A. Supplementary Experiments for Section 3.1

This section supplements Section 3.1 of the main paper by providing additional experimental details and results. We present the complete experimental setup described in Section 3.1, results for perturbation size of 8/255 in Table 5, and comprehensive results across five seeds for varying perturbation sizes as illustrated in Figures 6 and 7.

To explore the phenomenon of the transferability of adversarial perturbations following catastrophic overfitting, we engaged in an experimental study based on the settings established by He et al. [11]. We initiated our study by training a ResNet18 [10] model on the CIFAR-10 dataset using the Fast Gradient Sign Method Adversarial Training (FGSM-AT) [7] for 100 epochs. The training was conducted with perturbation sizes (ϵ) set to 8/255 and 16/255, employing a Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.1. The learning rate was programmed to diminish by a factor of 0.1 upon reaching the 80th and 90th epochs. Additionally, the model training incorporated a batch size of 128, and the images underwent preprocessing, which included padding of 4 pixels on each side, followed by random cropping and horizontal flipping. To precisely replicate the conditions leading to CO, we adopted zero initialization for generating adversarial samples and set the weight decay to zero. This setup was chosen to maintain consistency with He et al. [11] and to ensure the stable reproduction of CO, thereby facilitating a clear examination of the transferability of adversarial perturbations under these conditions.

We delve into the specifics of the experimental outcomes for each seed, as illustrated in Figures 6 and 7, to shed light on the underlying dynamics of $P_{abnormal}$ and PGD accuracy in relation to catastrophic overfitting. For all five seeds, we observed a gradual increase in $P_{abnormal}$ during the initial stages of training. However, a striking observation was made at the point of CO, where PGD accuracy plummeted to approximately 0, underscoring a sudden and severe degradation in the model's ability to counter adversarial attacks. Correspondingly, $P_{abnormal}$ experienced a sharp escalation, reinforcing the strong linkage between the onset of CO and the dramatic increase in $P_{abnormal}$. This pattern was consistent across different seeds. The detailed analysis for each seed further corroborates the significant impact of CO on the transferability of adversarial perturbations.

B. Pseudocode of the LIET Algorithm for Section 3.3

This section supplements Section 3.3 of the main paper by presenting the pseudocode for the LIET algorithm (Algorithm 1). To enhance clarity, the pseudocode simplifies certain operations. For instance, an implementation detail concerns the clipping of adversarial perturbations. Specifically, for a perturbation budget of $\epsilon = 8/255$, we clip the perturbation to stay within this bound. For larger budgets, however, we adopt the strategy from [3] and omit the clipping step to generate stronger adversaries. For a comprehensive implementation, we refer the reader to the provided source code.

Algorithm 1: LIET: Label Information Elimination Training

Input: A classifier f_θ with loss function $\mathcal{L}$; Dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$; Perturbation magnitude ϵ; Gray image $\boldsymbol{x_{gray}}$; Hyperparameter λ; Number of epochs E.

Output: Robust model parameters $\boldsymbol{\theta}$

for $e = 1$ **to** E **do**
  for *each class* c *in dataset* **do**
    $\boldsymbol{LI_c} = \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x_{gray}}} \mathcal{L}(f(\boldsymbol{x_{gray}}; \boldsymbol{\theta}), c))$ {Generate class-specific label information}
  end
  for *each batch* $\mathcal{B} = (\boldsymbol{x}, y) \subset \mathcal{D}$ **do**
    if *random*() < 0.5 **then**
      $\boldsymbol{x'} = \boldsymbol{x} + \boldsymbol{LI_y}$ {Randomly add label information}
    else
      $\boldsymbol{x'} = \boldsymbol{x} - \boldsymbol{LI_y}$ {Randomly subtract label information}
    end
    $\boldsymbol{\delta_x} = \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x'}} \mathcal{L}(f(\boldsymbol{x'}; \boldsymbol{\theta}), y))$ {Generate adversarial perturbation}
    $\boldsymbol{x_{adv}} = \boldsymbol{x} + \boldsymbol{\delta_x}$ {Create adversarial example}
    $\text{loss}_1 = \mathcal{L}(f(\boldsymbol{x_{adv}}; \boldsymbol{\theta}), y)$ {Standard adversarial training loss}
    $\text{loss}_2 = \lambda \cdot \mathcal{L}_{JSD}(f(\boldsymbol{x}; \boldsymbol{\theta}), f(\boldsymbol{x_{adv}}; \boldsymbol{\theta}))$ {JS divergence for smoother loss surface}
    $\text{total\_loss} = \text{loss}_1 + \text{loss}_2$ {Combined loss function}
  end
  $\boldsymbol{\theta} = \boldsymbol{\theta} - \eta \cdot \nabla_{\boldsymbol{\theta}} \text{total\_loss}$ {Update model parameters}
end

C. Experiment Details for Section 4.1

This section provides supplementary information to Section 4.1 of the main paper. Here, we present detailed experimental configurations and parameters that were utilized in our study but were omitted from the main paper for brevity.

Our research conducted experiments on three widely recognized datasets to evaluate the robustness against adversarial attacks, namely CIFAR-10, CIFAR-100, and

Input x	**Perturbation: 16/255**		**Perturbation: 8/255**	
	$P_{abnormal}$ (%)	$P_{dominate}$ (%)	$P_{abnormal}$ (%)	$P_{dominate}$ (%)
Uniform Gray (0)	42.20 ± 32.55	20.29 ± 28.62	49.89 ± 1.34	10.08 ± 0.16
Uniform Gray (0.5)	81.81 ± 20.00	44.47 ± 27.44	53.52 ± 4.13	10.52 ± 0.62
Training Mean	81.36 ± 20.47	45.45 ± 28.42	53.84 ± 4.31	10.55 ± 0.68
Uniform Noise	50.70 ± 2.35	11.08 ± 1.72	48.67 ± 0.93	9.87 ± 0.10
Test Sample 0	73.08 ± 18.95	26.06 ± 15.03	51.18 ± 0.49	10.12 ± 0.13

Table 5. Transferability of label information for different inputs on the CIFAR-10 dataset with perturbation sizes of 16/255 and 8/255.

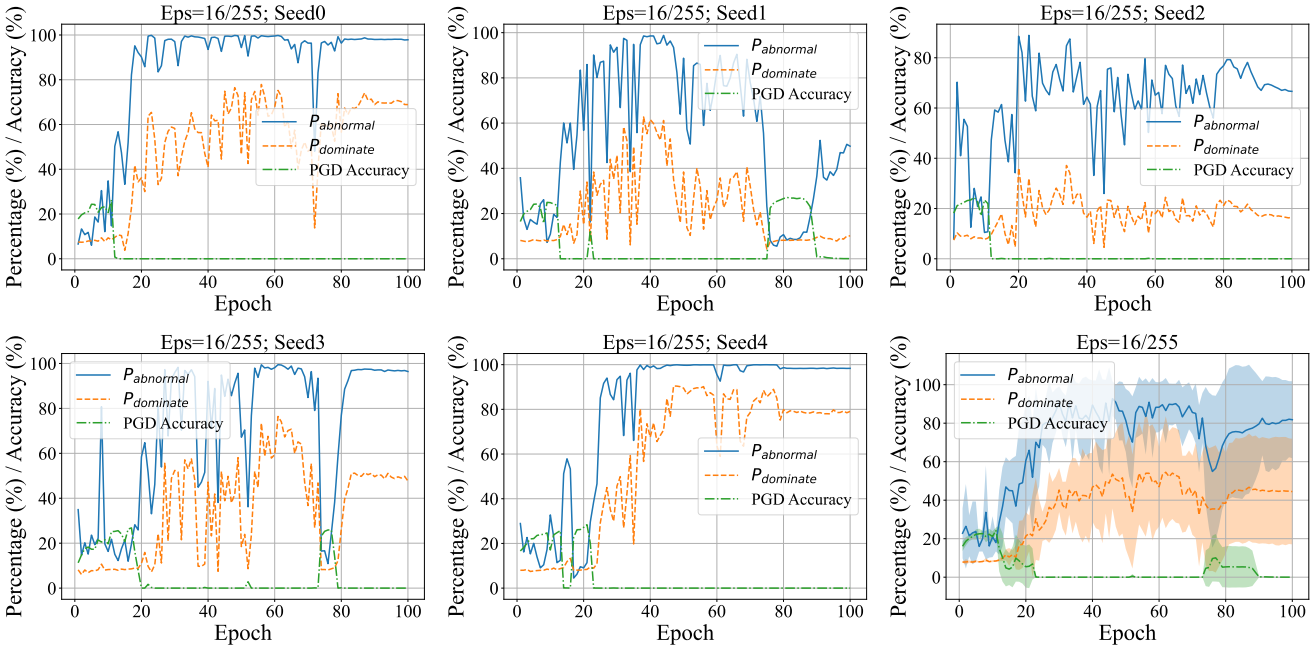

Figure 6. Transferability of label information for uniform gray image (value 0.5) on the CIFAR10 dataset with a perturbation size of 16/255.

Tiny ImageNet. For CIFAR-10 and CIFAR-100 datasets, we employed both the ResNet-18 architecture [10] and WideResNet-28-10 [29] as the network architectures, while for Tiny ImageNet, we opted for PreActResNet18 due to its enhanced performance on more complex datasets.

For the CIFAR-10, CIFAR-100, and Tiny ImageNet datasets, we carved out validation sets comprising 1000, 1000, and 2000 images, respectively, from the training data. During the training phase, we evaluated the model's performance on these validation sets using the PGD-10 accuracy metric. The model that achieved the highest accuracy on the validation set was selected as the final model. This validation strategy was consistently applied across all compared algorithms to maintain uniformity in model evaluation.

We set a batch size of 128 and applied a series of preprocessing steps on the images. These steps included padding the images with 4 pixels on each side, followed by random cropping and horizontal flipping to augment the dataset and improve model generalization.

We utilized the Stochastic Gradient Descent (SGD) as our optimization algorithm, with an initial learning rate set at 0.1, a weight decay parameter of 5e-4, and momentum of 0.9. The training process was conducted over 100 epochs, incorporating a OneCycleLR scheduler to adjust the learning rate dynamically. To stabilize the training process, we implemented a Weight Averaging (WA) [12] technique with a τ value of 0.9995. Each experiment was replicated three times under different random seeds to ensure the reliability of our results. For perturbation magnitude of $8/255$, we set λ values at 100, 200, and 100, respectively. Furthermore, we employed non-uniform label smoothing values of 0.6 for both CIFAR-10 and CIFAR-100, and 0.8 for Tiny ImageNet, to fine-tune the model's performance across diverse datasets. For perturbation magnitude of $16/255$, we set λ to

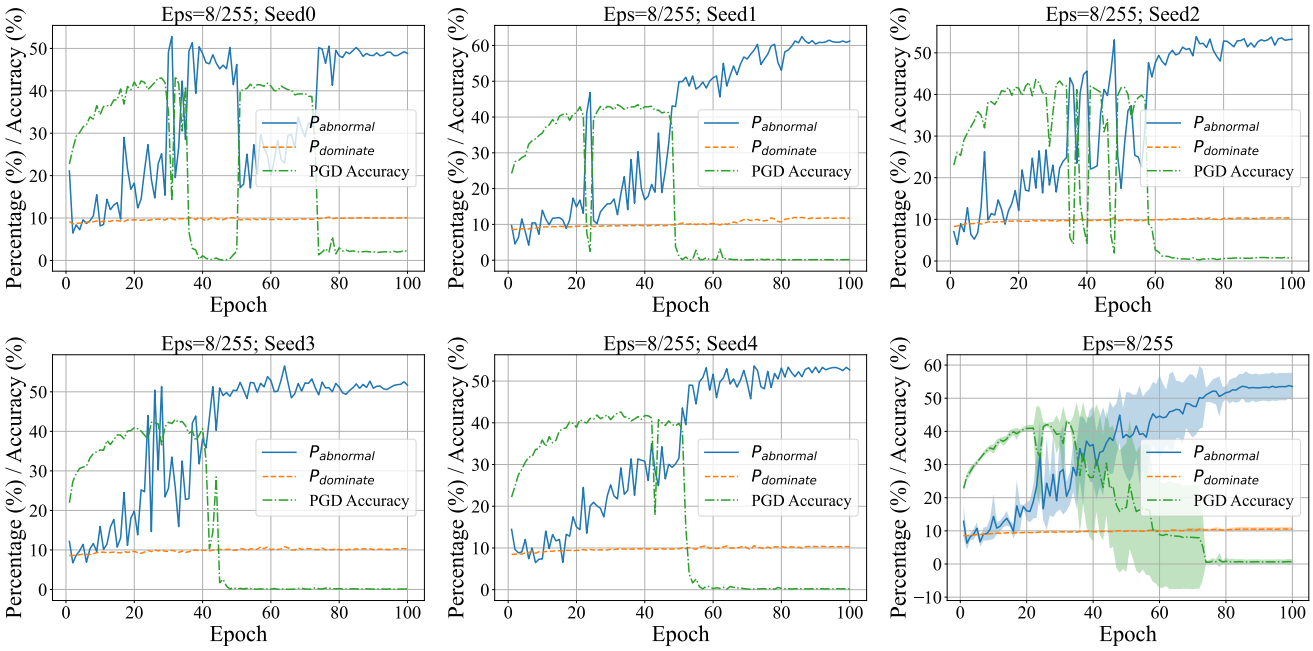

Figure 7. Transferability of label information for uniform gray image (value 0.5) on the CIFAR10 dataset with a perturbation size of 8/255.

20 and label smoothing value to 0.4.

As highlighted in our paper, initializing the training samples, $\boldsymbol{x}$, by either adding or subtracting $\boldsymbol{LI_c}$ proved effective in diminishing the label information, y, from the generated perturbation, $\boldsymbol{\delta_x}$. This initialization strategy was employed randomly to enhance the unpredictability of our defense mechanism against adversarial inputs.

In line with our strategy to boost diversity within the model, we randomly substituted 10% to 50% of the elements in $\boldsymbol{LI_c}$ with values uniformly distributed between $-\epsilon$ and ϵ when the perturbation size was 8/255. For a perturbation size of 16/255, we randomly substituted 0% to 100% of the elements in $\boldsymbol{LI_c}$ with values uniformly distributed between -2ϵ and 2ϵ. This approach was aimed at enriching the robustness of our model against adversarial attacks. To ensure our model adapts to evolving adversarial tactics, we updated $\boldsymbol{LI_c}$ at different intervals depending on the dataset: every 10 batches for CIFAR10 and every 20 batches for CIFAR100 and Tiny-ImageNet, aligning with our strategy to maintain model resilience over time.

To evaluate the robustness of our model, we subjected it to several adversarial attack methods, including PGD [16] and AutoAttack (AA) [2]. We varied the number of iterations for PGD attacks to 10, 20, and 50, which are henceforth referred to as PGD-20, and PGD-50, respectively.

The maximum allowed perturbation was set to 8/255 and 16/255 for CIFAR-10 and CIFAR-100 datasets, respectively. For Tiny-ImageNet, we used perturbation bounds of 8/255 and 12/255. These dataset-specific perturbation settings reflect the different sensitivity levels of each dataset to adversarial attacks and provide a more comprehensive evaluation of our defense mechanism.

D. Experiment Results for Section 4.2

This section supplements the results presented in Section 4.2 of the main paper by providing more detailed experimental findings. Specifically, Tables 6, 7, and 8 display the average results from three sets of experiments: CIFAR-10 and CIFAR-100 using ResNet-18, and Tiny ImageNet using PreActResNet-18. These results include clean accuracy and robust accuracy measured against PGD-10, PGD-20, PGD-50, and AutoAttack. Additionally, Tables 9 and 10 present experimental results for CIFAR-10 and CIFAR-100 using the larger WideResNet-28-10 architecture. Due to the computational demands of this larger model, these experiments were conducted only once.

For the training costs reported in Tables 6, 7, and 8, we measured training time on an NVIDIA V100 GPU. To calculate computational complexity (PFLOPs), we approximated the backward propagation cost as equivalent to the forward propagation cost. The total FLOPs for each method were determined by calculating the number of forward and backward passes required during training.

It is worth noting that while some methods have identical FLOPs calculations in our tables, their training times differ significantly. This discrepancy arises because a sub-

stantial portion of training time is consumed by parameter updates, and some methods require maintaining computational graphs in memory, which introduces additional overhead not captured in FLOPs measurements alone.

Method	Clean (%) ↑	PGD-10 (%) ↑	PGD-20 (%) ↑	PGD-50 (%) ↑	AA (%) ↑	Training Cost
PGD-AT [16]	82.32 ± 0.39 70.91 ± 1.37	53.76 ± 0.18 37.10 ± 0.25	52.83 ± 0.11 28.41 ± 0.09	52.60 ± 0.13 25.80 ± 0.36	48.68 ± 0.12 20.07 ± 0.05	353.70 min 59.87 PFLOPs
PGD-AT-WA [24]	82.00 ± 0.38 71.10 ± 0.31	54.90 ± 0.15 41.09 ± 0.07	54.21 ± 0.15 33.22 ± 0.25	54.08 ± 0.12 31.17 ± 0.21	50.26 ± 0.14 25.09 ± 0.21	358.21 min 59.87 PFLOPs
N-FGSM [3]	78.79 ± 0.46 63.09 ± 2.38	53.78 ± 0.13 **39.29 ± 0.45**	53.28 ± 0.07 33.00 ± 1.05	53.17 ± 0.09 31.65 ± 1.36	47.97 ± 0.12 22.50 ± 1.33	74.15 min 10.89 PFLOPs
FGSM-RS [28]	73.06 ± 10.15 66.21 ± 7.19	48.22 ± 7.64 15.11 ± 5.15	47.77 ± 7.46 9.33 ± 6.20	47.68 ± 7.46 5.42 ± 5.12	42.82 ± 6.70 0.00 ± 0.00	74.36 min 10.89 PFLOPs
FGSM-PGI [13]	80.32 ± 1.09 88.06 ± 0.35	56.36 ± 0.19 19.02 ± 0.34	55.81 ± 0.13 12.96 ± 0.38	55.70 ± 0.10 8.78 ± 0.28	49.73 ± 0.17 0.11 ± 0.00	99.71 min 10.89 PFLOPs
FGSM-UAP [20]	79.17 ± 0.27 88.07 ± 0.25	56.60 ± 0.03 15.56 ± 1.35	56.10 ± 0.04 9.72 ± 1.11	55.89 ± 0.02 6.03 ± 0.91	49.36 ± 0.11 0.03 ± 0.00	108.71 min 16.33 PFLOPs
NuAT [27]	80.78 ± 0.55 **91.82 ± 0.12**	55.43 ± 0.08 14.93 ± 0.61	54.79 ± 0.04 7.29 ± 0.63	54.64 ± 0.05 3.51 ± 0.37	**50.04 ± 0.13** 0.13 ± 0.02	127.81 min 16.33 PFLOPs
Grad-Align [1]	78.69 ± 1.13 46.93 ± 26.13	53.52 ± 0.16 27.72 ± 12.53	52.99 ± 0.20 23.09 ± 9.28	52.93 ± 0.20 21.82 ± 8.42	47.99 ± 0.15 15.43 ± 3.94	228.64 min 16.33 PFLOPs
FGSM-AT [7]	**90.98 ± 0.46** 79.75 ± 1.64	38.66 ± 1.92 15.15 ± 2.49	28.91 ± 1.36 10.13 ± 2.89	19.47 ± 1.65 5.98 ± 2.41	0.00 ± 0.00 0.00 ± 0.00	74.28 min 10.89 PFLOPs
Free-AT [25]	81.99 ± 0.95 89.15 ± 0.54	52.23 ± 0.15 31.91 ± 2.17	51.60 ± 0.09 21.66 ± 1.98	51.42 ± 0.08 12.47 ± 0.50	47.43 ± 0.14 0.00 ± 0.00	70.53 min 10.89 PFLOPs
COAT [14]	83.93 ± 0.34 84.58 ± 5.83	53.10 ± 0.22 36.27 ± 6.44	52.38 ± 0.17 24.64 ± 6.94	52.33 ± 0.09 18.17 ± 1.52	37.99 ± 0.31 4.58 ± 3.24	90.03 min 13.61 PFLOPs
GAT [26]	85.12 ± 0.04 65.02 ± 38.91	55.03 ± 0.05 9.48 ± 0.79	54.23 ± 0.20 5.98 ± 2.88	54.05 ± 0.21 4.44 ± 3.94	49.39 ± 0.22 3.14 ± 4.41	126.75 min 16.33 PFLOPs
LIET (Ours)	80.61 ± 0.44 52.72 ± 3.51	**56.70 ± 0.06** 37.17 ± 2.11	**56.14 ± 0.05** **33.55 ± 1.59**	**56.08 ± 0.07** **33.10 ± 1.44**	50.01 ± 0.09 **25.22 ± 0.41**	101.29 min 10.93 PFLOPs

Table 6. Comparison of clean accuracy, robust accuracy and training cost (time in minutes and computation in PFLOPs) on CIFAR-10 using ResNet-18. Each method is evaluated with perturbation sizes of 8/255 (first row) and 16/255 (second row). Best results are highlighted in **bold**.

Method	Clean (%) ↑	PGD-10 (%) ↑	PGD-20 (%) ↑	PGD-50 (%) ↑	AA (%) ↑	Training Cost
PGD-AT [16]	57.52 ± 0.95 48.38 ± 2.04	29.60 ± 0.23 17.03 ± 0.18	28.99 ± 0.21 12.66 ± 0.09	28.87 ± 0.27 11.56 ± 0.16	25.48 ± 0.11 9.17 ± 0.12	353.65 min 59.88 PFLOPs
PGD-AT-WA [24]	56.48 ± 1.34 45.84 ± 1.58	32.51 ± 0.31 22.44 ± 0.10	32.23 ± 0.26 18.09 ± 0.16	32.22 ± 0.25 17.45 ± 0.27	26.84 ± 0.28 12.66 ± 0.15	358.80 min 59.88 PFLOPs
N-FGSM [3]	54.77 ± 1.68 39.39 ± 2.11	30.70 ± 0.27 19.90 ± 0.30	30.47 ± 0.26 16.52 ± 0.04	30.41 ± 0.29 15.95 ± 0.18	25.31 ± 0.28 11.18 ± 0.21	74.62 min 10.89 PFLOPs
FGSM-RS [28]	37.59 ± 15.43 3.34 ± 0.58	20.94 ± 9.59 1.73 ± 0.35	20.86 ± 9.54 1.62 ± 0.33	20.84 ± 9.55 1.61 ± 0.32	16.70 ± 7.84 1.05 ± 0.21	74.62 min 10.89 PFLOPs
FGSM-PGI [13]	56.02 ± 0.21 57.29 ± 9.61	32.70 ± 0.14 3.85 ± 1.87	32.32 ± 0.10 2.08 ± 1.23	32.31 ± 0.11 1.50 ± 1.33	26.76 ± 0.07 0.57 ± 0.70	99.85 min 10.89 PFLOPs
FGSM-UAP [20]	53.54 ± 0.49 44.39 ± 28.33	32.14 ± 0.05 3.18 ± 0.66	31.83 ± 0.04 1.89 ± 0.10	31.81 ± 0.05 1.35 ± 0.47	26.29 ± 0.03 0.55 ± 0.67	114.03 min 16.33 PFLOPs
NuAT [27]	57.72 ± 2.01 62.30 ± 0.08	25.82 ± 1.28 10.72 ± 0.06	22.99 ± 0.94 5.92 ± 0.03	21.62 ± 0.58 4.07 ± 0.11	13.77 ± 0.58 1.94 ± 0.12	131.49 min 16.33 PFLOPs
Grad-Align [1]	54.87 ± 1.17 22.97 ± 13.38	31.86 ± 0.19 11.36 ± 6.63	31.60 ± 0.19 9.63 ± 5.58	31.54 ± 0.23 9.50 ± 5.47	26.18 ± 0.07 6.43 ± 3.93	229.34 min 16.33 PFLOPs
FGSM-AT [7]	21.36 ± 8.36 1.45 ± 0.45	3.04 ± 0.67 0.96 ± 0.04	2.38 ± 0.44 0.85 ± 0.15	1.88 ± 0.32 0.82 ± 0.17	0.19 ± 0.03 0.51 ± 0.49	74.17 min 10.89 PFLOPs
Free-AT [25]	58.29 ± 2.10 20.93 ± 4.08	30.47 ± 0.35 11.10 ± 2.49	30.01 ± 0.30 9.87 ± 2.21	29.96 ± 0.20 9.79 ± 2.21	24.34 ± 0.42 5.84 ± 1.56	71.97 min 10.89 PFLOPs
COAT [14]	**67.56 ± 1.13** 65.67 ± 0.27	24.55 ± 0.16 10.85 ± 1.01	23.23 ± 0.34 6.05 ± 0.73	22.70 ± 0.33 4.45 ± 0.61	18.93 ± 0.07 2.75 ± 0.38	88.85 min 13.61 PFLOPs
GAT [26]	65.24 ± 0.26 **72.42 ± 0.58**	27.61 ± 0.14 3.19 ± 0.04	26.69 ± 0.19 1.42 ± 0.02	26.54 ± 0.16 0.57 ± 0.06	21.97 ± 0.15 0.06 ± 0.03	126.85 min 16.33 PFLOPs
LIET (Ours)	51.52 ± 0.38 35.00 ± 2.85	**32.92 ± 0.12** **20.34 ± 0.86**	**32.75 ± 0.15** **17.60 ± 0.45**	**32.74 ± 0.14** **17.22 ± 0.42**	**27.05 ± 0.09** **12.04 ± 0.30**	103.74 min 11.11 PFLOPs

Table 7. Comparison of clean accuracy, robust accuracy and training cost (time in minutes and computation in PFLOPs) on CIFAR-100 using ResNet-18. Each method is evaluated with perturbation sizes of 8/255 (first row) and 16/255 (second row). Best results are highlighted in **bold**.

Method	Clean (%) ↑	PGD-10 (%) ↑	PGD-20 (%) ↑	PGD-50 (%) ↑	AA (%) ↑	Training Cost
PGD-AT [16]	43.60 ± 2.45 41.27 ± 2.45	20.20 ± 1.82 14.85 ± 0.27	19.90 ± 1.41 13.44 ± 0.04	19.86 ± 1.23 13.15 ± 0.00	16.00 ± 1.02 9.54 ± 0.19	2432.63 min 479.01 PFLOPs
PGD-AT-WA [12]	46.23 ± 0.85 41.89 ± 0.20	26.09 ± 0.36 19.91 ± 0.17	26.06 ± 0.34 18.60 ± 0.03	26.06 ± 0.34 18.43 ± 0.14	19.62 ± 0.24 12.31 ± 0.02	2439.34 min 479.01 PFLOPs
N-FGSM [3]	47.73 ± 0.45 37.98 ± 1.31	25.30 ± 0.26 18.11 ± 0.05	25.18 ± 0.24 16.94 ± 0.08	25.10 ± 0.23 16.64 ± 0.14	18.76 ± 0.24 11.11 ± 0.18	498.95 min 87.09 PFLOPs
FGSM-RS [28]	43.10 ± 4.09 5.64 ± 0.03	22.91 ± 1.35 2.33 ± 0.01	22.71 ± 1.30 2.26 ± 0.00	22.70 ± 1.27 2.25 ± 0.01	16.28 ± 1.03 1.33 ± 0.01	493.77 min 87.09 PFLOPs
FGSM-PGI [13]	48.59 ± 0.19 24.81 ± 1.29	**26.68 ± 0.25** 5.24 ± 0.06	26.46 ± 0.15 4.29 ± 0.04	26.39 ± 0.17 4.02 ± 0.14	19.52 ± 0.07 1.33 ± 0.27	652.09 min 87.09 PFLOPs
FGSM-UAP [20]	45.69 ± 0.99 16.89 ± 9.00	26.12 ± 0.05 2.55 ± 0.10	25.91 ± 0.02 2.05 ± 0.27	25.84 ± 0.04 1.89 ± 0.40	19.45 ± 0.14 0.71 ± 0.59	717.3 min 130.64 PFLOPs
NuAT [27]	45.55 ± 0.99 47.98 ± 4.95	26.51 ± 0.17 13.83 ± 2.69	26.38 ± 0.17 11.25 ± 4.00	26.37 ± 0.14 10.59 ± 4.58	19.55 ± 0.12 5.47 ± 3.37	865.64 min 130.64 PFLOPs
Grad-Align [1]	46.16 ± 2.03 36.83 ± 2.61	24.61 ± 0.51 15.57 ± 0.76	24.30 ± 0.40 14.43 ± 0.79	24.21 ± 0.41 14.22 ± 0.72	17.65 ± 0.28 8.71 ± 0.59	1514.23 min 130.64 PFLOPs
FGSM-AT [7]	34.68 ± 15.85 5.78 ± 1.95	17.12 ± 8.69 1.75 ± 0.08	16.92 ± 8.60 1.62 ± 0.16	16.84 ± 8.55 1.58 ± 0.21	16.60 ± 0.36 0.84 ± 0.29	492.96 min 87.09 PFLOPs
Free-AT [25]	48.19 ± 1.78 32.57 ± 3.72	23.85 ± 0.24 14.89 ± 1.00	23.63 ± 0.25 14.07 ± 0.75	23.58 ± 0.26 13.99 ± 0.73	16.34 ± 0.32 8.42 ± 0.58	474.87 min 87.09 PFLOPs
COAT [14]	**59.30 ± 0.38** **58.52 ± 1.41**	18.45 ± 0.08 10.65 ± 0.96	17.57 ± 0.16 8.15 ± 0.84	17.25 ± 0.22 7.38 ± 0.86	11.56 ± 0.02 3.71 ± 0.60	578.07 min 108.87 PFLOPs
GAT [26]	57.68 ± 0.25 33.84 ± 0.91	17.97 ± 0.23 1.02 ± 0.07	17.26 ± 0.24 0.69 ± 0.04	17.01 ± 0.28 0.60 ± 0.02	11.53 ± 0.15 0.20 ± 0.01	851.30 min 130.64 PFLOPs
LIET (Ours)	44.73 ± 0.33 35.18 ± 0.76	**26.68 ± 0.11** **18.43 ± 0.13**	**26.54 ± 0.12** **17.47 ± 0.10**	**26.54 ± 0.11** **17.31 ± 0.19**	**19.79 ± 0.04** **11.29 ± 0.15**	682.47 min 90.56 PFLOPs

Table 8. Comparison of clean accuracy, robust accuracy and training cost (time in minutes and computation in PFLOPs) on Tiny-ImageNet using PreResNet-18. Each method is evaluated with perturbation sizes of 8/255 (first row) and 12/255 (second row). Best results are highlighted in **bold**.

Method	Clean (%) ↑	PGD-10 (%) ↑	PGD-20 (%) ↑	PGD-50 (%) ↑	AA (%) ↑
N-FGSM [3]	80.60 68.10	56.06 39.17	55.76 31.62	55.67 29.86	50.31 21.66
FGSM-RS [28]	89.23 9.99	15.94 9.82	8.94 9.67	5.50 9.56	0.00 7.51
FGSM-PGI [13]	84.25 89.75	**60.03** 22.92	59.36 18.08	59.23 14.02	52.99 0.30
NuAT [27]	82.68 93.17	57.55 19.34	56.99 10.61	56.94 5.71	52.50 0.12
Grad-Align [1]	82.55 10.00	56.55 9.97	55.99 9.73	55.91 9.12	50.57 2.95
FGSM-AT [7]	**89.99** 11.46	27.07 9.73	18.51 9.51	11.01 9.34	0.01 7.20
Free-AT [25]	76.98 35.74	51.43 21.77	51.04 20.12	50.93 20.11	46.30 15.48
COAT [14]	86.69 **94.51**	53.47 34.91	53.08 20.97	53.11 8.29	41.90 0.04
GAT [26]	79.08 93.23	45.45 11.16	45.02 4.59	44.93 1.53	40.39 0.03
LIET (Ours)	82.49 68.32	59.99 **42.76**	**59.53** **35.96**	**59.48** **33.98**	**53.15** **23.84**

Table 9. Comparison of clean accuracy and robust accuracy on CIFAR-10 using WideResNet-28-10. Each method is evaluated with perturbation sizes of 8/255 (first row) and 16/255 (second row). Best results are highlighted in **bold**.

Method	Clean (%) ↑	PGD-10 (%) ↑	PGD-20 (%) ↑	PGD-50 (%) ↑	AA (%) ↑
N-FGSM [3]	57.55 43.49	**33.33** **20.50**	32.91 16.54	32.85 15.76	27.17 12.17
FGSM-RS [28]	64.43 2.82	8.31 1.07	5.94 0.91	4.06 0.88	0.00 0.40
FGSM-PGI [13]	**70.84** 57.42	19.50 9.99	15.89 6.31	13.39 5.29	1.54 3.50
NuAT [27]	66.23 67.24	22.87 12.45	19.50 6.97	17.98 4.41	12.84 1.68
Grad-Align [1]	26.80 1.80	12.25 1.34	12.19 1.13	12.20 1.01	10.00 0.64
FGSM-AT [7]	65.09 1.00	4.23 0.97	1.95 0.60	1.11 0.48	0.00 0.00
Free-AT [25]	49.72 20.94	27.57 10.64	27.49 9.21	27.48 9.28	22.20 5.67
COAT [14]	70.81 **77.49**	25.97 13.93	23.96 7.98	23.02 3.21	20.16 0.00
GAT [26]	70.66 74.93	27.62 4.30	26.67 1.84	26.62 0.93	23.09 0.04
LIET (Ours)	52.72 35.36	33.30 20.17	**33.16** **17.47**	**33.04** **17.19**	**27.33** **12.72**

Table 10. Comparison of clean accuracy and robust accuracy on CIFAR-100 using WideResNet-28-10. Each method is evaluated with perturbation sizes of 8/255 (first row) and 16/255 (second row). Best results are highlighted in **bold**.