We thank all reviewers for their helpful and constructive comments. We'll further improve the paper in the final version. Below we address their detailed comments.

R1: The results for AT$_{\text{PGD}}$ seem below the state-of-the-art: We need to clarify that the AT$_{\text{PGD}}$ model is trained by following the experimental settings in [36]. We found that the training configuration of the state-of-the-art AT$_{\text{PGD}}$ in [*1] pointed out by R1 differs from [36] in several aspects, including early stopping, weight decay factor, and the number of PGD steps. We also need to point out that the model which achieves 56% robust accuracy and 87% natural accuracy in [*1] is a Wide-ResNet-34-10 model (Table 1 in [*1]). Their smaller model (i.e., PreActResNet18) achieves 53% robust accuracy (Table 2 in [*1]). Besides, the robust accuracy is evaluated by PGD-10 in [*1], which is a weaker adversary than we used in experiments. To fairly compare with the state-of-the-art, we reproduce the results of [*1] and train ADT based models using the same settings/hyperparameters as in [*1]. The results of those models on CIFAR-10 are shown in Table A. By using the same training settings, our models can also improve the performance over AT$_{\text{PGD}}$. We'll include the results in the final version.

Table A: Model accuracy (%) on CIFAR-10 following [*1].

Model	$\mathcal{A}_{\text{nat}}$	PGD-10	PGD-20	PGD-100
AT$_{\text{PGD}}$	86.41	55.90	54.52	54.20
ADT$_{\text{EXP}}$	86.49	**56.84**	**55.43**	**55.01**
ADT$_{\text{EXP-AM}}$	87.27	56.28	54.88	54.58
ADT$_{\text{IMP-AM}}$	**87.38**	56.63	55.10	54.43

R1: Confidence intervals/multiple trials: In Table B, we show the mean and standard deviation of accuracy of AT$_{\text{PGD}}$ and ADT based models over 3 runs (using the submitted code). The standard deviation is small compared with the performance gap. We'll include the full results in the final version.

Table B: Model accuracy (%) on CIFAR-10 over 3 runs.

Model	$\mathcal{A}_{\text{nat}}$	PGD-20	PGD-100
AT$_{\text{PGD}}$	86.50±0.14	49.77±0.21	49.34±0.27
ADT$_{\text{EXP}}$	87.15±0.13	52.38±0.23	51.89±0.22
ADT$_{\text{EXP-AM}}$	87.30±0.09	53.01±0.22	52.45±0.28
ADT$_{\text{IMP-AM}}$	87.58±0.14	51.90±0.15	50.94±0.16

R1: ℓ_2 adversarial constraint: We need to clarify that we consider the ℓ_∞ norm constraint in this paper. However, our methods can be easily extended to the ℓ_2 norm. We agree that PGD is effective to find local maxima of the inner problem, but we show in Fig. 1 that the adversarial distributions can better explore the space of possible perturbations and characterize more diverse adversarial examples, resulting in more robust models, as discussed in Sec. 2.2.1.

R1: A new robustness constraint: Thanks for the insightful comment. We think that the proposed ADT framework is flexible to integrate a new robustness constraint. We'll consider this in future work.

R2: ADT is trained by one attack that operates on probability measures instead of individual samples: Yes, ADT uses a single attack which can find a distribution of adversarial examples instead of an individual sample. We have discussed in Sec. 2.2.1 the superiority of our approach upon others which generate individual adversarial examples by a single attack. We'll further polish our arguments in the final version to make them not misleading.

R2: To what extend the entropic regularization allows to find adversarial and sufficiently diverse examples: When using no entropic regularization, ADT degenerates into AT such that the adversarial examples are not diverse. When using a very large entropic regularization, the generated examples are diverse, but are not adversarial enough. Thus, we use a hyperparameter λ to control the strength of the entropy term in Eq. (5). As it's hard to derive the optimal value for λ, we did an ablation study on the effects of λ in Fig. 5. Our results suggest that choosing an appropriate λ (e.g., 0.01) can ensure the generated examples being both adversarial and diverse for learning a robust model.

R2: Another attack might be developed that performs well against ADT: Just like other empirical defenses, we cannot guarantee that there aren't any attacks that can beat our defenses. However, we have tried our best to evaluate the robustness of our defenses, including adopting a plenty of attacks, calculating the per-example accuracy, evaluating black-box attacks, and visualizing the loss landscape. Experiments suggest that the common failure modes [2,6,56] of previous defenses do not occur in our method. We'll also release our code and pre-trained models for future evaluations.

R2: Being clear about the attacks known when each of the baseline methods were proposed: One of the challenges of adversarial robustness research is that there exists a "cat-and-mouse" game between attacks and defenses, i.e., the defenses were later shown to be ineffective against new attacks, which has drawn much attention in this field [2,6,56]. Therefore, it's important to develop robust models that not only are robust to existing attacks but can also generalize to new ones [49], which is also the main motivation of our work. Although FeaAttack was proposed later than FeaScatter, it can also prove the ineffectiveness of FeaScatter. As above, we have tried our best to evaluate the worst-case robustness of our defenses following the guidelines in [6], and we're willing to test our models by future attacks continuously. We do believe that our defenses can generalize to new attacks better than the baselines.

R3: Related works on worst-case distribution: Thanks for the suggestion. We'll discuss them in the final version.

R4: The degenerated solution of ADT: When $\lambda = 0$, the adversarial distribution degenerates into a Dirac distribution and ADT becomes AT. So we expect that the performance of ADT ($\lambda = 0$) matches the performance of AT$_{\text{PGD}}$. As can be seen from Fig. 5, the model trained with $\lambda = 0$ gets about 50% accuracy against attacks, which is similar to the results of AT$_{\text{PGD}}$. But with the entropic regularization, ADT obtains more than 2% accuracy improvements, as shown in Fig. 5. We'll also show the results of ADT$_{\text{EXP}}$ with different λ in the final version.

[*1] L. Rice, E. Wong, J.Z. Kolter. Overfitting in adversarially robust deep learning. ICML 2020.