We thank the reviewers for their efforts and suggestions.

R1. There lacks of comparison with existing backdoor attacks. In terms of methodology, we discussed the previous attack methods in Sec 2.2. We pointed out their common weakness of using universal backdoor trigger pattern(s). Our proposed method overcomes this problem by using a trigger generator targeting on diverse and non-reusable patterns.

Empirically, there is no direct way to compare our method to others, and no such comparison was presented in the previous attack papers [3,2,10]. However, we can compare them indirectly by examining with backdoor defenses. The popular attacks such as BadNets and TrojanAttack can be detected by NeuralCleanse, as reported in that paper, while ours can surpass it. As for recent methods, we re-implemented them on the GTSRB dataset and examined them with GradCAM. The results are shown in Fig. 1a with Dynamic Backdoor [10] (top row) and Programmable Attack (bottom row). Unlike our method, both get caught by GradCAM since the classification prediction depends entirely on the small backdoor region. We will add this result to our revised paper.

R1. The idea of dynamic backdoor is similar to Programmable Attack. In that work, the trigger pattern is generated from the target image; it is still independent of the source image content. A pattern generated on one target image can be reused on any input image. Hence, it is still vulnerable to many defense methods. For example, GradCAM can spot the trigger as shown in Fig. 1a. Instead, our paper proposes generating non-reusable trigger patterns conditioned on the source images, achieved by novel constraints, leading to the capability to dodge all the mentioned defense mechanisms.

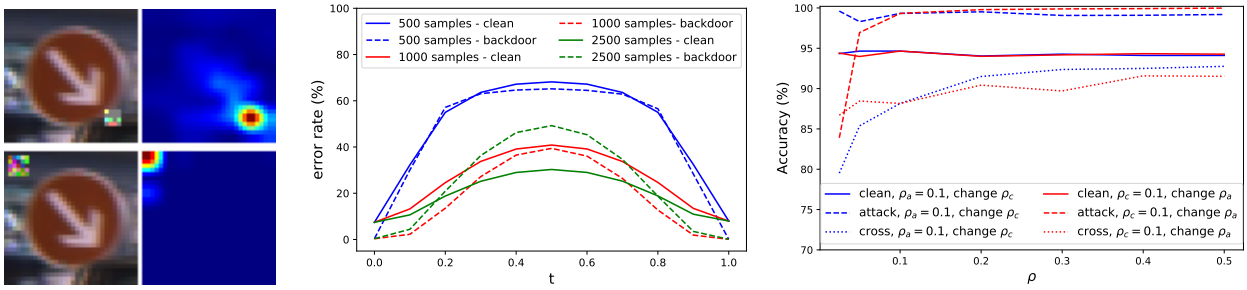
R2. Any algorithm that efficiently generates adversarial sample can be served as a dynamic backdoor algorithm. There are several differences between adversarial attacks and the proposed backdoor method. First, in adversarial attacks, we have no control over the deep model, making attack success uncertain. That model may be trained to adversarially robust to a wide range of attack algorithms. In backdoor attacks, we have total control on the model, guaranteeing a near 100% attack success rate. Second, adversarial attacks rely on expensive optimization processes. The computation is even more costly in black-box configurations. With our method, the attack process is fast and straightforward by running GPU-accelerated generators. Finally, due to universal adversarial examples, the adversarial generation on different input images can sometimes converge to the same adversarial pattern, violating our requirement. We will clarify this in our revised paper.

R2. The added backdoor is very visible according to Fig. 4. While the backdoor patterns are apparent to humans, like most of the previous backdoor attacks, they are still effective in fooling automated systems. As discussed in Sec 5, we plan to make them more realistic and imperceptible to humans in future studies.

R3. Can the proposed attack surpass the defense in the Bridging Mode Connectivity paper? We ran that method on our backdoor models on CIFAR-10 and reported the error rates in Fig. 1b. As can be seen, backdoor error rate is close to the corresponding clean error rate. The largest gap between them is only 20% when using 2500 training samples and t = 0.5. Hence, this defense method cannot mitigate our backdoor.

R3. Unclear if the reported results are from training or testing data. All of our results were computed on testing data, and details of data splits were reported in the supplementary PDF. It means our backdoored classifier, just by studying on training data, can recognize unseen backdoor patterns generated from unseen test images. This generalization ability is, in fact, a surprising and very important feature of our method. We will highlight it in the final version if this paper gets accepted.

R3. Ablation studies on ρ_a and ρ_c. We report the results on CIFAR-10 in Fig. 1c. Overall, our method is quite stable with high accuracy on clean, attack, and cross tests. When increasing ρ_c, the cross accuracy increases from 80% to 93%. When increasing ρ_a, the attack success rate goes up to near 100%, and the cross accuracy also surprisingly increases.



(a) (b) (c)

Figure 1: Extra experimental results