We thank all three reviewers for their constructive comments. Our responses are as follows:

Reviewer 1.

1. The parameters (including ensemble sizes) for the models are presented in Table 3 in Appendix B in the supplemental materials.

2. Figure 3 in the main text shows the running time of MILP and our method on GBDT models with different T, L and numbers of trees on the ijcnn1 dataset. There, we see that T and L have a modest impact on running time, with a larger (though still slower than MILP) impact from the number of trees. We see that T (clique size per level) has a smaller impact on running time than L (the number of levels).

3. Thank you for the question and reminder to upload our code. To comply with NeurIPS rules, we have sent an anonymous GitHub link with our code to AC and will let AC decide whether to release it to reviewers.

Reviewer 2.

1. We note that our perturbation-sensitive notion of feature importance is complementary to the conventional tree/forest feature importance, with several critical differences. In Figure 1 below we show the feature importance map of the standard and robust models used in Figure 4 in the main text. A feature's importance is measured by the average gain across all the splits it is used in. Pixels with darker color have larger importance and yellow pixels have zero importance. Our single-feature robustness bounds for different features shown in Figure 4 are different from importance scores in the following ways:

- Feature importance scores only depend on the model itself. But our single-feature robustness bound depends on both the model and the testing data point, and for different data points, the model may be sensitive to different features.
- Feature importance is a relatively heuristic score. But our robustness bound can give a formal guarantee that the model output would not change if this single feature is perturbed within the given range.
- A feature importance score assigns non-zero importance to more pixels than our method in general.

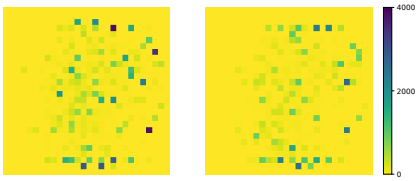

Figure 1: Feature importance of the model in Figure 4 in the main text. Left: standard GBDT model; Right: robust GBDT model. Yellow pixels have zero feature importance while darker pixels have larger importance.

2. The reviewer asks about the relationship of different attacks and methods seeking to improve robustness ([8], [12], [18]), with respect to impacts on our verification bounds. In response, we first confirm that, as suggested in [8], adversarial examples indeed exist in tree-based models. In Figure 1 of [8] the authors used a black-box attack developed in [12] to generate adversarial examples on MNIST and Fashion-MNIST. The noise added to generate these adversarial examples are so small that is indistinguishable for human eyes but can completely fool the model. Authors in [18] proposed an even stronger MILP based white-box attack designed specifically for GBDT models to find the best adversarial example with the exact minimum distortion. The robust training strategies introduced in [8] and [18] can indeed increase model robustness, and thus our verification bounds also increase in value correspondingly. In this case, our bounds are still a lower bound with respect to the exact MILP bound. Note that our verification bound is used to evaluate model robustness, instead of attacking models. Table 2 shows our bounds applied to robust models trained using [8] and the bounds indeed increase compared to standard model bounds in Table 1. The tightness of the bound depends on the model, and one cannot guarantee that robust model bounds are tighter than of a standard model. For example, in Tables 1 and 2, on covtype dataset, the standard model's bound is tighter than robust model's bound when we use the same T and L values. However, intuitively, the robust training method in [8] has a regularization effect, so the model tends to be sparser with fewer boxes and simpler overlapping structures. We believe that this also contributes to robustness of the models and tighter bounds.

3. MILP is also an adversarial attack method, and the bound obtained by MILP is the **exact** minimum distortion with formal mathematical guarantees. Therefore MILP is used as a ground truth benchmark to test our verification bounds, and it is not possible to obtain a tighter bound than it. However, MILP can be very slow in large models. Our bound is much faster than MILP, and very tight compared to it as shown in the experiments.

Reviewer 3.

1. ℓ_∞ balls are larger than all ℓ_p balls with the same radius. Since our current bound can guarantee that no adversarial example exists in an ℓ_∞ ball with radius R, we can also guarantee that no adversarial example exists in any ℓ_p balls with the same radius. We will leave the development of bounds designed specifically for other ℓ_p norms for future studies.

2. If you mean "why LP relaxation bound is much worse than MILP and our bound?", the reason is that LP relaxation relaxes all 0-1 constraints in MILP to [0,1] and the search space is much larger than MILP. Therefore a much smaller distortion is found by LP relaxation. However, since this relaxation is so loose, the data point found by LP relaxation is very likely not an adversarial example.