bold red**, except in the Clean columns.

Method	CIFAR-10							CIFAR-100						
	Clean	PGD-50	CW	AA	CR	PMA	Avg.	Clean	PGD-50	CW	AA	CR	PMA	Avg.
FGSM-RS	79.94	52.74	49.56	47.29	47.36	47.48	54.06	46.02	22.24	19.37	17.51	17.48	17.72	23.39
FGSM-RS★	82.74	**54.51**	**51.47**	**49.51**	**49.77**	**49.60**	**56.27**	47.06	**23.68**	**20.46**	**18.68**	**18.68**	**18.90**	**24.58**
N-FGSM	79.88	53.02	50.13	48.00	48.09	48.19	54.55	54.48	30.66	27.51	25.40	25.44	25.66	31.53
N-FGSM★	80.23	**54.73**	**51.51**	**49.80**	**49.82**	**49.94**	**56.00**	55.23	**31.67**	**28.35**	**26.53**	**26.52**	**26.75**	**32.51**
FGSM-MEP	79.77	55.72	51.47	50.00	49.98	50.18	56.19	56.75	31.71	27.92	26.24	26.26	26.50	32.56
FGSM-MEP★	80.10	**55.92**	**51.61**	**50.21**	**50.23**	**50.36**	**56.41**	57.99	**32.28**	**28.39**	**26.82**	**26.81**	**27.08**	**33.23**
GAT	85.07	53.36	50.35	48.70	48.66	48.87	55.84	65.73	26.32	23.81	21.91	21.84	22.09	30.28
GAT★	85.02	**54.78**	**50.54**	**48.97**	**48.95**	**49.17**	**56.24**	64.76	**28.19**	**25.12**	**23.43**	**23.41**	**23.66**	**31.43**
GRAD-ALIGN	80.11	52.75	50.30	48.19	48.24	48.39	54.66	55.09	30.85	27.47	25.30	25.29	25.58	31.60
GRAD-ALIGN★	79.95	**55.03**	**51.40**	**49.72**	**49.71**	**49.90**	**55.95**	56.51	**31.73**	**28.31**	**26.42**	**26.43**	**26.72**	**32.69**
NU-AT	80.91	54.74	51.81	50.38	50.36	50.49	56.45	57.92	20.77	19.26	13.28	16.72	16.96	24.15
NU-AT★	80.31	**55.95**	**52.44**	**51.20**	**51.21**	**51.29**	**57.07**	52.43	**21.79**	**19.40**	**14.71**	**18.08**	**17.24**	23.94
LIET	80.22	56.22	51.44	49.97	49.93	50.13	56.32	51.38	32.49	27.74	26.67	26.68	26.93	31.98
LIET★	81.70	**56.94**	**52.06**	**50.38**	**50.36**	**50.56**	**57.00**	53.13	**33.44**	**28.87**	**27.72**	**27.70**	**27.98**	**33.14**

Layer-wise Feature Entropy. Fig. 5 compares the layer-wise entropy profiles of clean and FGSM-adversarial examples. DC-FGSM maintains similar profiles across layers, as does multi-step PGD-AT-WA, whereas the FAT baselines show larger differences between clean and adversarial inputs. These results suggest that DC-FGSM limits changes in feature entropy under FGSM perturbations. Additional comparisons are provided in Appendix F.

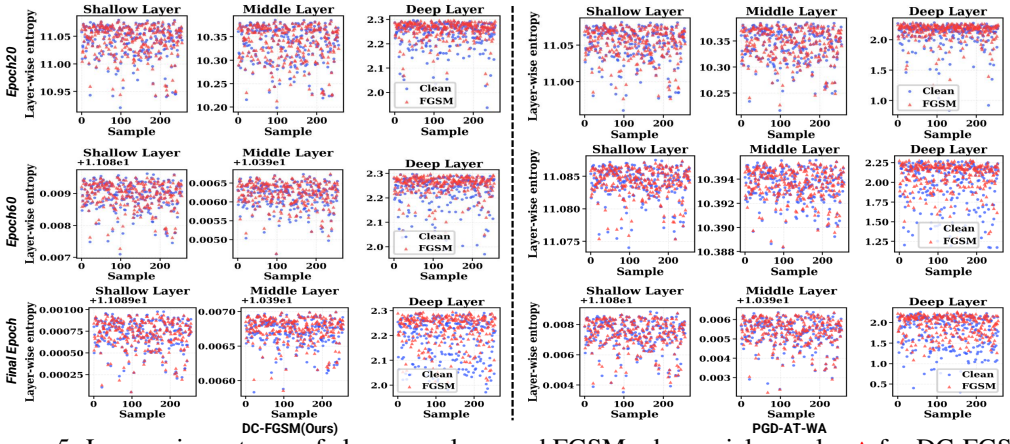

Figure 5: Layer-wise entropy of clean samples • and FGSM-adversarial samples ▲ for DC-FGSM and multi-step PGD-AT-WA during training under $\epsilon = 16/255$.

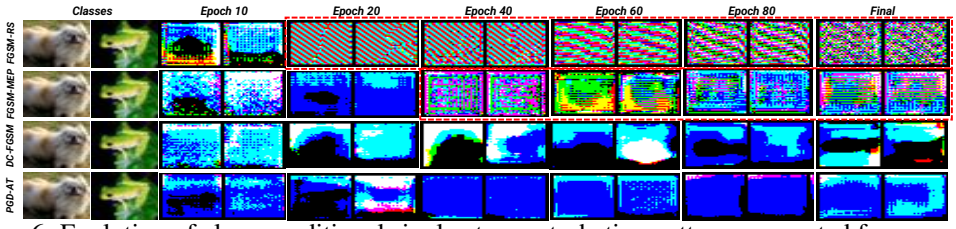

Figure 6: Evolution of class-conditional single-step perturbation patterns generated from a neutral gray input. Red boxes highlight persistent class-specific spatial structures that emerge after CO.

Evolution of Class-Specific Trigger Patterns. We visualize single-step perturbations generated from a neutral gray input for selected target classes throughout training. Fig. 6 shows that DC-FGSM largely avoids persistent class-specific spatial patterns. This observation supports the role of spatial debiasing in limiting self-induced trigger formation. Perturbation patterns for the other baselines are provided in Appendix B.1.

3.4 Ablation Study

Impact of Spatial Transformation Strategies. Fig. 7 compares random-noise initialization with spatial debiasing using fixed, random, or first-order alignment-based transformation selection. The alignment-based strategy achieves the strongest and most consistent robustness across the evaluated attacks, supporting the benefit of adaptive transformation selection.

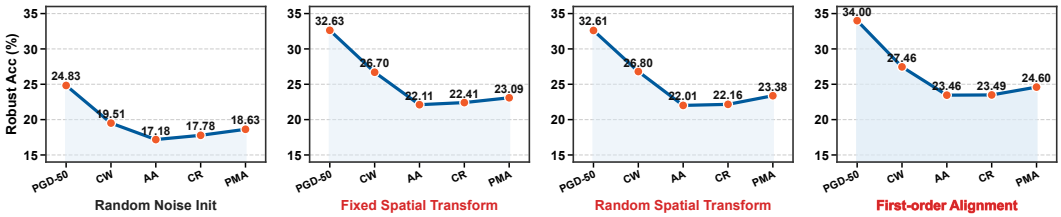

Figure 7: Ablation of spatial transformation strategies on CIFAR-10 under $\epsilon = 16/255$.

Impact of Gradient Correction. Fig. 8 compares D-FGSM and DC-FGSM on three datasets, each evaluated at two perturbation budgets under five attacks. Gradient correction consistently improves robustness, generally yielding larger gains at the higher budgets of $\epsilon = 16/255$ on CIFAR-10 and CIFAR-100 and $\epsilon = 12/255$ on Tiny-ImageNet. The training curves also show reduced late-stage gradient drift and more stable robust accuracy. Together, these results support the benefit of gradient correction beyond spatial debiasing alone.

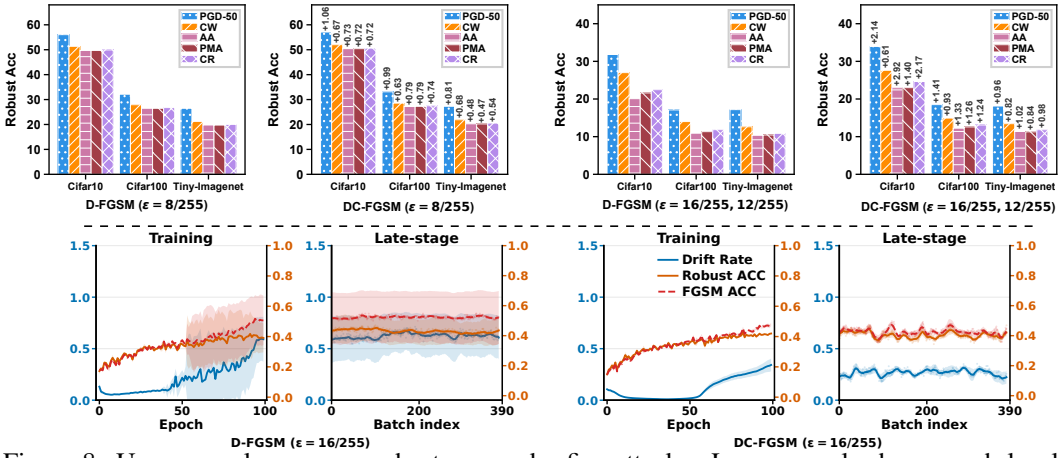

Figure 8: Upper panels compare robustness under five attacks. Lower panels show epoch-level training dynamics and batch-level behavior during the first epoch after CO. Shaded regions indicate standard deviations across random seeds.

Efficiency and Hyperparameter Sensitivity. Debiasing initialization adds no extra forward/backward passes. Gradient correction adds one forward–backward pass per batch while active. Restricting the correction to the last 50 epochs ($E_c = 50$) instead of all epochs reduces the cost from 69.57 to 62.08 minutes with comparable robustness. This late-stage variant is our default. A detailed complexity analysis is provided in Appendix D. Based on the sensitivity analysis, we use $(r, \lambda) = (0.6\epsilon, 0.8)$ for $\epsilon = 8/255$ and $(r, \lambda) = (0.4\epsilon, 0.5)$ for $\epsilon = 16/255$.

4 Conclusion

We identify class-specific trigger-like perturbations and gradient drift as key factors contributing to catastrophic overfitting in fast adversarial training. To address these factors, we propose DC-FGSM, which combines class-adaptive spatial debiasing with gradient correction. Spatial debiasing disrupts persistent class-specific patterns, while gradient correction compensates for directional drift during single-step adversarial example generation. Experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet show that DC-FGSM achieves strong robustness across the evaluated perturbation budgets. The gradient correction module also improves robustness across a range of FAT baselines, supporting its use as a plug-and-play component.

REFERENCES

Elias Abad Rocamora, Fanghui Liu, Grigorios Chrysos, Pablo M. Olmos, and Volkan Cevher. Efficient local linearity regularization to overcome catastrophic overfitting. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 48418–48447, 2024.

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16048–16059. Curran Associates, Inc., 2020.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017. doi: 10.1109/SP.2017.49.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. In *NIPS 2017 Workshop on Machine Learning and Computer Security*, 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 13–18 Jul 2020.

Pau de Jorge Aranda, Adel Bibi, Riccardo Volpi, Amartya Sanyal, Philip Torr, Grégory Rogez, and Puneet Dokania. Make some noise: Reliable and efficient single-step adversarial training. *Advances in Neural Information Processing Systems*, 35:12881–12893, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2299–2308. PMLR, 09–15 Jun 2019.

Zeinab Golgooni, Mehrdad Saberi, Masih Eskandar, and Mohammad Hossein Rohban. Zerograd : Mitigating and explaining catastrophic overfitting in fgsm adversarial training, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Zhengbao He, Tao Li, Sizhe Chen, and Xiaolin Huang. Investigating catastrophic overfitting in fast adversarial training: A self-fitting perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2314–2321, June 2023.

Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Prior-guided adversarial initialization for fast adversarial training. In *European Conference on Computer Vision*, pp. 567–584. Springer, 2022.

Hoki Kim, Woojin Lee, and Jaewook Lee. Understanding catastrophic overfitting in single-step adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8119–8127, 2021. doi: 10.1609/aaai.v35i9.16989.

Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 301–310, 2020.

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

Vishesh Kumar and Akshay Agarwal. Feature collapse under corruption: An entropy perspective on robust neural networks. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *PMLR*, Seoul, South Korea, 2026.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses*, pp. 273–294. Springer, 2018a. doi: 10.1007/978-3-030-00470-9_13.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xinyan Zhang. Trojaning attack on neural networks. In *Proceedings of the 25th Annual Network and Distributed System Security Symposium*. Internet Society, 2018b. doi: 10.14722/ndss.2018.23291.

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.

Nay Myat Min, Long H. Pham, Yige Li, Yige Li, and Jun Sun. Crow: Eliminating backdoors from large language models via internal consistency regularization. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021.

Chao Pan, Qing Li, and Xin Yao. Adversarial initialization with universal adversarial perturbation: A new approach to fast adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21501–21509, 2024.

Chao Pan, Ke Tang, Qing Li, and Xin Yao. Mitigating catastrophic overfitting in fast adversarial training via label information elimination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2991–3000, 2025a.

Chao Pan, Yu Wu, Ke Tang, Qing Li, and Xin Yao. Efficient robustness evaluation via constraint relaxation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6263–6271, 2025b. doi: 10.1609/aaai.v39i6.32670.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R. Venkatesh Babu. Guided adversarial attack for evaluating and enhancing adversarial defenses. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20297–20308, 2020.

Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and Venkatesh Babu R. Towards efficient and effective adversarial training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11821–11833. Curran Associates, Inc., 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks, 2019.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

Yong Xie, Weijie Zheng, Hanxun Huang, Guangnan Ye, and Xingjun Ma. Towards million-scale adversarial robustness evaluation with stronger individual attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30702–30711, June 2025.

Zhiguo Yang, Dongsheng Xu, Ruizhi Zhong, Jiacheng Pi, Xingxing Huang, and Wenjie Ruan. Logit-margin repulsion for backdoor defense. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34918–34928, June 2026.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith (eds.), *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26693–26712. PMLR, 17–23 Jul 2022.

APPENDIX

In this appendix, we provide additional explanations, experiments, and analyses to further support our findings.

A RELATED WORK

A.1 ADVERSARIAL ATTACKS AND ADVERSARIAL TRAINING

Adversarial Attacks. Adversarial attacks add small, bounded perturbations to an input image to cause a well-trained model to misclassify it. The resulting inputs, known as adversarial examples, expose a key vulnerability of deep neural networks: nearly imperceptible perturbations can substantially alter model predictions (Szegedy et al., 2014; Goodfellow et al., 2015). Given an input-label pair $(\boldsymbol{x}, y)$, their generation is formulated as a constrained optimization problem that maximizes the classification loss within a prescribed threat model:

$$\boldsymbol{\delta}^* = \arg \max_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \mathcal{L}\big(f(\boldsymbol{x} + \boldsymbol{\delta}), y\big). \tag{13}$$

Common attacks used in adversarial training include the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). FGSM efficiently generates adversarial examples with a single gradient step (Goodfellow et al., 2015), whereas PGD uses multiple projected gradient updates to produce stronger and more reliable adversarial examples. As a result, PGD has become a standard method for adversarial training. Specifically, PGD iteratively updates adversarial examples as

$$\boldsymbol{x}_{adv}^{t+1} = \Pi_\epsilon \left(\boldsymbol{x}_{adv}^t + \eta \operatorname{sign} \left(\nabla_{\boldsymbol{x}_{adv}^t} \mathcal{L}\big(f(\boldsymbol{x}_{adv}^t), y\big) \right) \right), \tag{14}$$

where η denotes the step size and $\Pi_\epsilon(\cdot)$ projects the updated example onto $\mathcal{B}_\epsilon(\boldsymbol{x})$.

Adversarial Training. Adversarial training improves robustness by optimizing models against adversarially perturbed inputs. Following the robust optimization framework of Madry et al. (Madry et al., 2018), its objective is

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}\big(f(x+\delta), y\big) \right]. \tag{16}$$

PGD-based adversarial training is among the most effective empirical defenses, but its multi-step inner maximization incurs substantial computational cost. Moreover, adversarial training commonly involves a trade-off between clean accuracy and adversarial robustness (Zhang et al., 2019).

A.2 FAST ADVERSARIAL TRAINING AND CATASTROPHIC OVERFITTING

Catastrophic Overfitting. Catastrophic overfitting (CO) is a major limitation of fast adversarial training (FAT). During training, a model may remain robust to the single-step attack used for optimization while its robustness against stronger multi-step attacks suddenly collapses (Wong et al., 2020; de Jorge Aranda et al., 2022; Jia et al., 2022).

Fast Adversarial Training. Fast adversarial training (FAT) reduces this overhead by approximating the inner maximization with a single-step attack. A common choice is FGSM (Szegedy et al., 2014; Goodfellow et al., 2015; Jia et al., 2022), which generates an adversarial example as

$$\boldsymbol{x}_{adv} = \Pi_\epsilon \left[\boldsymbol{x} + \epsilon \operatorname{sign} \left(\nabla_{\boldsymbol{x}} \mathcal{L}\big(f(\boldsymbol{x}), y\big) \right) \right]. \tag{15}$$

Because FGSM requires only a single gradient step, FAT substantially reduces the cost of adversarial example generation. Existing FAT methods for mitigating CO can be broadly categorized into two groups: initialization-based approaches and regularization-based approaches.

Initialization-based fast adversarial training methods use carefully designed initial perturbations to improve the starting points and diversity of single-step attacks, thereby mitigating CO. Given an initial perturbation η, the initial adversarial input is defined as $\boldsymbol{x}_0 = \boldsymbol{x} + \eta$. FGSM then produces

$$\boldsymbol{x}_{adv} = \Pi_\epsilon \left[\boldsymbol{x}_0 + \boldsymbol{\epsilon} \operatorname{sign} \left(\nabla_{\boldsymbol{x}_0} \mathcal{L}\big((\boldsymbol{x}_0), y\big) \right) \right], \tag{16}$$

where η denotes the initial perturbation and $\boldsymbol{\epsilon}$ is the attack step size.

FGSM-RS (Wong et al., 2020) samples a random initial perturbation within the admissible perturbation set before the FGSM update. N-FGSM (de Jorge Aranda et al., 2022) injects stronger random noise around clean inputs and omits clipping of the noisy initialization, which has been shown to mitigate catastrophic overfitting under larger perturbation budgets. FGSM-MEP (Jia et al., 2022) uses historical perturbation information to construct momentum-guided initializations. LIET identifies the transferability of label information and mitigates catastrophic overfitting by disrupting this information with noise (Pan et al., 2025a).

Regularization-based FAT improves the effectiveness of single-step attacks by constraining model behavior near clean or adversarial examples from a local linearity perspective. A representative approach, gradient alignment, encourages input gradients at clean and adversarial examples to have consistent directions (Andriushchenko & Flammarion, 2020). Other methods enforce local linearity or smoothness (Abad Rocamora et al., 2024) or regularize model outputs and feature representations (Zhang et al., 2022). Prior work has also attributed catastrophic overfitting (CO) to self-fitting (He et al., 2023) and unstable small gradients (Golgooni et al., 2021).

A.3 Backdoor Threat Models

Backdoor attacks pose a serious security threat by implanting a hidden association between an attacker-chosen trigger and a target prediction Liu et al. (2018b); Shafahi et al. (2018); Min et al. (2025). Early attacks, such as BadNets Gu et al. (2017), demonstrate that poisoning training samples with fixed triggers can induce targeted misclassification. Backdoor attacks take many forms. Recent studies use spatial transformations Nguyen & Tran (2021), image blending, and learnable or sample-specific triggers to improve attack effectiveness and stealth Chen et al. (2017); Liu et al. (2018a). The common objective of these methods is to preserve normal predictions on clean inputs while inducing a target prediction when the trigger is present.

For a poisoned sample $\mathcal{A}(\boldsymbol{x}, \tau)$, where τ denotes the trigger and $\mathcal{A}$ denotes trigger insertion, the poisoned-training objective can be written as

$$\min_{\theta} \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\mathcal{L}\big(f_\theta(\boldsymbol{x}), y\big)\right] + \lambda \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{poison}}} \left[\mathcal{L}\big(f_\theta(\mathcal{A}(\boldsymbol{x}, \tau)), y_t\big)\right], \tag{17}$$

where y_t is the attacker-specified target label and λ controls the contribution of poisoned samples. Attack performance is typically evaluated using clean accuracy and attack success rate, measuring the preservation of normal behavior and the effectiveness of the trigger-induced association.

B Theoretical Analysis

B.1 Formation of Self-Induced Trigger-Like Patterns

This section first analyzes the class-specific trigger perturbations induced by single-step adversarial training and then explains how our spatial debiasing operations suppress the resulting trigger responses.

For a training pair $(\boldsymbol{x}, y)$, let $\boldsymbol{\delta} = \boldsymbol{x}_{adv} - \boldsymbol{x}$ denote the perturbation generated by the single-step inner maximization. Because $\boldsymbol{\delta}$ is computed from the label-dependent gradient $\nabla_{\boldsymbol{x}}\mathcal{L}(f(\boldsymbol{x}), y)$, perturbations generated from samples of the same class may share common class-conditional structures. For each class $k \in \{1, \dots, K\}$, we define the corresponding class-specific perturbation prototype as $\boldsymbol{\delta}^{(k)} = \mathbb{E}[\boldsymbol{\delta} \mid y = k]$. When the n-th candidate transformation is applied, where $n \in \{1, \dots, N\}$, we denote the resulting perturbation for class k by $\boldsymbol{p}^{(k,n)}$ and the corresponding adversarial input by $\boldsymbol{x}_{adv}^{(k,n)} = \boldsymbol{x} + \boldsymbol{p}^{(k,n)}$.

To characterize the predictive effect of a class-conditional perturbation, we consider the logit margin between the true class k and a competing class $c \neq k$:

$$M^{(k,c)}(\boldsymbol{x}) = f_k(\boldsymbol{x}) - f_c(\boldsymbol{x}). \tag{18}$$

Assume that $M^{(k,c)}$ has an $L^{(k,c)}$-Lipschitz continuous input gradient within $\mathcal{B}_\epsilon(\boldsymbol{x})$. For a perturbation $\boldsymbol{p}^{(k,n)}$ generated for class k using transformation n, a first-order Taylor expansion gives

$$\begin{aligned} &M^{(k,c)}\left(\boldsymbol{x}_{adv}^{(k,n)}\right) - M^{(k,c)}(\boldsymbol{x}) \\ &= \left\langle \nabla_{\boldsymbol{x}} M^{(k,c)}(\boldsymbol{x}), \boldsymbol{p}^{(k,n)} \right\rangle + \mathrm{Rem}_{\mathrm{Tay}}^{(k,c,n)}, \end{aligned} \tag{19}$$

where the remainder satisfies

$$\left|\mathrm{Rem}_{\mathrm{Tay}}^{(k,c,n)}\right| \leq \frac{L^{(k,c)}}{2} \left\|\boldsymbol{p}^{(k,n)}\right\|_2^2. \tag{20}$$

Taking the conditional expectation over samples with label k yields

$$\begin{aligned} &\mathbb{E}\Big[M^{(k,c)}\left(\boldsymbol{x}_{adv}^{(k,n)}\right) - M^{(k,c)}(\boldsymbol{x}) \mid y = k\Big] \\ &= \mathbb{E}\Big[\left\langle \nabla_{\boldsymbol{x}} M^{(k,c)}(\boldsymbol{x}), \boldsymbol{p}^{(k,n)} \right\rangle \mid y = k\Big] + \mathbb{E}\Big[\mathrm{Rem}_{\mathrm{Tay}}^{(k,c,n)} \mid y = k\Big]. \end{aligned} \tag{21}$$

equation 21 thus suggests a first-order mechanism by which repeated optimization may reinforce this shared structure, encouraging the model to exploit self-induced class-specific shortcuts rather than learn genuinely robust features. the emergence of persistent class-specific patterns after CO onset is consistent with this mechanism.

Why Spatial Debiasing Reduces Class-Specific Trigger Responses. The trigger response induced by a class-conditional perturbation is primarily determined by its alignment with a class-specific predictive direction. Spatial debiasing reduces this alignment by redistributing the perturbation structure while preserving its admissible magnitude.

For class k, let $\boldsymbol{\delta}^{(k)}$ denote the corresponding class-wise perturbation prototype. Let $T^{(n)}$ denote the n-th candidate spatial transformation, where $n \in \{1, \ldots, N\}$, sampled according to the class-adaptive policy $\boldsymbol{\pi}^{(k)}$. The transformed prototype is defined as

$$\boldsymbol{p}^{(k,n)} = \mathcal{N}_\epsilon\left(T^{(n)}\left(\boldsymbol{\delta}^{(k)}\right)\right), \tag{22}$$

where $\mathcal{N}_\epsilon$ rescales the transformed perturbation and projects it onto the feasible set

$$\mathcal{S}_\epsilon = \{\boldsymbol{\delta} \mid \|\boldsymbol{\delta}\|_\infty \leq \epsilon\}. \tag{23}$$

For a competing class $c \neq k$, let $M^{(k,c)}(\boldsymbol{x}) = f_k(\boldsymbol{x}) - f_c(\boldsymbol{x})$ denote the logit margin between the true class k and the competing class c. We define the class-conditional predictive alignment of a perturbation $\boldsymbol{\delta}$ as $\mathcal{A}^{(k,c)}(\boldsymbol{\delta}) = \mathbb{E}\big[\langle \nabla_{\boldsymbol{x}} M^{(k,c)}(\boldsymbol{x}), \boldsymbol{\delta}\rangle \mid y = k\big]$.

Assume that the class-adaptive transformation policy satisfies the spatial decorrelation condition

$$\left|\mathbb{E}_{n\sim\boldsymbol{\pi}^{(k)}}\left[\mathcal{A}^{(k,c)}\left(\boldsymbol{p}^{(k,n)}\right)\right]\right| \leq \varrho_{k,c}\left|\mathcal{A}^{(k,c)}\left(\boldsymbol{\delta}^{(k)}\right)\right|. \tag{24}$$

Assume that $M^{(k,c)}$ has an $L^{(k,c)}$-Lipschitz continuous input gradient within $\mathcal{B}_\epsilon(\boldsymbol{x})$. Applying a first-order Taylor expansion to the transformed prototype gives

$$\begin{aligned} &\mathbb{E}\left[M^{(k,c)}\left(\boldsymbol{x} + \boldsymbol{p}^{(k,n)}\right) - M^{(k,c)}(\boldsymbol{x}) \mid y = k\right] \\ &= \mathcal{A}^{(k,c)}\left(\boldsymbol{p}^{(k,n)}\right) + \mathbb{E}\left[\mathrm{Rem}_{\mathrm{Tay}}^{(k,c,n)} \mid y = k\right], \end{aligned} \tag{25}$$

where the Taylor remainder is defined as

$$\mathrm{Rem}_{\mathrm{Tay}}^{(k,c,n)} = M^{(k,c)}\left(\boldsymbol{x} + \boldsymbol{p}^{(k,n)}\right) - M^{(k,c)}(\boldsymbol{x}) - \left\langle \nabla_{\boldsymbol{x}} M^{(k,c)}(\boldsymbol{x}), \boldsymbol{p}^{(k,n)} \right\rangle, \tag{26}$$

and satisfies

$$\left|\mathrm{Rem}_{\mathrm{Tay}}^{(k,c,n)}\right| \leq \frac{L^{(k,c)}}{2} \left\|\boldsymbol{p}^{(k,n)}\right\|_2^2. \tag{27}$$

Taking the expectation over the transformation policy and combining Eqs. equation 24 and equation 25 yields

$$\begin{aligned}&\left|\mathbb{E}_{n\sim\boldsymbol{\pi}^{(k)}}\mathbb{E}\left[M^{(k,c)}\left(\boldsymbol{x}+\boldsymbol{p}^{(k,n)}\right)-M^{(k,c)}(\boldsymbol{x})\mid y=k\right]\right|\\&\leq \varrho_{k,c}\left|\mathcal{A}^{(k,c)}\left(\boldsymbol{\delta}^{(k)}\right)\right|+\frac{L^{(k,c)}}{2}\mathbb{E}_{n\sim\boldsymbol{\pi}^{(k)}}\mathbb{E}\left[\left\|\boldsymbol{p}^{(k,n)}\right\|_2^2\mid y=k\right].\end{aligned}\tag{28}$$

If the transformation preserves the perturbation budget, $\left\|\boldsymbol{p}^{(k,n)}\right\|_\infty\leq\epsilon$, then

$$\frac{L^{(k,c)}}{2}\mathbb{E}_{n\sim\boldsymbol{\pi}^{(k)}}\mathbb{E}\left[\left\|\boldsymbol{p}^{(k,n)}\right\|_2^2\mid y=k\right]\leq\frac{L^{(k,c)}d\epsilon^2}{2},\tag{29}$$

where d is the input dimension. When the original class-specific alignment dominates the second-order remainder, any policy satisfying $\varrho_{k,c}<1$ reduces the bound on the class-conditional margin response. Thus, under the spatial decorrelation condition, spatial debiasing reduces class-specific predictive alignment, which is consistent with weaker hard- and soft-trigger responses.

To isolate perturbation structure from image semantics, we visualize class-wise single-step perturbations on semantically neutral grayscale inputs throughout training (Pan et al., 2025a). Figure 9 and Figure 10 compares existing FAT variants, single-step adversarial baselines and the multi-step PGD-AT at epochs 10, 20, 40, 60, 80, and the final epoch. FAT methods rapidly develop persistent periodic patterns, whereas DC-FGSM progressively disrupts these fixed spatial structures. These observations support our analysis that single-step training can reinforce self-induced class-specific shortcuts, while spatial debiasing reduces their persistence.

Multi-step inner maximization continually adapts perturbations to the evolving loss landscape. In contrast, a single-step attack relies on one local gradient and may produce biased patterns that are subsequently reinforced by the outer minimization.

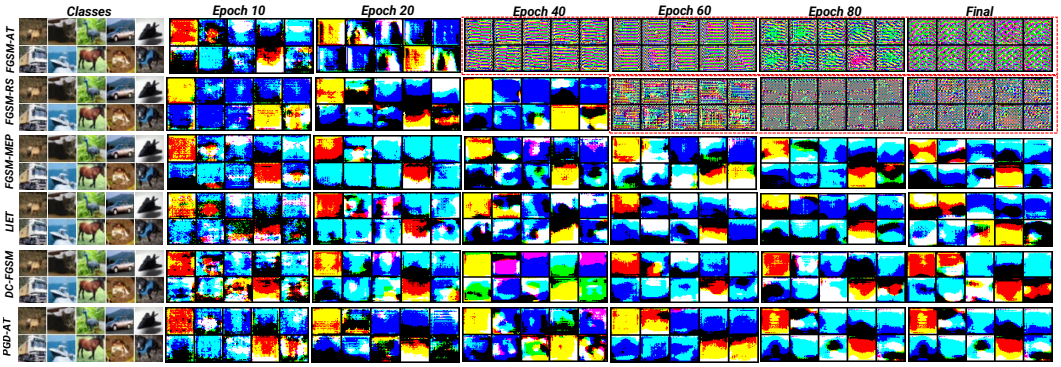

Figure 9: Evolution of class-specific perturbations during training. Red boxes mark class-related patterns emerging after CO under $\epsilon = 8/255$.

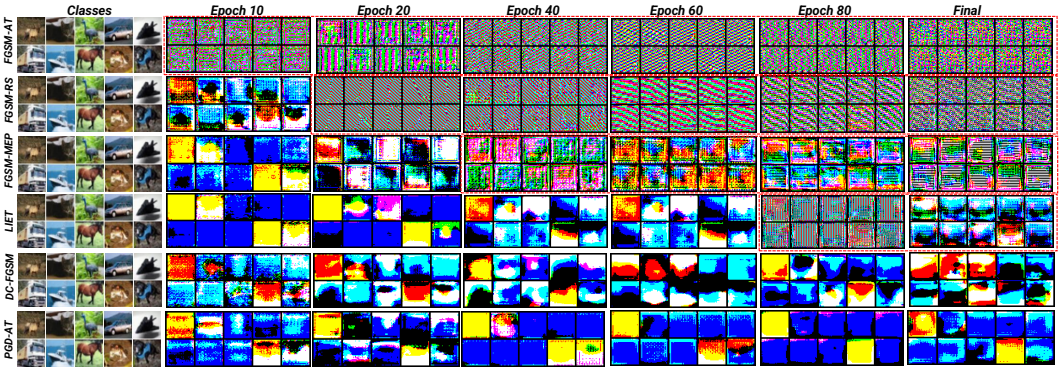

Figure 10: Evolution of class-specific perturbations during training. Red boxes mark class-related patterns emerging after CO under $\epsilon = 16/255$.

B.2 GRADIENT-DRIFT RESIDUAL

This section first analyzes the gradient-drift residual in initialization-based FAT. It then analyzes how a finite-difference correction between gradients at two nearby points incorporates the observed local gradient variation into the subsequent update.

Let $\boldsymbol{\delta}$ be an arbitrary initialization perturbation satisfying $\boldsymbol{x} + \boldsymbol{\delta} \in \mathcal{B}_\epsilon(\boldsymbol{x})$. Define $\boldsymbol{x}_0 = \boldsymbol{x} + \boldsymbol{\delta}$ and $\boldsymbol{g}_0 = \nabla_{\boldsymbol{x}'} \mathcal{L}\big(f(\boldsymbol{x}'), y\big)\big|_{\boldsymbol{x}'=\boldsymbol{x}_0}$. We then take a nearby probing step $\boldsymbol{x}_r = \Pi_\epsilon\big(\boldsymbol{x}_0 + \epsilon_r \operatorname{sign}(\boldsymbol{g}_0)\big)$, where ϵ_r denotes the probing step size (as specified in Section 2.2), and define $\boldsymbol{g}_r = \nabla_{\boldsymbol{x}'} \mathcal{L}\big(f(\boldsymbol{x}'), y\big)\big|_{\boldsymbol{x}'=\boldsymbol{x}_r}$. Let $\boldsymbol{h} = \boldsymbol{x}_r - \boldsymbol{x}_0$ denote the resulting offset. The gradient-drift residual is $\mathcal{R} = \|\boldsymbol{g}_r - \boldsymbol{g}_0\|_2$.

Proposition 1 (Gradient-Drift Residual Decomposition). The gradient-drift residual $\mathcal{R} = \|\boldsymbol{g}_r - \boldsymbol{g}_0\|_2$ satisfies

$$\mathcal{R} \le \big|\|\boldsymbol{g}_r\|_2 - \|\boldsymbol{g}_0\|_2\big| + \sqrt{2\,\|\boldsymbol{g}_r\|_2\,\|\boldsymbol{g}_0\|_2\,[1 - \cos(\boldsymbol{g}_0, \boldsymbol{g}_r)]}. \tag{30}$$

Proof. The squared residual admits the exact decomposition

$$\mathcal{R}^2 = \|\boldsymbol{g}_r - \boldsymbol{g}_0\|_2^2 = \big(\|\boldsymbol{g}_r\|_2 - \|\boldsymbol{g}_0\|_2\big)^2 + 2\,\|\boldsymbol{g}_r\|_2\,\|\boldsymbol{g}_0\|_2\,[1 - \cos(\boldsymbol{g}_0, \boldsymbol{g}_r)]\,. \tag{31}$$

The first term measures the change in gradient magnitude, whereas the second term measures the change in gradient direction. Applying $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$ yields equation 30. □

The directional component has an exact geometric interpretation:

$$\left\| \frac{\boldsymbol{g}_r}{\|\boldsymbol{g}_r\|_2} - \frac{\boldsymbol{g}_0}{\|\boldsymbol{g}_0\|_2} \right\|_2^2 = 2\,[1 - \cos(\boldsymbol{g}_0, \boldsymbol{g}_r)]\,. \tag{32}$$

Thus, $1 - \cos(\boldsymbol{g}_0, \boldsymbol{g}_r)$ is not merely a heuristic drift indicator, but exactly quantifies the squared Euclidean distance between the normalized initialization and probing directions. Because sign-based updates are invariant to uniform positive rescaling, this directional component is more directly related to the change in the single-step update path than the magnitude component.

Proposition 2 (Curvature-Induced Gradient Drift). Let $\mathcal{L}(\boldsymbol{x}') = \mathcal{L}(f(\boldsymbol{x}'), y)$, and assume that $\mathcal{L}$ is twice continuously differentiable on the line segment connecting $\boldsymbol{x}_0$ and $\boldsymbol{x}_r$. Let $\boldsymbol{h} = \boldsymbol{x}_r - \boldsymbol{x}_0$. Then

$$\boldsymbol{g}_r - \boldsymbol{g}_0 = \left[\int_0^1 \mathbf{H}\,(\boldsymbol{x}_0 + t\boldsymbol{h})\ \mathrm{d}t \right] \boldsymbol{h}, \tag{33}$$

where $\mathbf{H}(\boldsymbol{x}') = \nabla^2_{\boldsymbol{x}'} \mathcal{L}(\boldsymbol{x}')$.

Proof. Applying the fundamental theorem of calculus to the gradient field along the path $\boldsymbol{x}_0 + t\boldsymbol{h}$ gives

$$\boldsymbol{g}_r - \boldsymbol{g}_0 = \nabla \mathcal{L}\,(\boldsymbol{x}_0 + \boldsymbol{h}) - \nabla \mathcal{L}\,(\boldsymbol{x}_0) = \int_0^1 \nabla^2 \mathcal{L}\,(\boldsymbol{x}_0 + t\boldsymbol{h})\,\boldsymbol{h}\,\mathrm{d}t. \tag{34}$$

□

equation 33 shows that gradient drift is the accumulated action of the local Hessian along the probing direction. In particular, if

$$\sup_{\boldsymbol{z} \in [\boldsymbol{x}_0, \boldsymbol{x}_r]} \|\mathbf{H}(\boldsymbol{z})\|_2 \le H,$$

then

$$\mathcal{R} \le H\,\|\boldsymbol{h}\|_2\,. \tag{35}$$

Therefore, larger probing displacements or stronger local curvature can increase the upper bound on the deviation from the initialization gradient. This provides one explanation for why the initialization gradient may become less reliable as the perturbation budget increases.

Proposition 3 (Residual Reduction by Gradient Correction). Define the corrected gradient by

$$\widehat{\boldsymbol{g}} = \boldsymbol{g}_0 + \lambda(\boldsymbol{g}_r - \boldsymbol{g}_0), \qquad \lambda \in [0, 1]. \tag{36}$$

Then the residual of the corrected gradient with respect to the probing gradient satisfies

$$\|\boldsymbol{g}_r - \widehat{\boldsymbol{g}}\|_2 = (1 - \lambda)\mathcal{R}. \tag{37}$$

Proof. By substituting equation 36, we obtain

$$\boldsymbol{g}_r - \widehat{\boldsymbol{g}} = (1-\lambda)(\boldsymbol{g}_r - \boldsymbol{g}_0). \tag{38}$$

Taking the ℓ_2 norm gives equation 37. Thus, the correction coefficient λ directly controls the fraction of the observed local gradient drift that is incorporated into the update. When $\lambda = 0$, the corrected gradient reduces to the initialization gradient; when $\lambda = 1$, it equals the probing gradient. Intermediate values yield a curvature-aware interpolation without explicitly computing the Hessian.

Figures 11 and 12 visualize gradient drift and adversarial accuracy across FAT methods at $\epsilon = 8/255$ and $16/255$, showing epoch-level trends (left) and batch-level behavior before and after CO (middle, right). Methods undergoing CO exhibit volatile post-CO drift and sharp robust-accuracy degradation despite high FGSM accuracy, whereas DC-FGSM maintains smooth drift and stable robust accuracy, especially at the larger budget.

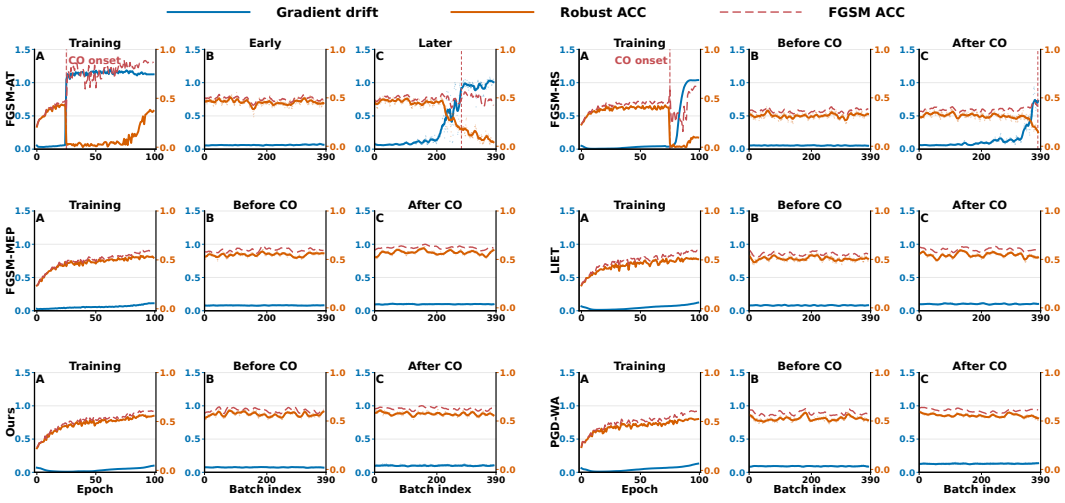

Figure 11: Gradient-drift and robust-accuracy dynamics across FAT methods at $\epsilon = 8/255$. From left to right: epoch-level evolution, batch-level behavior before CO, and batch-level behavior after CO.

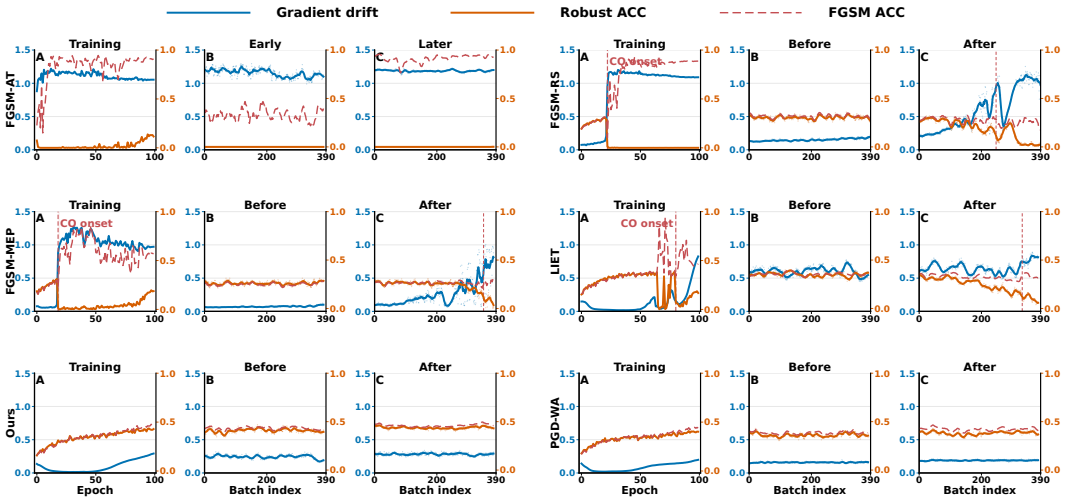

Figure 12: Gradient-drift and robust-accuracy dynamics across FAT methods at $\epsilon = 16/255$. From left to right: epoch-level evolution, batch-level behavior before CO, and batch-level behavior after CO.

C FULL ALGORITHM

DC-FGSM combines class-adaptive spatial debiasing and drift-aware gradient correction within a unified single-step training framework. Algorithm 1 summarizes the procedure. For each minibatch, a class-wise prior initializes the adversarial examples, while a stochastic policy selects spatial transformations based on their accumulated utilities. The selected transformations update the class-wise priors online. After the correction stage begins, a nearby probing step estimates local gradient drift and adjusts the final update direction. Before the correction stage begins, the original gradient is used. The resulting adversarial examples jointly optimize the model with a clean–adversarial consistency regularizer.

Algorithm 1 Debiasing and Correction Adversarial Training

Input: Dataset $\mathcal{D}$, model f_θ, perturbation budget ϵ, number of epochs E, correction start epoch E_c, transformation bank $\mathcal{T}$, policy parameters (ρ, ω, ξ), momentum coefficient μ, probing step size ϵ_r, correction coefficient λ, and consistency coefficient β

Output: Robust model f_θ

Initialize the class-wise buffers $\{\mathcal{P}^{(k)}\}_{k=1}^{K}$, historical priors $\{\boldsymbol{m}^{(k)}\}_{k=1}^{K}$, utilities $\{u^{(k,n)}\}$, and policies $\{\boldsymbol{\pi}^{(k)}\}_{k=1}^{K}$
for $e = 1, \ldots, E$ **do**
  for each minibatch $(\boldsymbol{x}, y) \sim \mathcal{D}$ **do**
    $\tilde{\boldsymbol{x}} \leftarrow \Pi_\epsilon\left(\boldsymbol{x} + \epsilon \operatorname{sign}\left(\boldsymbol{m}^{(y)}\right)\right)$ ▷ Class-adaptive debiased initialization
    $\tilde{\boldsymbol{g}} \leftarrow \nabla_{\tilde{\boldsymbol{x}}} \mathcal{L}\big(f(\tilde{\boldsymbol{x}}), y\big)$
    if $e \geq E_c$ **then**
      $\boldsymbol{x}_r \leftarrow \Pi_\epsilon\left(\tilde{\boldsymbol{x}} + \epsilon_r \operatorname{sign}(\tilde{\boldsymbol{g}})\right)$
      $\boldsymbol{g}_r \leftarrow \nabla_{\boldsymbol{x}_r} \mathcal{L}\big(f(\boldsymbol{x}_r), y\big)$
      $\widehat{\boldsymbol{g}} \leftarrow \tilde{\boldsymbol{g}} + \lambda\left(\boldsymbol{g}_r - \tilde{\boldsymbol{g}}\right)$ ▷ Drift-aware gradient correction
    else
      $\widehat{\boldsymbol{g}} \leftarrow \tilde{\boldsymbol{g}}$
    end if
    $\tilde{\boldsymbol{x}}_{adv} \leftarrow \Pi_\epsilon\left(\tilde{\boldsymbol{x}} + \epsilon \operatorname{sign}(\widehat{\boldsymbol{g}})\right)$
    for each class k present in the minibatch **do**
      Compute the class-wise statistics:

$$\boldsymbol{\delta}^{(k)} \leftarrow \operatorname{Mean}\left(\tilde{\boldsymbol{x}}_{adv} - \boldsymbol{x} \,|\, y = k\right),$$
$$\boldsymbol{g}^{(k)} \leftarrow \operatorname{Mean}\left(\tilde{\boldsymbol{g}} \,|\, y = k\right).$$

      if policy update step **then**
        for $n = 1, \ldots, N$ **do**
          $\boldsymbol{p}^{(k,n)} \leftarrow \mathcal{N}_\epsilon\left(T^{(n)}\left(\boldsymbol{\delta}^{(k)}\right)\right)$
          $q^{(k,n)} \leftarrow \dfrac{\left\langle \boldsymbol{g}^{(k)}, \boldsymbol{p}^{(k,n)} - \bar{\boldsymbol{p}}^{(k)} \right\rangle}{\left\|\boldsymbol{g}^{(k)}\right\|_1}$
        end for
        $\widetilde{q}^{(k,n)} \leftarrow \dfrac{q^{(k,n)} - \mu_q^{(k)}}{\sigma_q^{(k)}}$
        $u^{(k,n)} \leftarrow (1-\rho) u^{(k,n)} + \rho \widetilde{q}^{(k,n)}$ ▷ Update the class-adaptive policy
        $\pi^{(k,n)} \leftarrow (1-\xi)\dfrac{\exp\left(u^{(k,n)}/\omega\right)}{\sum_{j=1}^{N} \exp\left(u^{(k,j)}/\omega\right)} + \dfrac{\xi}{N}$
      end if
      Sample $n_k \sim \boldsymbol{\pi}^{(k)}$ and insert $\mathcal{N}_\epsilon\left(T^{(n_k)}\left(\boldsymbol{\delta}^{(k)}\right)\right)$ into $\mathcal{P}^{(k)}$
      $\boldsymbol{m}^{(k)} \leftarrow \mu \boldsymbol{m}^{(k)} + (1-\mu) \operatorname{sign}\left(\operatorname{Mean}\left(\mathcal{P}^{(k)}\right)\right)$ ▷ Update the class-wise prior
    end for
    Update f_θ using $\mathcal{L}_{\text{total}} = \mathcal{L}\big(f(\tilde{\boldsymbol{x}}_{adv}), y\big) + \beta \mathcal{L}_{\text{JSD}}\big(f(\tilde{\boldsymbol{x}}), f(\tilde{\boldsymbol{x}}_{adv})\big)$
  end for
end for
  return f_θ

D COMPUTATIONAL COMPLEXITY

As shown in Figure 13, the two components of DC-FGSM incur different computational overheads. Debiasing initialization adds negligible training time, as it applies lightweight masking, replacement, and spatial transformations directly to the stored perturbation prior without additional model evaluations or forward/backward passes.

In contrast, gradient correction requires additional gradient probing and incurs modest computational overhead. Nevertheless, DC-FGSM provides a favorable robustness–efficiency trade-off. As shown in Fig. 13, DC-FGSM requires 62.08 minutes on CIFAR-10, compared with approximately 190 minutes for PGD-AT-WA, yielding a 67.3% reduction in training time while achieving higher robust accuracy. Its training cost is also comparable to, and in some cases lower than, that of classical single-step baselines, despite providing substantially stronger robustness. The same trend holds on CIFAR-100 and Tiny-ImageNet, showing that DC-FGSM achieves competitive robustness with substantially lower training cost than multi-step training.

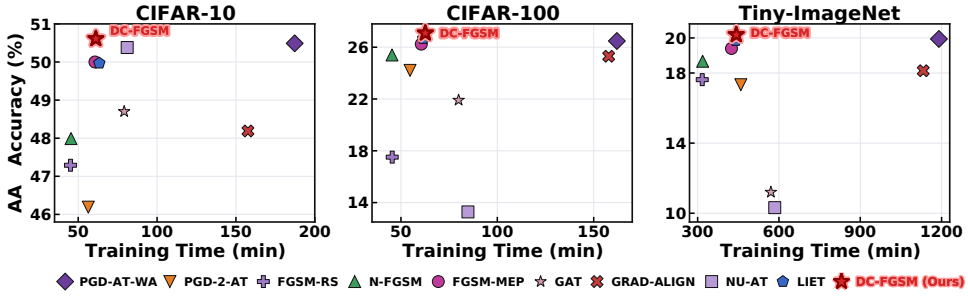

Figure 13: Comparison of training time and adversarial robustness on CIFAR-10, CIFAR-100, and Tiny-ImageNet under $\epsilon = 8/255$. Robustness is measured by AutoAttack (AA) accuracy at the Best checkpoint. A method closer to the upper-left corner achieves a better trade-off between training efficiency and robustness.

Comparison with Two-Step Adversarial Training. To determine whether DC-FGSM's performance gains stem from its core mechanisms rather than simply from additional gradient steps, we compare it with PGD-2. PGD-2 uses two-step adversarial training but does not incorporate DC-FGSM's debiasing initialization or gradient-residual correction. Figure 14 reports robust accuracy under PGD-50, CW, AutoAttack, PMA, and CR attacks at both perturbation budgets. DC-FGSM consistently outperforms PGD-2 across all attack types and checkpoint settings, including the more challenging $\epsilon = 16/255$ setting. These results indicate that DC-FGSM's gains primarily arise from its proposed correction mechanisms rather than simply from using more optimization steps.

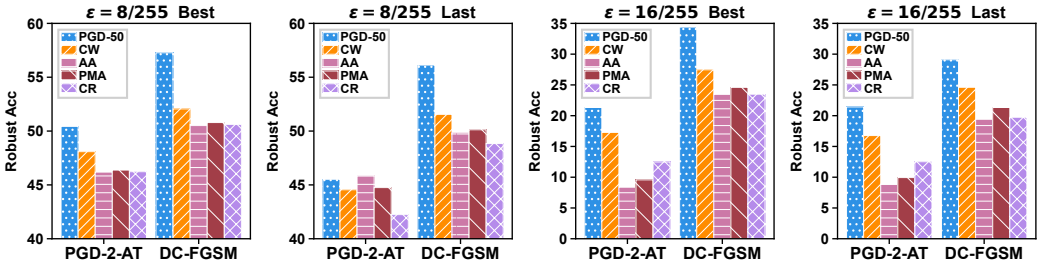

Figure 14: Robust accuracy comparison between PGD-2 and DC-FGSM under PGD-50, CW, AutoAttack (AA), PMA, and CR attacks at $\epsilon = 8/255$ and $\epsilon = 16/255$, evaluated at the best and final checkpoints.

Computational Overhead: Full-Stage vs. Late-Stage Correction. Gradient correction adds one extra forward–backward pass per mini-batch. We therefore compare full-stage correction, applied throughout training, with late-stage correction applied only during the final 50 epochs ($E_c = 50$). As shown in Fig. 15, late-stage correction closely matches, and in some cases exceeds, full-stage correction in mean robustness across the five evaluated attacks at both perturbation budgets and

checkpoint selections. This indicates that gradient correction is most beneficial during later training, when gradient drift becomes more pronounced.

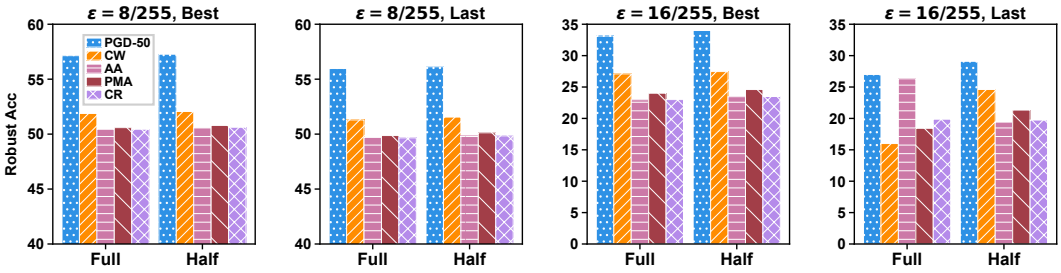

Figure 15: Comparison of full-stage and late-stage gradient correction under $\epsilon = 8/255$ and $\epsilon = 16/255$, evaluated at the best and final checkpoints. Late-stage correction is applied only during the final 50 epochs ($E_c = 50$) and achieves robustness comparable to full-stage correction across the evaluated attacks.

E HYPERPARAMETER SENSITIVITY ANALYSIS

We study the sensitivity of DC-FGSM to the probe distance r and correction coefficient λ using PGD-50 robust accuracy. Figure 16 shows the results at the best and final checkpoints. Performance remains stable across a broad range of hyperparameter values, with the strongest results generally obtained at moderate probe distances and correction coefficients. The selected setting, marked by the dashed boxes, lies within this stable high-performing region and maintains comparable performance at both checkpoints.

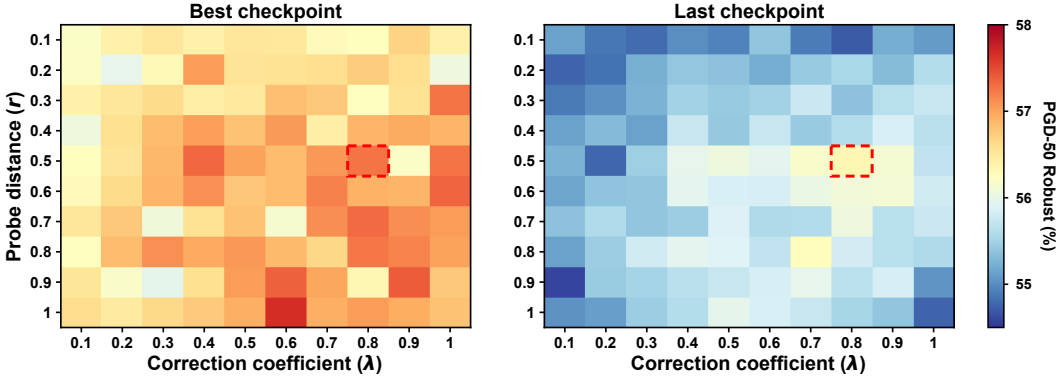

Figure 16: Sensitivity analysis of the hyperparameters r and λ.

F LAYER-WISE ENTROPY ANALYSIS OF BACKDOOR-LIKE BEHAVIOR IN FAT

This section provides layer-wise entropy evidence for the first mechanism and examines whether the resulting clean–adversarial feature separation is mitigated by DC-FGSM.

Layer-wise entropy serves as a coarse proxy for the diversity of feature activations at each depth. Higher entropy indicates a more distributed activation response, whereas lower entropy indicates a more concentrated response. Comparing clean and perturbed inputs therefore reveals whether they induce similar activation distributions or are processed through distinct feature patterns. A pronounced entropy gap is consistent with perturbation-induced feature separation.

Figure 17 shows that CO and backdoored models develop increasingly separated entropy profiles, supporting the view that CO induces a trigger-like processing pathway. In contrast, Fig. 19 shows that DC-FGSM maintains more consistent entropy profiles between clean and FGSM-adversarial examples, closely resembling multi-step PGD-AT-WA. This result is consistent with reduced perturbation-induced feature separation and supports the role of spatial debiasing in reducing self-induced trigger-like behavior. All experiments are conducted on CIFAR-10 with a perturbation budget of $\epsilon = 16/255$. At each checkpoint, all methods are evaluated on the same fixed subset of test images. For a clean image $\boldsymbol{x}$ with label y, we generate its FGSM counterpart using the current model: $\boldsymbol{x}_{adv} = \Pi_\epsilon \left[\boldsymbol{x} + \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} \mathcal{L}\big(f(\boldsymbol{x}), y\big)\right)\right]$. We then extract activations from representative shallow, intermediate, and deep layers for both $\boldsymbol{x}$ and $\boldsymbol{x}_{adv}$. For layer l, we flatten the activation tensor into a feature vector $\boldsymbol{a}_l(\boldsymbol{x})$ and normalize it into a probability vector. The corresponding probability vector and layer-wise entropy are defined as

$$\boldsymbol{h}_l(\boldsymbol{x}) = \operatorname{softmax}\big(\boldsymbol{a}_l(\boldsymbol{x})\big), \quad H_l(\boldsymbol{x}) = -\sum_j h_{l,j}(\boldsymbol{x}) \log h_{l,j}(\boldsymbol{x}). \tag{39}$$

Each row in the figures corresponds to a monitored epoch, and each column corresponds to a selected layer. Blue circles denote clean examples, whereas red triangles denote their FGSM counterparts. The horizontal axis indexes the fixed evaluation subset, and the vertical axis reports the layer-wise entropy $H_l(\boldsymbol{x})$.

A large clean–FGSM entropy gap indicates that the perturbation substantially changes the distribution of layer activations, whereas a small gap is consistent with more similar processing complexity at that layer. In the context of our analysis, the former provides evidence of perturbation-induced feature separation that may accompany trigger-like behavior, while the latter is consistent with more stable clean–adversarial processing. Because entropy is a marginal statistic, we interpret it as evidence of feature separation rather than a complete measure of representation similarity.

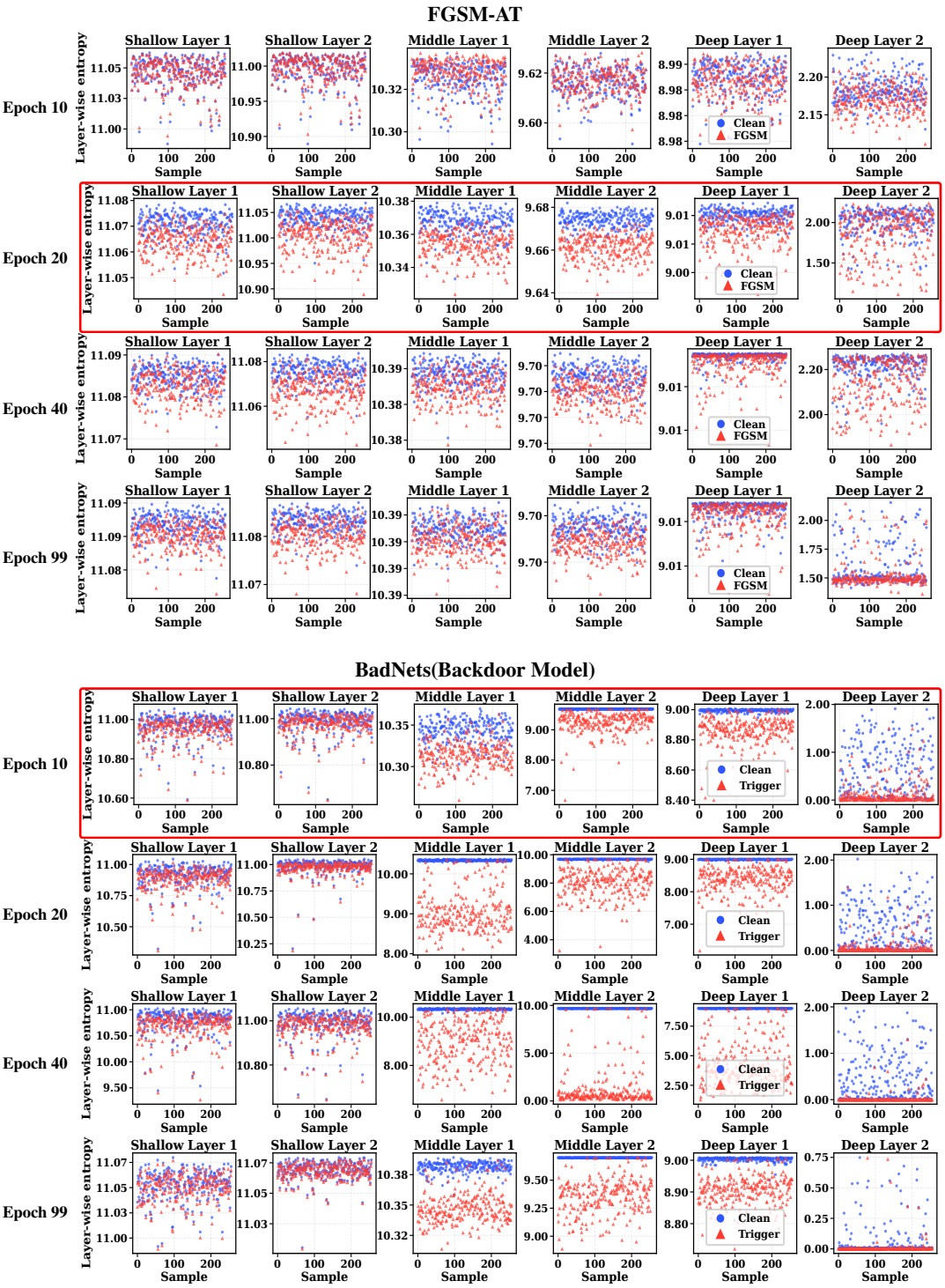

Figure 17: Layer-wise entropy of clean and FGSM-adversarial examples under FGSM-AT, and clean and trigger-embedded examples in the backdoor model on CIFAR-10. Red boxes highlight epochs with pronounced entropy-profile separation.

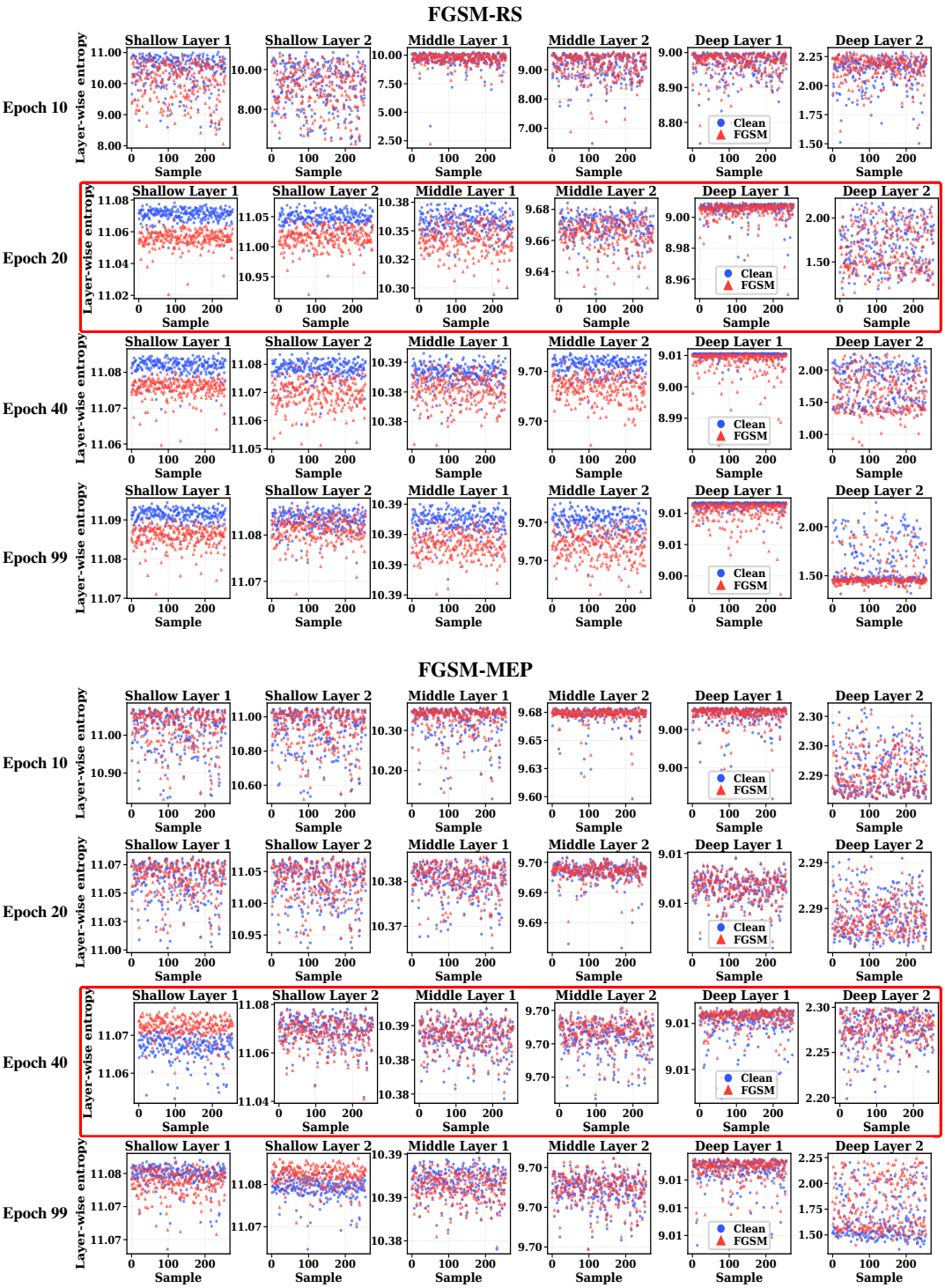

Figure 18: Layer-wise entropy dynamics of clean and FGSM-adversarial examples under FGSM-RS and FGSM-MEP on CIFAR-10. Red boxes mark epochs with pronounced separation between the clean and adversarial entropy profiles.

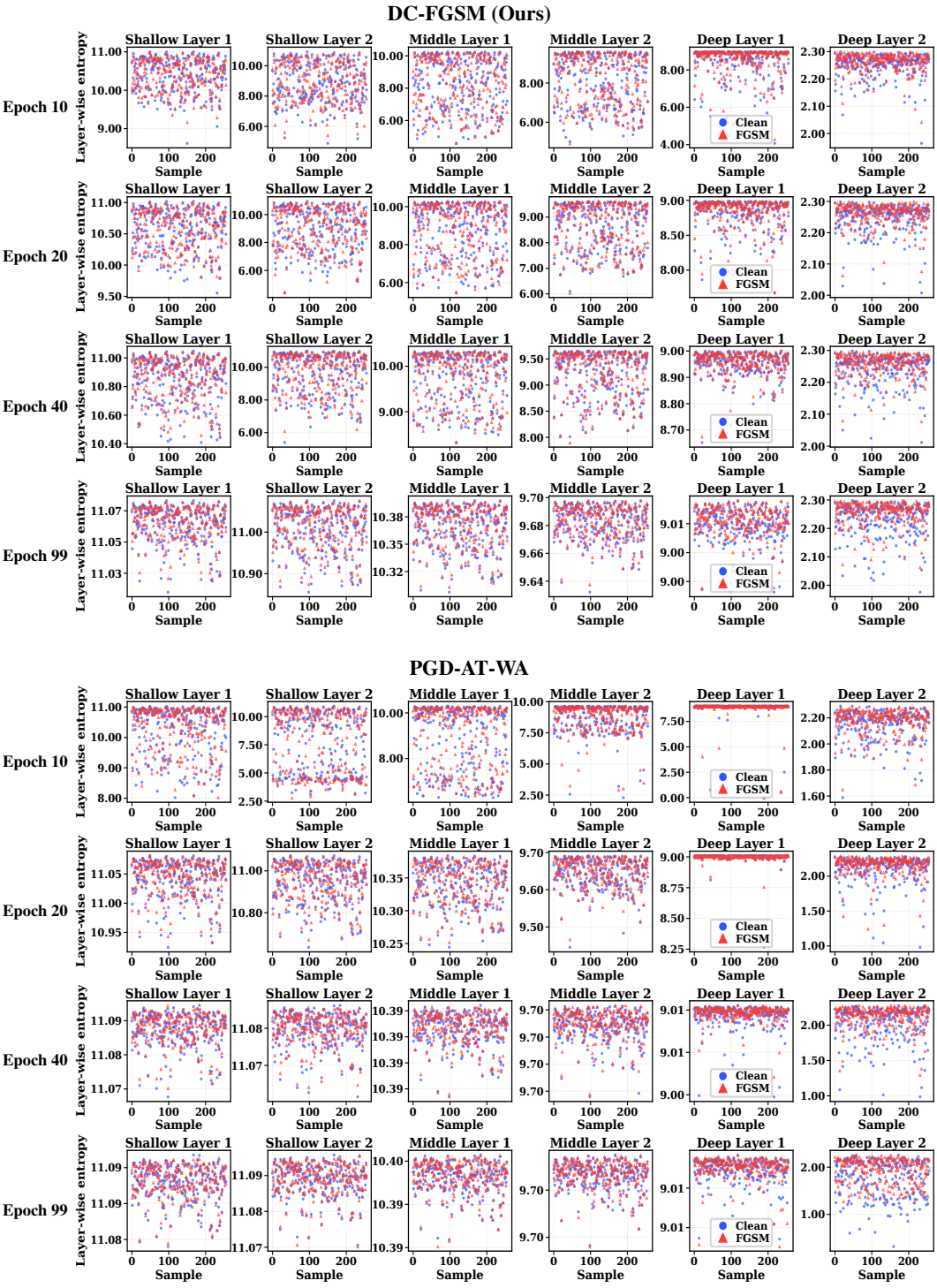

Figure 19: Layer-wise entropy dynamics of clean and FGSM adversarial examples under DC-FGSM (Ours) and PGD-AT-WA on CIFAR-10. DC-FGSM maintains stable clean-adversarial entropy profiles across shallow, middle, and deep layers, showing a response pattern comparable to multi-step adversarial training with weight averaging. This suggests that the proposed debiasing and correction strategy can stabilize intermediate feature representations within an efficient fast adversarial training framework.

G ADDITIONAL EXPERIMENTAL RESULTS

This section presents comprehensive supplementary evaluations, including results under standard white-box attacks such as 10 and 20 step PGD.

G.1 EXPERIMENTAL SETTINGS

This section details the training configurations used in our experiments, including the datasets, model architectures, optimization schedules, perturbation budgets, and evaluation protocols.

Training setup. Unless stated otherwise, all models are trained for 100 epochs using SGD with an initial learning rate of 0.1, momentum of 0.9, weight decay of 5×10^{-4}, and a batch size of 128. The learning rate is adjusted using OneCycleLR. Standard data augmentation consists of four-pixel padding followed by random cropping and random horizontal flipping, and is applied only to clean samples. Label smoothing is consistently applied to all baseline methods. We use weight averaging with 0.9995. Each experiment is repeated with three random seeds.

Model selection is based on validation PGD-10 accuracy. The checkpoint with the highest validation PGD-10 accuracy is reported as *Best*, whereas the checkpoint from the final training epoch is reported as *Last*. Both checkpoints are evaluated using identical attack protocols. CIFAR-10 and CIFAR-100 use 32×32 images with 49,000 training images, whereas Tiny ImageNet uses 64×64 images with 98,000 training images. CIFAR-10 and CIFAR-100 are evaluated under $\epsilon = 8/255$ and $\epsilon = 16/255$, while Tiny ImageNet is evaluated under $\epsilon = 8/255$ and $\epsilon = 12/255$.

Policy update strategy. The policy considers four spatial transformations, initialized with equal probabilities of 25%: crop-and-resize using a patch of $3/4$ of the input resolution, translation by at most $1/8$ of the image size along each axis, rotation within $[-15°, 15°]$, and random shuffling of a 2×2 block grid. During training, operation selection combines 90% learned class-wise probabilities with 10% uniform exploration. To improve perturbation diversity, we replace 10%–50% of the entries with values uniformly sampled from $[-\epsilon, \epsilon]$.

Debiasing initialization. The momentum coefficient for the historical perturbation prior is set to 0.9995, and the mask ratio is uniformly sampled from $[0.1, 0.5]$. The patch size is set to 16 for CIFAR-10 and CIFAR-100 and to 32 for Tiny ImageNet.

Gradient correction. The gradient correction module uses the same configuration across datasets within each perturbation-budget regime. For $\epsilon = 8/255$, the probing radius and correction strength are set to $r = 0.6\epsilon$ and $\lambda = 0.8$, respectively. For the larger-budget settings, namely $\epsilon = 16/255$ for CIFAR-10 and CIFAR-100 and $\epsilon = 12/255$ for Tiny ImageNet, we use the more conservative configuration $r = 0.4\epsilon$ and $\lambda = 0.5$. This design reduces overly aggressive corrections and improves training stability under stronger perturbations.

Robustness evaluation. Robustness is evaluated using PGD, AutoAttack (AA), CW, CR, and PMA. PGD is run for 10, 20, and 50 iterations, denoted as PGD-10, PGD-20, and PGD-50, respectively. All baseline methods use the same preprocessing, optimization settings, checkpoint-selection protocol, and attack configurations.

Architectural experiments. To evaluate generalization across architectures, we use the same training protocol and DC-FGSM hyperparameters unless stated otherwise. ResNet-34 and WideResNet-28-10 are evaluated on CIFAR-10 and CIFAR-100 to determine whether the performance gains remain consistent across deeper and wider network architectures.

G.2 MULTI-ATTACK ROBUSTNESS AND MULTI-SEED STABILITY

Additional Results on CIFAR-10. Table 4 reports the complete evaluation on CIFAR-10 under $\epsilon = 8/255$, including the Best and Last checkpoints. DC-FGSM achieves the highest average accuracy among single-step methods at both checkpoints, reaching 57.18% and 56.70%, respectively, outperforming the strongest competing baselines by 0.80 and 0.89 percentage points. At the Best checkpoint, DC-FGSM achieves the highest accuracy across all seven robust evaluation metrics. At the Last checkpoint, it again attains the best single-step average, exceeding FGSM-MEP and LIET by 1.70 and 0.89 percentage points, respectively. The relatively small standard deviations over three random seeds further support the training stability of DC-FGSM.

Additional Results on CIFAR-100. For completeness, Table 5 reports the full CIFAR-100 evaluation under $\epsilon = 8/255$ for both the Best and Last checkpoints, including mean $\pm$ standard deviation over three random seeds. DC-FGSM achieves the highest single-step average at both checkpoints, reaching 32.89% and 32.73%, respectively. It also attains the best robust accuracy across all evaluated attacks at the Best checkpoint and remains competitive at the Last checkpoint. The relatively small standard deviations further indicate reproducible and stable performance on this multi-class dataset.

Additional Results on Tiny-ImageNet. Table 6 reports complete results under $\epsilon = 8/255$ for both the Best and Last checkpoints. DC-FGSM achieves the highest single-step average accuracy at both checkpoints, reaching 26.02% and 25.97%, respectively. At the Best checkpoint, it exceeds the strongest single-step baseline by 0.26 percentage points and improves PGD-50 and AutoAttack accuracy by 0.67 and 0.33 points, respectively. At the Last checkpoint, it retains the best single-step average, with corresponding PGD-50 and AutoAttack gains of 0.54 and 0.66 points. The relatively small standard deviations over three random seeds further indicate reproducible and stable performance on Tiny-ImageNet.

Table 4: Clean and robust accuracy (%) on CIFAR-10 under $\epsilon = 8/255$. Avg. denotes the average of clean and robust accuracy. Best single-step results are shown in **bold red**.

Method	CIFAR-10								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
					$\epsilon = 8/255$, Best checkpoint				
PGD-AT-WA***	82.15	55.16	54.49	54.36	52.34	50.49	50.54	50.71	56.28
Std.	±0.52	±0.10	±0.12	±0.10	±0.20	±0.20	±0.21	±0.24	±0.21
FGSM-RS	79.94	53.55	52.93	52.74	49.56	47.29	47.36	47.48	53.86
Std.	±1.06	±0.21	±0.08	±0.13	±0.32	±0.30	±0.26	±0.29	±0.28
N-FGSM	79.88	53.67	53.14	53.02	50.13	47.99	48.09	48.20	54.27
Std.	±1.15	±0.07	±0.16	±0.21	±0.17	±0.07	±0.08	±0.11	±0.16
FGSM-MEP	79.77	56.34	55.82	55.72	51.47	50.00	49.98	50.18	56.16
Std.	±0.99	±0.23	±0.13	±0.09	±0.26	±0.16	±0.17	±0.16	±0.25
GAT	85.07	54.42	53.53	53.36	50.35	48.70	48.66	48.87	55.37
Std.	±0.08	±0.36	±0.17	±0.13	±0.25	±0.20	±0.20	±0.22	±0.16
GRAD-ALIGN	80.11	53.56	52.90	52.75	50.30	48.19	48.24	48.39	54.31
Std.	±1.30	±0.17	±0.18	±0.20	±0.37	±0.34	±0.32	±0.39	±0.35
NU-AT	80.91	55.49	54.86	54.74	51.81	50.38	50.36	50.49	56.13
Std.	±0.47	±0.16	±0.32	±0.34	±0.24	±0.33	±0.30	±0.33	±0.19
LIET	80.22	56.82	56.31	56.22	51.44	49.97	49.93	50.13	56.38
Std.	±0.59	±0.13	±0.06	±0.11	±0.20	±0.17	±0.17	±0.18	±0.17
DC-FGSM (Ours)	80.96	**57.85**	**57.30**	**57.27**	**52.06**	**50.61**	**50.62**	**50.79**	**57.18**
Std.	±0.80	±0.31	±0.12	±0.15	±0.16	±0.12	±0.11	±0.14	±0.17
					$\epsilon = 8/255$, Last checkpoint				
PGD-AT-WA***	85.17	52.30	50.90	50.55	49.45	47.44	47.60	47.62	53.88
Std.	±0.11	±0.18	±0.18	±0.15	±0.11	±0.00	±0.01	±0.03	±0.10
FGSM-RS	90.84	30.78	20.93	12.80	9.27	0.00	10.74	18.36	24.22
Std.	±0.59	±4.36	±4.48	±4.04	±2.41	±0.00	±3.13	±4.77	±2.80
N-FGSM	83.62	49.35	47.84	47.32	46.41	43.86	43.87	44.13	50.80
Std.	±0.08	±0.18	±0.21	±0.21	±0.09	±0.23	±0.24	±0.19	±0.14
FGSM-MEP	83.14	54.23	53.31	53.14	50.42	48.51	48.50	48.74	55.00
Std.	±0.23	±0.11	±0.14	±0.18	±0.25	±0.01	±0.01	±0.05	±0.04
GAT	85.33	53.43	52.19	51.98	49.66	47.62	47.54	47.81	54.45
Std.	±0.15	±0.13	±0.10	±0.11	±0.02	±0.00	±0.03	±0.05	±0.03
GRAD-ALIGN	85.10	48.73	47.02	46.55	45.99	43.47	43.62	43.80	50.54
Std.	±0.05	±0.38	±0.25	±0.26	±0.15	±0.10	±0.07	±0.11	±0.16
NU-AT	81.96	54.90	54.03	53.91	51.22	49.56	49.55	49.69	55.60
Std.	±0.07	±0.12	±0.22	±0.22	±0.18	±0.14	±0.13	±0.12	±0.11
LIET	82.71	55.81	54.96	54.74	50.77	49.08	49.09	49.31	55.81
Std.	±0.03	±0.25	±0.36	±0.34	±0.21	±0.08	±0.13	±0.08	±0.18
DC-FGSM (Ours)	82.71	**57.10**	**56.28**	**56.13**	**51.55**	**49.85**	**49.84**	**50.12**	**56.70**
Std.	±0.14	±0.21	±0.25	±0.27	±0.13	±0.14	±0.14	±0.08	±0.14

Table 5: Clean and robust accuracy (%) on CIFAR-10 under $\epsilon = 8/255$. *Avg.* denotes the average of clean and robust accuracy. Best single-step results are shown in **bold red**.

Method	CIFAR-100								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
					$\epsilon = 8/255$, Best checkpoint				
PGD-AT-WA***	58.23	31.91	31.51	31.38	28.33	26.48	26.48	26.74	32.63
Std.	±0.41	±0.13	±0.18	±0.19	±0.10	±0.14	±0.15	±0.16	±0.08
FGSM-RS	46.02	22.56	22.28	22.24	19.37	17.51	17.48	17.72	23.15
Std.	±2.15	±1.17	±1.11	±1.10	±1.07	±0.87	±0.89	±0.86	±1.15
N-FGSM	54.48	30.95	30.68	30.66	27.51	25.40	25.44	25.66	31.35
Std.	±0.21	±0.17	±0.17	±0.14	±0.31	±0.20	±0.23	±0.22	±0.19
FGSM-MEP	56.75	32.11	31.79	31.71	27.92	26.24	26.26	26.50	32.41
Std.	±0.80	±0.17	±0.16	±0.16	±0.08	±0.05	±0.04	±0.07	±0.10
GAT	65.73	27.36	26.48	26.32	23.81	21.91	21.84	22.09	29.44
Std.	±0.45	±0.22	±0.25	±0.25	±0.29	±0.19	±0.20	±0.26	±0.15
GRAD-ALIGN	55.09	31.25	30.94	30.85	27.47	25.30	25.29	25.58	31.47
Std.	±1.67	±0.13	±0.14	±0.14	±0.36	±0.35	±0.31	±0.33	±0.41
NU-AT	57.92	25.56	22.41	20.77	19.26	13.28	16.72	16.96	24.11
Std.	±2.28	±0.56	±0.33	±0.39	±0.19	±0.66	±0.68	±0.24	±0.34
LIET	51.38	32.72	32.54	32.49	27.74	26.67	26.68	26.93	32.14
Std.	±1.09	±0.16	±0.16	±0.10	±0.06	±0.11	±0.12	±0.12	±0.15
DC-FGSM (Ours)	53.67	**33.37**	**33.07**	**33.01**	**28.39**	**27.09**	**27.09**	**27.40**	**32.89**
Std.	±0.85	±0.11	±0.16	±0.15	±0.18	±0.23	±0.26	±0.24	±0.12
					$\epsilon = 8/255$, Last checkpoint				
PGD-AT-WA***	59.17	28.13	27.41	27.24	24.38	22.70	22.67	22.94	29.33
Std.	±0.28	±0.19	±0.26	±0.27	±0.20	±0.17	±0.16	±0.18	±0.21
FGSM-RS	61.72	0.78	0.41	0.15	0.10	0.00	1.50	0.74	8.17
Std.	±1.28	±0.04	±0.10	±0.02	±0.05	±0.00	±0.40	±0.08	±0.19
N-FGSM	57.37	27.44	26.82	26.72	23.87	22.16	22.15	22.32	28.61
Std.	±0.12	±0.15	±0.18	±0.16	±0.09	±0.02	±0.02	±0.05	±0.07
FGSM-MEP	58.14	28.48	27.88	27.73	24.17	22.59	22.61	22.85	29.31
Std.	±0.13	±0.04	±0.12	±0.09	±0.02	±0.17	±0.15	±0.12	±0.07
GAT	64.80	27.22	26.37	26.13	23.55	21.73	21.68	21.96	29.18
Std.	±0.10	±0.07	±0.11	±0.07	±0.29	±0.13	±0.13	±0.14	±0.12
GRAD-ALIGN	57.73	24.58	23.72	23.46	20.72	18.56	18.55	18.99	25.79
Std.	±1.13	±1.88	±1.93	±1.97	±2.18	±2.18	±2.17	±2.13	±1.94
NU-AT	62.98	24.33	20.67	18.44	16.96	11.24	16.20	15.75	23.32
Std.	±1.28	±1.12	±1.09	±1.00	±1.03	±0.65	±0.43	±0.57	±0.57
LIET	52.58	32.50	32.22	32.18	27.50	26.35	26.35	26.57	32.03
Std.	±0.30	±0.09	±0.06	±0.07	±0.15	±0.12	±0.13	±0.13	±0.03
DC-FGSM (Ours)	54.60	**33.17**	**32.84**	**32.76**	**28.07**	**26.70**	**26.67**	**27.06**	**32.73**
Std.	±0.33	±0.14	±0.09	±0.08	±0.09	±0.05	±0.07	±0.06	±0.04

Table 6: Clean and robust accuracy (%) on Tiny-ImageNet under $\epsilon = 8/255$. Avg. denotes the average of clean and robust accuracy. Best single-step results are shown in **bold red**.

Method	Tiny-ImageNet								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
					$\epsilon = 8/255$, Best checkpoint				
PGD-AT-WA* **	49.07	26.57	26.34	26.29	21.96	19.94	19.94	20.29	26.30
Std.	±0.22	±0.11	±0.14	±0.17	±0.13	±0.09	±0.07	±0.04	±0.04
FGSM-RS	46.42	24.56	24.29	24.24	19.70	17.62	17.57	17.89	24.04
Std.	±4.27	±1.32	±1.27	±1.24	±1.35	±1.11	±1.10	±1.13	±1.59
N-FGSM	46.12	25.25	25.11	25.11	20.58	18.67	18.63	18.96	24.80
Std.	±1.69	±0.17	±0.11	±0.08	±0.26	±0.24	±0.21	±0.22	±0.37
FGSM-MEP	49.27	26.46	26.17	26.11	21.27	19.40	19.38	19.73	25.97
Std.	±0.11	±0.21	±0.26	±0.33	±0.20	±0.05	±0.07	±0.18	±0.15
GAT	57.43	17.88	17.10	16.88	12.98	11.20	11.13	11.36	19.50
Std.	±0.22	±0.20	±0.17	±0.14	±0.18	±0.24	±0.25	±0.26	±0.19
GRAD-ALIGN	46.17	24.63	24.40	24.37	20.17	18.13	18.09	18.50	24.31
Std.	±1.96	±0.52	±0.52	±0.52	±0.51	±0.54	±0.54	±0.55	±0.70
NU-AT	27.68	15.40	15.30	15.29	11.54	10.32	10.33	10.62	14.56
Std.	±1.10	±0.10	±0.11	±0.11	±0.09	±0.14	±0.16	±0.13	±0.11
FREE-AT	36.23	19.97	19.90	19.88	15.26	13.59	13.55	13.78	19.02
Std.	±0.02	±0.31	±0.32	±0.31	±0.10	±0.11	±0.10	±0.15	±0.15
LIET	44.92	26.77	26.64	26.61	21.35	19.85	19.84	20.13	25.76
Std.	±0.15	±0.22	±0.23	±0.18	±0.01	±0.06	±0.08	±0.07	±0.11
DC-FGSM (Ours)	45.55	**27.29**	**26.71**	**26.68**	**21.56**	**20.18**	**20.27**	**20.47**	**26.02**
Std.	±0.38	±0.15	±0.17	±0.17	±0.12	±0.19	±0.21	±0.19	±0.07
					$\epsilon = 8/255$, Last checkpoint				
PGD-AT-WA* **	51.27	24.05	23.72	23.60	19.69	17.60	17.61	17.87	24.42
Std.	±0.18	±0.05	±0.03	±0.05	±0.08	±0.07	±0.06	±0.09	±0.05
FGSM-RS	35.93	7.25	7.07	7.00	5.56	4.71	4.76	4.80	9.63
Std.	±16.89	±10.25	±9.99	±9.89	±7.86	±6.66	±6.65	±6.79	±8.80
N-FGSM	49.37	22.56	22.11	21.98	17.60	15.68	15.66	15.95	22.61
Std.	±0.02	±0.34	±0.34	±0.40	±0.10	±0.02	±0.03	±0.02	±0.16
FGSM-MEP	53.93	10.22	9.08	8.40	6.30	5.46	9.53	6.01	13.62
Std.	±2.25	±9.24	±9.77	±10.19	±8.28	±7.71	±4.80	±7.49	±6.92
GAT	57.39	17.99	17.18	17.02	13.09	11.34	11.28	11.54	19.60
Std.	±0.10	±0.03	±0.08	±0.03	±0.08	±0.18	±0.20	±0.17	±0.09
GRAD-ALIGN	47.89	7.58	7.43	7.36	5.94	5.23	5.23	5.34	11.50
Std.	±2.18	±10.71	±10.50	±10.40	±8.40	±7.40	±7.39	±7.53	±8.00
NU-AT	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Std.	±0.00	±0.00	±0.00	±0.00	±0.00	±0.00	±0.00	±0.00	±0.00
FREE-AT	32.41	18.32	18.20	18.19	13.80	12.40	12.41	12.60	17.29
Std.	±0.05	±0.15	±0.10	±0.09	±0.04	±0.04	±0.06	±0.06	±0.06
LIET	46.01	26.15	25.99	25.93	20.80	19.24	19.24	19.52	25.36
Std.	±0.23	±0.06	±0.04	±0.04	±0.19	±0.04	±0.04	±0.06	±0.03
DC-FGSM (Ours)	46.86	**26.65**	**26.48**	**26.47**	**21.35**	**19.90**	**19.86**	**20.18**	**25.97**
Std.	±0.32	±0.13	±0.12	±0.13	±0.18	±0.20	±0.20	±0.21	±0.07

Compatibility of Gradient Correction. Tables 7 and 8 evaluate gradient correction (GC) combined with seven representative FAT baselines at $\epsilon = 8/255$, reporting mean $\pm$ standard deviation over three random seeds. On CIFAR-10, GC improves the average robust accuracy of most baselines, with particularly large gains for FGSM-RS, N-FGSM, GAT, and NU-AT; LIET, for example, improves from 56.30% to 57.09%. The same trend holds on the more challenging CIFAR-100, where GC improves FGSM-RS, N-FGSM, GAT, GRAD-ALIGN, NU-AT, and LIET, with the latter increasing from 32.15% to 33.25%. These consistent gains suggest that GC can serve as a general mechanism for correcting post-initialization gradient mismatch across different single-step adversarial training methods. Comparable standard deviations across seeds further support the reproducibility of these improvements.

Table 7: Clean and robust accuracy (%) on CIFAR-10 under $\epsilon = 8/255$. Methods marked with ★ use gradient correction. Improvements over the corresponding original methods are shown in **bold red**.

Method	CIFAR-10								
	Clean	**PGD-10**	**PGD-20**	**PGD-50**	**CW**	**AA**	**CR**	**PMA**	**Avg.**
FGSM-RS	79.94	53.55	52.93	52.74	49.56	47.29	47.36	47.48	53.86
Std.	±1.29	±0.26	±0.10	±0.16	±0.39	±0.37	±0.32	±0.36	±0.34
FGSM-RS★	82.74	**55.17**	**54.47**	**54.51**	**51.53**	**49.51**	**49.67**	**49.67**	**55.88**
Std.	±0.95	±0.02	±0.09	±0.20	±0.06	±0.09	±0.11	±0.11	±0.12
N-FGSM	79.88	53.67	53.14	53.02	50.13	48.00	48.09	48.19	54.27
Std.	±1.40	±0.09	±0.20	±0.26	±0.21	±0.09	±0.10	±0.13	±0.19
N-FGSM★	80.23	**55.19**	**54.79**	**54.73**	**50.39**	**50.89**	**49.90**	**50.37**	**55.81**
Std.	±2.64	±0.38	±0.42	±0.52	±0.93	±1.17	±0.41	±0.69	±0.51
FGSM-MEP	79.77	56.34	55.82	55.72	51.47	50.00	49.98	50.18	56.16
Std.	±1.22	±0.28	±0.16	±0.11	±0.32	±0.20	±0.21	±0.19	±0.31
FGSM-MEP★	80.35	**56.56**	**56.04**	**55.92**	**51.61**	**50.21**	**50.36**	**50.36**	**56.61**
Std.	±2.38	±0.57	±0.53	±0.48	±0.41	±0.50	±0.40	±0.40	±0.40
GAT	85.07	54.42	53.53	53.36	50.35	48.70	48.66	48.87	55.37
Std.	±0.10	±0.44	±0.20	±0.16	±0.30	±0.25	±0.24	±0.26	±0.20
GAT★	85.02	**55.87**	**54.99**	**54.78**	**50.54**	**49.52**	**49.09**	**50.12**	**56.24**
Std.	±0.09	±0.12	±0.17	±0.17	±0.90	±0.79	±0.19	±0.77	±0.34
GRAD-ALIGN	80.11	53.56	52.90	52.75	50.30	48.19	48.24	48.39	54.31
Std.	±1.59	±0.21	±0.22	±0.24	±0.46	±0.41	±0.39	±0.47	±0.43
GRAD-ALIGN★	79.95	**55.51**	**55.09**	**55.03**	**50.30**	**49.82**	**49.82**	**51.49**	**55.88**
Std.	±1.64	±0.37	±0.24	±0.21	±1.27	±0.28	±0.28	±0.31	±0.34
NU-AT	80.91	55.49	54.86	54.74	51.81	50.38	50.36	50.49	56.13
Std.	±0.58	±0.20	±0.40	±0.42	±0.29	±0.40	±0.37	±0.40	±0.23
NU-AT★	80.29	**56.70**	**56.09**	**55.95**	**52.44**	**51.27**	**51.27**	**52.44**	**57.06**
Std.	±0.40	±0.11	±0.18	±0.19	±0.65	±0.24	±0.24	±0.01	±0.14
LIET	80.22	56.82	56.31	56.22	50.77	49.97	49.93	50.13	56.30
Std.	±0.72	±0.16	±0.07	±0.13	±0.24	±0.21	±0.21	±0.22	±0.21
LIET★	81.70	**57.72**	**57.08**	**56.94**	**50.38**	**50.36**	**50.52**	**52.06**	**57.09**
Std.	±0.83	±0.28	±0.39	±0.39	±0.32	±0.45	±0.45	±0.27	±0.22

Table 8: Clean and robust accuracy (%) on CIFAR-100 under $\epsilon = 8/255$. Methods marked with ★ use gradient correction. Improvements over the corresponding original methods are shown in **bold red**.

Method	CIFAR-100								
	Clean	**PGD-10**	**PGD-20**	**PGD-50**	**CW**	**AA**	**CR**	**PMA**	**Avg.**
FGSM-RS	46.02	22.56	22.28	22.24	19.37	17.51	17.48	17.72	23.15
Std.	±2.15	±1.17	±1.11	±1.10	±1.07	±0.87	±0.89	±0.86	±8.89
FGSM-RS★	51.54	**29.75**	**29.56**	**29.58**	**25.56**	**23.92**	**23.90**	**24.13**	**29.74**
Std.	±4.79	±2.45	±2.32	±2.29	±2.65	±2.31	±2.29	±2.27	±2.35
N-FGSM	54.48	30.95	30.68	30.66	27.51	25.40	25.44	25.66	31.35
Std.	±0.21	±0.17	±0.17	±0.14	±0.31	±0.20	±0.23	±0.22	±9.04
N-FGSM★	55.23	**31.98**	**31.72**	**31.67**	**28.35**	**26.53**	**26.52**	**26.75**	**32.34**
Std.	±0.35	±0.10	±0.13	±0.11	±0.10	±0.16	±0.15	±0.10	±0.11
FGSM-MEP	56.75	32.11	31.79	31.71	27.92	26.24	26.26	26.50	32.41
Std.	±0.80	±0.17	±0.16	±0.16	±0.08	±0.05	±0.04	±0.07	±9.52
FGSM-MEP★	57.32	**32.14**	31.76	**31.72**	**28.02**	**26.37**	22.67	**26.57**	29.52
Std.	±1.03	±0.15	±0.23	±0.21	±0.16	±0.10	±4.11	±0.13	±0.13
GAT	65.73	27.36	26.48	26.32	23.81	21.91	21.84	22.09	29.44
Std.	±0.45	±0.22	±0.25	±0.25	±0.29	±0.19	±0.20	±0.26	±13.88
GAT★	64.76	**29.12**	**28.34**	**28.19**	**25.12**	**23.43**	**23.41**	**23.66**	**30.75**
Std.	±0.19	±0.14	±0.17	±0.15	±0.15	±0.14	±0.20	±0.24	±0.17
GRAD-ALIGN	55.09	31.25	30.94	30.85	27.47	25.30	25.29	25.58	31.47
Std.	±1.67	±0.13	±0.14	±0.14	±0.36	±0.35	±0.31	±0.33	±9.26
GRAD-ALIGN★	56.51	**32.16**	**31.84**	**31.73**	**28.31**	**26.42**	**26.43**	**26.72**	**32.51**
Std.	±1.00	±0.38	±0.49	±0.51	±0.46	±0.44	±0.47	±0.46	±0.46
NU-AT	57.92	25.56	22.41	20.77	19.26	13.28	16.72	16.96	24.11
Std.	±2.28	±0.56	±0.33	±0.39	±0.19	±0.66	±0.68	±0.24	±13.26
NU-AT★	52.43	24.61	**22.79**	**21.79**	**19.40**	**14.71**	**18.08**	**17.24**	23.88
Std.	±9.25	±1.07	±0.49	±1.29	±0.15	±2.53	±1.14	±0.71	±0.84
LIET	51.38	32.72	32.54	32.49	27.74	26.67	26.68	26.93	32.14
Std.	±1.09	±0.16	±0.16	±0.10	±0.06	±0.11	±0.12	±0.12	±7.72
LIET★	53.13	**33.64**	**33.49**	**33.44**	**28.87**	**27.72**	**27.70**	**27.98**	**33.25**
Std.	±0.50	±0.15	±0.15	±0.12	±0.09	±0.15	±0.14	±0.14	±0.09

G.3 Results Under Wider and Deeper Surrogate Models

Deeper Model: ResNet-34. To assess backbone generality, Tables 9 and 10 evaluate the methods with a deeper ResNet-34 backbone under $\epsilon = 8/255$. DC-FGSM remains competitive across both datasets and checkpoints, achieving 58.28%/51.63% PGD-50/AutoAttack accuracy on CIFAR-10 and 33.99%/28.38% on CIFAR-100 at the best checkpoint. Its strong final-checkpoint performance further indicates that the observed robustness and stability are not specific to the shallow backbone.

Table 9: Clean and robust accuracy (%) on CIFAR-10 using **ResNet-34** under $\epsilon = 8/255$. Avg. denotes the average of clean and robust accuracy. Best single-step results are shown in **bold red**.

Method	CIFAR-10								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
$\epsilon = 8/255$, Best checkpoint									
PGD-AT-WA***	82.67	56.43	55.81	55.66	53.65	51.88	51.89	52.02	57.50
Std.	±0.45	±0.12	±0.11	±0.05	±0.19	±0.18	±0.17	±0.14	±0.17
FGSM-RS	74.57	43.59	42.89	42.75	40.13	38.02	38.12	38.33	44.80
Std.	±2.64	±1.04	±0.89	±0.86	±1.51	±1.25	±1.24	±1.31	±1.32
N-FGSM	78.98	55.05	54.73	54.64	51.03	48.91	49.02	49.14	55.19
Std.	±1.55	±0.30	±0.22	±0.20	±0.65	±0.48	±0.44	±0.50	±0.54
FGSM-MEP	82.42	57.86	57.32	57.23	**53.11**	51.04	51.10	51.34	57.68
Std.	±0.53	±0.22	±0.21	±0.24	±0.17	±0.15	±0.09	±0.08	±0.09
GAT	85.70	55.96	55.29	55.14	51.69	50.16	50.13	50.27	56.79
Std.	±0.34	±0.10	±0.32	±0.34	±0.25	±0.15	±0.10	±0.06	±0.07
GRAD-ALIGN	80.67	55.24	54.74	54.62	51.67	49.47	49.58	49.73	55.72
Std.	±0.82	±0.28	±0.33	±0.29	±0.27	±0.22	±0.26	±0.24	±0.23
NU-AT	82.07	56.61	56.05	55.88	52.81	51.45	51.43	51.60	57.24
Std.	±0.16	±0.32	±0.26	±0.26	±0.20	±0.15	±0.14	±0.16	±0.18
LIET	81.83	57.57	57.03	56.96	52.47	50.93	50.91	51.13	57.35
Std.	±0.62	±0.27	±0.40	±0.38	±0.10	±0.22	±0.27	±0.21	±0.19
DC-FGSM (Ours)	81.82	**58.75**	**58.30**	**58.28**	53.09	**51.63**	**51.60**	**51.80**	**58.16**
Std.	±0.80	±0.08	±0.05	±0.06	±0.41	±0.37	±0.36	±0.36	±0.27
$\epsilon = 8/255$, Last checkpoint									
PGD-AT-WA***	85.88	52.41	50.87	50.46	50.89	48.54	48.72	48.87	54.58
Std.	±0.17	±0.01	±0.04	±0.07	±0.10	±0.03	±0.03	±0.02	±0.01
FGSM-RS	90.99	9.80	6.58	5.05	3.90	0.00	2.87	3.76	15.37
Std.	±0.42	±4.27	±3.41	±3.14	±3.18	±0.00	±2.61	±1.47	±2.19
N-FGSM	83.82	48.86	46.95	46.42	47.60	44.05	44.37	44.50	50.82
Std.	±0.36	±0.09	±0.06	±0.04	±0.27	±0.07	±0.10	±0.06	±0.09
FGSM-MEP	85.20	53.35	52.18	51.87	50.36	47.40	47.51	47.95	54.48
Std.	±0.21	±0.19	±0.18	±0.14	±0.26	±0.15	±0.13	±0.05	±0.12
GAT	85.94	53.49	52.13	51.78	50.19	47.96	47.96	48.23	54.71
Std.	±0.14	±0.19	±0.17	±0.19	±0.20	±0.30	±0.27	±0.24	±0.16
GRAD-ALIGN	84.81	49.64	47.74	47.26	48.22	44.75	45.05	45.23	51.59
Std.	±0.06	±0.11	±0.15	±0.17	±0.17	±0.15	±0.15	±0.22	±0.14
NU-AT	83.43	54.26	52.95	52.62	50.91	49.02	49.03	49.22	55.18
Std.	±0.16	±0.10	±0.15	±0.12	±0.14	±0.06	±0.03	±0.06	±0.08
LIET	84.03	55.35	54.10	53.78	50.74	48.88	48.89	49.17	55.62
Std.	±0.20	±0.09	±0.18	±0.18	±0.08	±0.08	±0.08	±0.11	±0.06
DC-FGSM (Ours)	84.75	**56.87**	**55.92**	**55.66**	**51.75**	**49.77**	**49.80**	**50.06**	**56.82**
Std.	±0.03	±0.35	±0.24	±0.18	±0.17	±0.17	±0.16	±0.14	±0.17

Table 10: Clean and robust accuracy (%) on CIFAR-100 using **ResNet-34** under $\epsilon = 8/255$. Avg. denotes the average of clean and robust accuracy. Best single-step results are shown in **bold red**.

Method	CIFAR-100								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
$\epsilon = 8/255$, Best checkpoint									
PGD-AT-WA***	58.54	32.97	32.59	32.57	29.97	28.13	28.10	28.33	33.90
Std.	±0.81	±0.23	±0.26	±0.29	±0.02	±0.06	±0.09	±0.09	±0.03
FGSM-RS	47.13	22.60	22.14	22.05	19.38	17.36	17.31	17.56	23.19
Std.	±0.04	±0.55	±0.54	±0.52	±0.18	±0.20	±0.21	±0.26	±0.31
N-FGSM	54.84	31.30	31.08	31.08	28.32	26.38	26.38	26.61	32.00
Std.	±0.79	±0.14	±0.12	±0.13	±0.20	±0.30	±0.26	±0.34	±0.29
FGSM-MEP	57.30	32.84	32.48	32.48	29.47	27.71	27.72	27.95	33.49
Std.	±1.04	±0.07	±0.03	±0.05	±0.15	±0.05	±0.01	±0.07	±0.13
GAT	66.26	28.00	27.26	27.11	25.79	23.86	23.85	24.08	30.78
Std.	±0.53	±0.07	±0.12	±0.19	±0.21	±0.19	±0.22	±0.21	±0.10
GRAD-ALIGN	57.98	31.06	30.63	30.50	28.78	26.50	26.48	26.79	32.34
Std.	±0.17	±0.09	±0.11	±0.12	±0.22	±0.27	±0.26	±0.25	±0.14
NU-AT	53.27	24.21	21.96	20.67	19.50	15.12	18.33	17.55	23.83
Std.	±11.66	±0.19	±1.67	±2.58	±0.80	±2.88	±0.94	±1.41	±0.33
LIET	51.93	33.01	32.81	32.75	28.99	27.79	27.80	27.99	32.88
Std.	±1.18	±0.18	±0.09	±0.11	±0.16	±0.13	±0.12	±0.09	±0.23
DC-FGSM (Ours)	54.93	**34.33**	**34.06**	**33.99**	**30.06**	**28.38**	**28.36**	**28.84**	**34.12**
Std.	±0.11	±0.30	±0.34	±0.33	±0.15	±0.17	±0.10	±0.16	±0.18
$\epsilon = 8/255$, Last checkpoint									
PGD-AT-WA***	59.58	28.52	27.80	27.66	25.49	23.93	23.90	24.13	30.13
Std.	±0.11	±0.20	±0.22	±0.21	±0.19	±0.20	±0.21	±0.23	±0.18
FGSM-RS	62.70	4.87	3.03	1.53	0.78	0.00	2.46	4.92	10.03
Std.	±1.45	±0.47	±0.26	±0.12	±0.07	±0.00	±0.31	±0.01	±0.12
N-FGSM	57.51	27.63	27.09	26.96	25.16	23.48	23.45	23.62	29.36
Std.	±0.43	±0.12	±0.12	±0.05	±0.06	±0.18	±0.12	±0.07	±0.14
FGSM-MEP	58.71	28.86	28.27	28.17	26.21	23.97	24.18	24.37	30.34
Std.	±0.19	±0.51	±0.61	±0.60	±0.35	±0.08	±0.30	±0.20	±0.36
GAT	64.67	26.99	25.99	25.67	24.92	22.78	22.72	22.90	29.58
Std.	±0.24	±0.19	±0.21	±0.29	±0.15	±0.20	±0.20	±0.17	±0.15
GRAD-ALIGN	62.31	13.32	11.42	10.09	8.95	7.39	10.52	11.67	16.96
Std.	±2.87	±9.17	±9.83	±10.64	±10.52	±10.28	±8.12	±7.62	±7.99
NU-AT	67.22	22.87	19.03	16.74	16.23	10.20	16.75	15.07	23.01
Std.	±0.62	±0.54	±0.65	±0.69	±0.65	±0.18	±0.91	±0.49	±0.39
LIET	53.17	32.26	31.97	31.88	28.34	26.98	26.98	27.25	32.35
Std.	±0.23	±0.19	±0.17	±0.24	±0.08	±0.13	±0.11	±0.14	±0.12
DC-FGSM (Ours)	56.48	**33.26**	**32.85**	**32.68**	**29.27**	**27.16**	**27.20**	**27.79**	**33.34**
Std.	±0.70	±0.03	±0.07	±0.11	±0.02	±0.21	±0.20	±0.12	±0.05

Wider Model: WideResNet-28-10. To further assess backbone generality, Tables 11 and 12 evaluate DC-FGSM with a wider WideResNet-28-10 backbone under $\epsilon = 8/255$. DC-FGSM achieves strong single-step robustness on both datasets, reaching 60.02%/53.85% PGD-50/AutoAttack accuracy on CIFAR-10 and 35.89%/29.99% on CIFAR-100 at the best checkpoint. Its competitive final-checkpoint performance further shows that the method remains effective as model width increases.

Table 11: Clean and robust accuracy (%) on CIFAR-10 using **WideResNet-28-10** under $\epsilon = 8/255$. Avg. denotes the average of clean and robust accuracy. Best single-step results are shown in **bold red**.

Method	CIFAR-10								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
$\epsilon = 8/255$, Best checkpoint									
PGD-AT-WA***	85.36	58.41	57.67	57.50	55.84	54.17	54.18	54.27	59.68
Std.	±0.70	±0.05	±0.26	±0.24	±0.29	±0.19	±0.16	±0.20	±0.15
FGSM-RS	89.95	29.76	21.10	14.49	13.54	0.00	15.11	23.98	25.99
Std.	±0.52	±4.67	±4.20	±3.30	±1.11	±0.00	±1.98	±3.84	±2.37
N-FGSM	86.51	35.79	31.19	27.81	24.00	14.38	22.74	25.16	33.45
Std.	±5.67	±11.53	±15.02	±17.90	±18.38	±24.27	±17.06	±15.16	±14.18
FGSM-MEP	83.79	59.98	59.32	59.18	55.36	53.00	53.05	53.37	59.63
Std.	±0.62	±0.05	±0.10	±0.15	±0.11	±0.18	±0.17	±0.05	±0.04
GAT	82.20	52.45	51.92	51.80	47.95	46.75	46.69	46.94	53.34
Std.	±1.82	±4.84	±5.17	±5.18	±3.91	±4.31	±4.30	±4.17	±4.14
GRAD-ALIGN	82.54	56.35	55.82	55.64	52.92	50.72	50.80	50.94	56.97
Std.	±1.15	±0.72	±0.66	±0.61	±0.48	±0.31	±0.33	±0.34	±0.55
NU-AT	83.70	58.19	57.49	57.31	54.69	53.39	53.35	53.50	58.95
Std.	±0.43	±0.30	±0.18	±0.16	±0.49	±0.44	±0.46	±0.50	±0.35
LIET	83.75	59.94	59.28	59.23	54.74	53.30	53.28	53.51	59.63
Std.	±0.92	±0.04	±0.16	±0.14	±0.07	±0.06	±0.06	±0.05	±0.07
DC-FGSM (Ours)	84.32	**60.75**	**60.15**	**60.02**	**55.35**	**53.85**	**53.82**	**54.02**	**60.28**
Std.	±1.05	±0.29	±0.49	±0.51	±0.06	±0.19	±0.22	±0.14	±0.12
$\epsilon = 8/255$, Last checkpoint									
PGD-AT-WA***	86.58	52.31	50.40	49.84	49.07	47.42	47.67	47.61	53.86
Std.	±0.44	±0.23	±0.15	±0.15	±0.16	±0.18	±0.17	±0.23	±0.10
FGSM-RS	89.97	18.13	11.03	6.58	6.76	0.00	9.14	14.20	19.48
Std.	±0.58	±1.73	±1.06	±0.14	±0.82	±0.00	±0.67	±1.71	±0.67
N-FGSM	85.91	30.15	24.57	20.91	19.46	9.95	16.53	18.88	28.29
Std.	±7.69	±9.11	±12.15	±14.68	±18.12	±16.30	±11.29	±9.57	±10.43
FGSM-MEP	87.12	48.59	45.48	43.54	42.42	33.65	44.57	41.94	48.41
Std.	±2.06	±7.83	±10.74	±13.33	±14.40	±22.76	±5.15	±10.00	±10.27
GAT	93.66	39.07	30.40	23.20	25.45	7.05	26.67	21.44	33.37
Std.	±0.14	±3.07	±2.84	±2.45	±2.67	±0.76	±0.95	±0.70	±1.66
GRAD-ALIGN	85.82	49.42	46.93	46.02	46.88	43.37	43.85	43.92	50.78
Std.	±0.50	±0.62	±0.63	±0.72	±0.94	±0.34	±0.42	±0.51	±0.54
NU-AT	91.07	48.92	43.02	39.01	37.90	23.41	42.97	35.97	45.28
Std.	±1.63	±0.74	±0.81	±2.40	±4.73	±7.30	±7.15	±2.22	±2.75
LIET	86.12	55.61	54.28	53.93	51.83	50.15	50.14	50.36	56.55
Std.	±0.07	±0.30	±0.30	±0.21	±0.05	±0.08	±0.05	±0.06	±0.10
DC-FGSM (Ours)	86.55	**56.73**	**55.47**	**55.08**	**52.86**	**51.03**	**50.96**	**51.31**	**57.50**
Std.	±0.09	±0.22	±0.19	±0.21	±0.18	±0.14	±0.11	±0.06	±0.14

G.4 Evaluation Under Larger Perturbation Budgets

CIFAR-10 with $\epsilon = 16/255$. Table 13 reports the complete results of the evaluated adversarial training methods on CIFAR-10 under the larger perturbation budget $\epsilon = 16/255$. The table evaluates the robustness of DC-FGSM and the baseline methods under the same attack protocols.

CIFAR-100 under $\epsilon = 16/255$. Table 14 reports the results under the larger perturbation budget. DC-FGSM maintains competitive robustness across the evaluated attacks and preserves substantially stronger performance at the Last checkpoint, achieving 14.52% PGD-50 and 8.49% AutoAttack accuracy, compared with 8.05% and 4.10% for LIET, respectively. Its relatively small standard deviations across attacks and checkpoints further indicate stable and reproducible performance under this more challenging setting.

Tiny-ImageNet with $\epsilon = 12/255$. We also evaluate Tiny-ImageNet under $\epsilon = 12/255$, as shown in Table 15. DC-FGSM achieves the best FAT robustness at the best checkpoint under PGD-50 and CW, reaching 18.14% and 13.45%, respectively. At the last checkpoint, it maintains 17.46% PGD-50 and 10.97% AA accuracy, outperforming LIET and other FAT baselines. These results suggest that the stability advantage of DC-FGSM remains visible under stronger perturbations and larger-scale data.

Table 12: Clean and robust accuracy (%) on CIFAR-100 using **WideResNet-28-10** under $\epsilon = 8/255$. Avg. denotes the average of clean and robust accuracy. DC-FGSM results that are the best among single-step methods are shown in **bold red**.

Method	CIFAR-100								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
				$\epsilon = 8/255$, Best checkpoint					
PGD-AT-WA***	60.94	34.79	34.39	34.31	31.54	29.79	29.77	30.07	35.70
Std.	±1.19	±0.26	±0.37	±0.44	±0.23	±0.20	±0.21	±0.21	±0.10
FGSM-RS	36.77	18.21	17.91	17.82	15.15	13.33	13.32	13.59	18.26
Std.	±17.96	±8.18	±8.05	±8.10	±7.89	±7.21	±7.21	±7.27	±8.98
N-FGSM	58.75	33.00	32.71	32.65	30.35	28.14	28.19	28.42	34.03
Std.	±0.92	±0.34	±0.38	±0.40	±0.20	±0.07	±0.05	±0.04	±0.05
FGSM-MEP	52.54	29.18	28.84	28.77	25.54	23.63	23.66	23.95	29.51
Std.	±3.94	±3.77	±3.85	±3.89	±3.31	±3.15	±3.20	±3.23	±3.54
GAT	70.10	27.88	27.04	26.82	25.69	23.52	23.47	23.72	31.03
Std.	±0.87	±0.45	±0.53	±0.49	±0.51	±0.60	±0.65	±0.59	±0.37
GRAD-ALIGN	48.75	25.20	25.02	24.91	21.95	20.02	20.02	20.36	25.78
Std.	±0.06	±0.13	±0.17	±0.11	±0.44	±0.45	±0.46	±0.36	±0.26
NU-AT	45.84	24.95	23.72	23.06	20.23	17.38	18.77	18.76	24.09
Std.	±16.41	±0.34	±2.36	±3.47	±1.37	±4.31	±1.87	±2.52	±0.17
LIET	57.69	34.27	33.48	33.22	29.50	24.73	27.14	26.68	33.34
Std.	±5.32	±0.45	±0.70	±0.84	±0.60	±3.32	±1.05	±1.79	±0.49
DC-FGSM (Ours)	57.43	**36.20**	**35.90**	**35.89**	**31.52**	**29.99**	**29.93**	**30.37**	**35.90**
Std.	±0.83	±0.04	±0.11	±0.08	±0.16	±0.14	±0.09	±0.07	±0.04
				$\epsilon = 8/255$, Last checkpoint					
PGD-AT-WA***	61.76	29.07	28.36	28.19	26.12	24.77	24.73	24.93	30.99
Std.	±0.12	±0.32	±0.27	±0.31	±0.20	±0.20	±0.20	±0.20	±0.19
FGSM-RS	61.88	7.01	4.75	3.46	1.30	0.00	3.16	7.72	11.16
Std.	±1.98	±4.78	±4.55	±3.67	±1.00	±0.00	±1.71	±4.44	±2.44
N-FGSM	59.24	27.56	26.84	26.76	25.27	23.65	23.65	23.85	29.60
Std.	±0.38	±0.24	±0.30	±0.32	±0.23	±0.23	±0.20	±0.21	±0.18
FGSM-MEP	69.17	19.30	15.70	12.86	8.35	2.26	16.44	16.04	20.02
Std.	±1.41	±0.59	±0.51	±0.74	±1.48	±1.02	±1.23	±1.15	±0.47
GAT	70.92	25.58	20.75	16.87	15.41	6.28	14.63	12.66	22.89
Std.	±1.45	±0.13	±0.92	±1.47	±2.06	±2.12	±2.85	±2.02	±1.25
GRAD-ALIGN	61.57	5.39	3.54	2.38	0.86	0.00	3.22	4.92	10.24
Std.	±1.36	±0.89	±0.66	±0.58	±0.08	±0.00	±0.68	±1.07	±0.62
NU-AT	70.73	22.02	18.01	15.61	14.15	9.25	14.32	13.43	22.19
Std.	±0.35	±1.11	±0.83	±0.39	±0.60	±0.28	±1.37	±1.17	±0.64
LIET	64.74	28.91	26.32	24.56	22.53	10.87	27.06	17.42	27.80
Std.	±0.33	±1.83	±1.89	±1.80	±1.72	±2.41	±0.82	±1.86	±1.53
DC-FGSM (Ours)	63.78	**29.84**	**27.58**	**27.09**	**25.13**	**19.67**	25.76	**23.75**	**30.33**
Std.	±0.16	±0.18	±0.23	±0.21	±0.05	±0.17	±0.34	±0.31	±0.16

Table 13: Clean and robust accuracy (%) on CIFAR-10 under $\epsilon = 16/255$. Avg. denotes the average of clean and robust accuracy. Best single-step results are shown in **bold red**.

Method	CIFAR-10								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
				$\epsilon = 16/255$, Best checkpoint					
PGD-AT-WA***	72.45	40.34	32.22	30.03	28.32	24.12	24.70	25.02	34.65
Std.	±1.14	±0.42	±0.78	±0.94	±0.61	±0.65	±0.70	±0.69	±0.45
FGSM-RS	49.75	28.55	24.19	23.36	19.08	16.11	16.28	16.63	24.24
Std.	±3.94	±0.89	±0.59	±0.66	±0.57	±0.73	±0.79	±0.61	±0.22
N-FGSM	64.03	39.18	32.63	31.14	26.48	22.14	22.69	23.12	32.68
Std.	±2.79	±0.07	±0.75	±1.10	±0.44	±1.06	±0.99	±0.94	±0.32
FGSM-MEP	43.60	27.62	24.79	24.51	20.52	19.09	19.09	19.50	24.84
Std.	±0.57	±0.68	±0.68	±0.67	±0.86	±0.79	±0.79	±0.96	±0.66
GAT	70.07	26.71	19.84	18.68	14.75	12.96	12.79	13.16	23.62
Std.	±0.53	±0.21	±0.05	±0.02	±0.02	±0.06	±0.03	±0.02	±0.09
GRAD-ALIGN	62.54	36.45	30.24	28.93	24.10	19.69	20.17	20.75	30.36
Std.	±1.49	±0.58	±0.85	±1.00	±0.80	±1.24	±1.17	±1.05	±0.60
NU-AT	91.14	20.62	12.02	6.23	8.32	0.10	3.33	3.71	18.18
Std.	±0.14	±1.13	±0.96	±0.63	±0.70	±0.02	±0.63	±0.24	±0.39
LIET	60.40	39.44	34.34	33.14	26.56	22.98	23.19	23.92	33.00
Std.	±4.62	±0.98	±0.35	±0.66	±1.27	±1.83	±1.64	±1.57	±0.50
DC-FGSM(Ours)	72.54	**43.01**	**35.68**	**34.00**	**27.46**	**23.46**	**23.49**	**24.60**	**35.53**
Std.	±0.50	±0.17	±0.04	±0.19	±0.09	±0.05	±0.12	±0.21	±0.04
				$\epsilon = 16/255$, Last checkpoint					
PGD-AT-WA***	76.10	37.63	28.26	25.01	26.07	19.57	20.31	22.55	31.94
Std.	±0.18	±0.22	±0.32	±0.11	±0.09	±0.19	±0.19	±0.01	±0.03
FGSM-RS	69.57	6.87	2.01	0.36	0.54	0.00	1.84	3.17	10.54
Std.	±4.15	±1.46	±0.74	±0.18	±0.30	±0.00	±0.57	±0.71	±0.65
N-FGSM	70.69	35.33	26.17	23.00	23.00	15.80	16.18	18.86	28.63
Std.	±0.46	±0.15	±0.24	±0.31	±0.09	±0.25	±0.36	±0.17	±0.13
FGSM-MEP	88.48	17.55	11.52	6.43	3.61	0.13	16.07	11.69	19.44
Std.	±1.66	±1.03	±1.06	±0.99	±1.08	±0.03	±1.22	±1.40	±0.98
GAT	91.67	13.41	6.44	3.08	3.04	0.05	2.21	2.03	15.24
Std.	±0.25	±0.21	±0.53	±0.41	±0.58	±0.01	±0.14	±0.34	±0.16
GRAD-ALIGN	72.44	32.52	22.67	19.07	19.28	12.22	12.84	15.37	25.80
Std.	±0.85	±0.71	±0.78	±0.69	±0.58	±0.38	±0.43	±0.55	±0.52
NU-AT	91.27	19.80	11.28	5.64	5.57	0.11	3.02	2.50	17.40
Std.	±0.10	±1.31	±1.12	±0.89	±0.47	±0.01	±0.79	±0.13	±0.48
LIET	76.97	20.84	15.09	12.72	13.58	7.38	14.05	12.33	21.62
Std.	±7.99	±14.08	±13.78	±14.25	±10.25	±10.15	±7.12	±8.77	±8.61
DC-FGSM(Ours)	77.51	**41.37**	**31.83**	**29.09**	**24.61**	**19.41**	**19.74**	**21.33**	**33.11**
Std.	±0.18	±0.28	±0.42	±0.35	±0.23	±0.33	±0.27	±0.14	±0.23

Table 14: Clean and robust accuracy (%) on CIFAR-100 under $\epsilon = 16/255$. Avg. denotes the average of Clean, PGD-10, PGD-20, PGD-50, CW, AA, CR, and PMA. Best single-step results, excluding the Clean column, are shown in **bold red**.

Method	CIFAR-100								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
$\epsilon = 16/255$, Best checkpoint									
PGD-AT-WA***	46.38	22.06	17.75	17.12	14.32	12.30	12.33	12.73	19.37
Std.	±1.54	±0.04	±0.21	±0.32	±0.27	±0.26	±0.30	±0.30	±0.02
FGSM-RS	26.43	12.89	10.99	10.72	8.12	6.98	7.07	7.25	11.31
Std.	±0.69	±0.83	±0.64	±0.62	±0.43	±0.42	±0.44	±0.45	±0.53
N-FGSM	41.42	19.97	16.18	15.52	13.05	10.94	11.05	11.32	17.43
Std.	±0.60	±0.05	±0.17	±0.20	±0.17	±0.08	±0.12	±0.10	±0.13
FGSM-MEP	29.09	14.34	12.28	12.00	9.16	8.02	7.98	8.27	12.64
Std.	±5.22	±1.12	±0.80	±0.75	±0.91	±0.73	±0.71	±0.71	±1.37
GAT	50.03	5.93	3.48	2.91	2.69	1.39	1.48	1.64	8.69
Std.	±2.79	±0.66	±0.62	±0.56	±0.65	±0.40	±0.31	±0.36	±0.56
GRAD-ALIGN	31.75	14.99	12.20	11.64	9.05	7.53	7.58	7.87	12.83
Std.	±0.40	±0.43	±0.29	±0.31	±0.18	±0.23	±0.24	±0.22	±0.16
NU-AT	27.61	15.49	13.52	13.30	9.53	8.23	8.25	8.60	13.07
Std.	±0.37	±0.41	±0.37	±0.30	±0.20	±0.18	±0.14	±0.11	±0.14
LIET	34.99	20.14	17.40	15.87	12.55	11.11	11.23	11.62	16.86
Std.	±12.81	±0.43	±1.04	±1.56	±1.00	±0.56	±0.45	±0.54	±0.31
DC-FGSM (Ours)	47.38	**23.15**	**19.11**	**18.41**	**15.00**	**11.95**	**12.41**	**13.05**	**20.06**
Std.	±1.26	±0.10	±0.17	±0.23	±0.11	±0.23	±0.20	±0.29	±0.12
$\epsilon = 16/255$, Last checkpoint									
PGD-AT-WA***	50.44	19.00	14.29	13.42	12.27	9.55	12.33	10.69	17.75
Std.	±0.05	±0.28	±0.15	±0.10	±0.05	±0.01	±0.30	±0.02	±0.08
FGSM-RS	34.20	0.07	0.01	0.00	0.00	0.00	7.07	0.09	5.18
Std.	±9.19	±0.01	±0.00	±0.00	±0.00	±0.00	±0.44	±0.02	±1.20
N-FGSM	46.75	16.58	12.44	11.46	10.57	7.85	11.05	8.96	15.71
Std.	±0.81	±0.86	±0.71	±0.69	±0.50	±0.55	±0.12	±0.34	±0.48
FGSM-MEP	64.55	2.08	1.10	0.55	0.32	0.04	7.98	1.74	9.80
Std.	±4.01	±0.23	±0.11	±0.10	±0.06	±0.04	±0.71	±0.10	±0.34
GAT	70.77	3.10	1.27	0.58	0.48	0.06	1.48	0.60	9.79
Std.	±0.57	±0.15	±0.12	±0.05	±0.09	±0.00	±0.31	±0.14	±0.14
GRAD-ALIGN	52.77	0.07	0.01	0.00	0.01	0.00	7.58	0.12	7.57
Std.	±0.54	±0.04	±0.01	±0.00	±0.01	±0.00	±0.24	±0.04	±0.07
NU-AT	62.86	9.40	5.18	3.45	3.11	1.45	8.25	2.87	12.07
Std.	±0.25	±0.57	±0.41	±0.29	±0.17	±0.09	±0.14	±0.40	±0.23
LIET	53.19	10.47	8.37	6.25	5.25	4.00	5.81	5.66	12.38
Std.	±0.21	±0.58	±0.44	±0.37	±0.25	±0.41	±0.45	±0.23	±0.16
DC-FGSM (Ours)	50.58	**22.18**	**17.65**	**16.67**	**13.61**	**10.16**	**11.27**	**11.94**	**19.26**
Std.	±0.65	±0.36	±0.30	±0.18	±0.11	±0.37	±0.15	±0.07	±0.18

Table 15: Clean and robust accuracy (%) on Tiny-ImageNet under $\epsilon = 12/255$. Avg. denotes the average of clean and robust accuracy. Best results are shown in **bold red**.

Method	Tiny-ImageNet								
	Clean	PGD-10	PGD-20	PGD-50	CW	AA	CR	PMA	Avg.
$\epsilon = 12/255$, Best checkpoint									
PGD-AT-WA***	41.55	19.80	18.71	18.53	14.38	12.28	12.26	12.74	18.78
Std.	±0.61	±0.15	±0.10	±0.05	±0.13	±0.03	±0.05	±0.06	±0.09
FGSM-RS	24.76	10.66	9.98	9.89	6.88	5.67	5.64	5.88	9.92
Std.	±3.31	±2.12	±2.09	±2.11	±1.65	±1.41	±1.40	±1.40	±1.93
N-FGSM	40.19	18.31	17.00	16.73	13.08	11.00	10.93	11.30	17.32
Std.	±0.22	±0.14	±0.05	±0.07	±0.16	±0.18	±0.18	±0.15	±0.05
FGSM-MEP	23.81	11.30	10.64	10.58	7.34	6.13	6.06	6.45	10.29
Std.	±0.86	±0.46	±0.42	±0.41	±0.28	±0.46	±0.44	±0.39	±0.45
GAT	43.46	1.20	0.69	0.49	0.40	0.12	0.13	0.18	5.83
Std.	±2.23	±0.07	±0.04	±0.03	±0.05	±0.06	±0.08	±0.07	±0.25
GRAD-ALIGN	35.65	15.21	14.11	13.93	10.49	8.52	8.46	8.86	14.40
Std.	±1.38	±0.56	±0.51	±0.48	±0.46	±0.38	±0.40	±0.39	±0.37
NU-AT	41.28	10.80	8.44	7.64	5.56	3.50	4.02	4.48	10.72
Std.	±15.25	±0.31	±1.23	±1.78	±1.09	±1.85	±1.46	±1.32	±0.84
LIET	33.93	18.32	17.44	17.34	13.23	11.29	11.62	11.93	16.93
Std.	±0.53	±0.18	±0.17	±0.21	±0.25	±0.09	±0.10	±0.12	±0.20
DC-FGSM (Ours)	41.11	**19.60**	**18.18**	**18.14**	**13.45**	**11.37**	**11.29**	**11.85**	**18.01**
Std.	±1.05	±0.23	±0.27	±0.27	±0.30	±0.21	±0.25	±0.30	±0.14
$\epsilon = 12/255$, Last checkpoint									
PGD-AT-WA***	44.51	18.52	16.93	16.65	13.34	11.07	11.01	11.49	17.94
Std.	±0.15	±0.14	±0.16	±0.08	±0.07	±0.10	±0.07	±0.10	±0.06
FGSM-RS	18.38	0.05	0.03	0.01	0.00	0.00	0.00	0.00	2.31
Std.	±14.03	±0.08	±0.04	±0.01	±0.00	±0.00	±0.00	±0.00	±1.77
N-FGSM	42.73	16.85	15.11	14.76	11.30	9.29	9.24	9.58	16.11
Std.	±0.21	±0.09	±0.13	±0.19	±0.10	±0.04	±0.08	±0.05	±0.06
FGSM-MEP	40.59	0.09	0.04	0.00	0.00	0.00	3.04	0.00	5.47
Std.	±20.87	±0.06	±0.02	±0.00	±0.00	±0.00	±0.35	±0.00	±2.59
GAT	63.96	1.06	0.31	0.12	0.06	0.00	0.07	0.06	8.21
Std.	±0.29	±0.11	±0.02	±0.03	±0.02	±0.00	±0.01	±0.02	±0.05
GRAD-ALIGN	39.67	8.91	8.19	8.00	5.85	4.63	4.60	4.86	10.59
Std.	±5.50	±6.41	±5.91	±5.79	±4.28	±3.41	±3.40	±3.60	±3.47
NU-AT	56.02	8.78	5.47	4.40	2.77	1.30	1.94	2.53	10.40
Std.	±0.24	±0.45	±0.38	±0.33	±0.12	±0.01	±0.24	±0.23	±0.19
LIET	47.61	8.59	6.26	5.31	3.46	1.62	3.36	3.54	9.97
Std.	±1.84	±1.66	±1.70	±1.63	±1.23	±0.83	±0.88	±1.15	±0.91
DC-FGSM (Ours)	42.48	**19.28**	**17.73**	**17.46**	**13.15**	**10.85**	**10.16**	**11.27**	**17.83**
Std.	±0.16	±0.10	±0.05	±0.01	±0.13	±0.05	±0.09	±0.14	±0.06