We thank all reviewers for their insightful feedback, please see answers inlined below.

R1: Limited novelty compared to HAWQ, paper is just adding Hessian trace. A: Please note that we are proposing a fully automated approach whereas HAWQ manually chooses bit precision for each layer. Also, despite being fully automated, our accuracy is still higher than HAWQ. Also we enabled mixed-precision activation which was not covered by HAWQ and is beneficial for object detection tasks.

R1: What if top eig. is highly dominating? A: That is an excellent question. We show top 20 Hessian eigenvalues in Fig. 1 for ResNet50 and InceptionV3. As one can see, the top eig. is not an outlier. Theoretically it is possible that top eig. be very large, but we did not observe it in practice. We will clarify this in the paper.

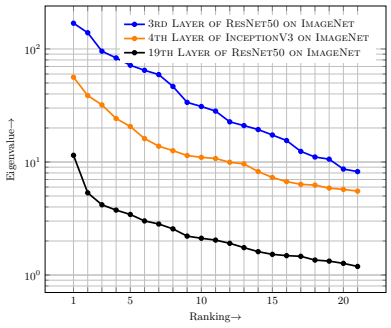

Figure 1: *The top-20 eigenvalues of ResNet50 (3rd/19th layers) and InceptionV3 (4th layer).*

R1, R3: Need to compare with HAQ. A: We had indeed included comparison with HAQ in Tab. 3 in Appendix (p.12 in Supplementary). We compared performance for ResNet50, InceptionV3, SqueezeNext. We found it's hard for HAQ agents to search very complicated models, and for all cases AMQ achieves higher precision with smaller model size with up to $100\times$ speedup for finding the bit precision setting.

R1: Tab1-2 comparisons are not fair, some methods use fixed bit. A: Please note that we do compare with HAQ and HAWQ which are mixed-precision methods. Also please note that prior work with low precision achieve very low accuracy. Also comparing with fixed bit precision is standard as HAQ also makes similar comparison (for example they compare with PACT which is fixed-bit quantization in Table 3 of their paper).

R1: It is difficult to understand Fig. 4. A: For a given compressed model size there are huge amount of different bit settings that obey the ordering based on Hessian sensitivity. Pareto Frontier method is a simple and efficient way to find the bit precision setting that results in the smallest second-order perturbation, jointly for every model size. We include detailed explanation of how to generate Fig. 4 in Appendix E.

R1: Missing relationship between total perturbation and accuracy. A: That is an excellent point. We performed an ablation study for SqueezeNext between the setting with high total perturbation (which achieves 67.46% accuracy with model size 1.09MB) and lowest perturbation (which achieves 68.68% accuracy with model size 1.07MB). As expected, configuration with lower perturbation results in better accuracy. We followed same training for these two runs. We will add detailed version of this ablation study for different models to the final version.

R2: Theory assumes strong assumption that $||\Delta W_i^*|| = ||\Delta W_j^*||$ A: We apologize for the confusion. Please note that the lemma still holds for cases where $||\Delta W_i^*|| \neq ||\Delta W_j^*||$ (this will introduce a constant coefficient in Eq. 4). We only used it in the proof for simplicity. Actually because of the $||\Delta W_i^*|| \neq ||\Delta W_j^*||$ cases, the lemma suggests to use $\overline{Tr}(H_i)||\Delta W_i^*||_2^2$ to determine the sensitivity, which is the exact expression in Eq 9. We will clarify in final version.

R2: Strong assumption that coefficients are identical for all the eigenvectors. A: The assumption can be relaxed to random coefficients as far as those random coefficients have the same 2nd moment, i.e., α_{bit} can be random variable for different directions but with same $\mathbf{E}[\alpha_{bit}^2]$. We will clarify this in the final version.

R2: Necessary to include ablation study with $||Q(W_i) - W_i||$. A: That is an excellent point. We conducted this ablation study for SqueezeNext (which is a hard task since it is already compact) following exactly the same fine-tuning settings. Using $||Q(W_i) - W_i||$ results in a final accuracy of 67.83% (1.10MB model size), where as AMQ achieves 68.68% (1.07 MB model size). We will add this ablation result for other models to the final version.

R3, R4: The improvement in performance seems moderate. A: While $0.5 \sim 0.6\%$ accuracy gain for ImageNet is actually considerable but we kindly note that the main point of our paper is a fully automated approach as opposed to the manual approach of HAWQ. The latter only provides relative sensitivity of layers and requires a data scientist to manually choose the bit precision for each layer. AMQ is fully automated and still achieves better results. It also up to $100\times$ faster for its end-to-end time to find the bit precision setting of each layer as compared to HAQ (Table 3 Appendix). We think this is quite significant improvement.

R3: It would be good to have HAWQ result for RetinaNet as well. A: RetinaNet with HAWQ achieves 33.5 mAP with model size 17.9MB while AMQ achieves 34.1 mAP with the same model size. We will add this to Tab. 2.

R4: Why does the fine-tuning perturbation have to lie in direction of the sum of H's eigenvectors? A: Since Hessian should be PSD at convergence, its eigenvectors form a complete space for the parameters.

R4: Additional assumptions on higher-order curvatures may be necessary. A: That is correct, and we will clarify this. We should also mention the prohibitive computational cost of evaluating higher order terms.