We thank the reviewers for their time and valuable feedback on our paper. We've responded to comments below:

Motivation and Contribution: In this study, our primary goal was to understand whether winning tickets contain generic inductive biases which apply to multiple problems or rather, whether winning tickets are simply overfit to the particular dataset and optimizer used to generate them. One of the most exciting aspects of the lottery ticket hypothesis is that it suggests that we may be able to generate new initialization schemes which can substantially improve training of neural networks from scratch without requiring any pruning or fine-tuning. However, such initialization schemes are only possible if winning tickets contain generic inductive biases. We therefore view the primary contribution of this study as a scientific one – understanding the generality of winning tickets – rather than a prescriptive one, though we hope that this study will help to lay the groundwork for improved training methods in the near future. Importantly, the absolute performance of winning tickets is independent of the generality of winning tickets.

Comparisons against other pruning methods (R1, R2): We completely agree with the reviewers that comparisons between winning ticket trained networks and more traditional model compression approaches are interesting and important problems, but we feel this an orthogonal question to the primary question of this study (how generic are winning tickets), especially since these comparisons are difficult to make fairly. Pruning approaches generally require the full model to be trained from scratch, after which the model is generally iteratively pruned and fine-tuned. In contrast, winning tickets (especially transferred winning tickets) can be trained from scratch, and do not require any further fine-tuning. As such, the costs of generating these models are quite different, making fair comparisons challenging. We will add a discussion point clarifying these differences in section 5.1.

As Reviewer 1 acknowledged, our study required massive amounts of compute, especially since we performed six replicates of each experiment. For example, we trained over 1500 ImageNet/Places365 models from scratch to generate the results in Figs. 3e,f and 4e,f. We therefore view such comparisons as beyond the scope of the present work.

Analysis of lottery tickets (R1) : We wholeheartedly agree with the reviewer that this an exciting and important line of inquiry (and one that we are currently pursuing!). However, detailed analysis of the structure of winning tickets is quite challenging since winning tickets and random tickets are often statistically quite similar (at least to the first few moments). We also agree that comparing pruning patterns to those derived from L_0 or L_1 regularization are interesting, but we haven't analyzed these yet. As such, we leave this problem for future work.

Rescaling of weights (R2): To be clear, after each pruning iteration, a new mask is generated, after which the subsequent model is trained from scratch with the new mask. Because converged weights are not used after pruning (as in typical compression approaches), there is no need to re-scale the weights. Additionally, the relevant comparisons for our primary question are between the transfer ticket performance, same dataset ticket performance, and random ticket performance, all of which have the same scaling of initialization.

Differences in input dimension (R2): For both the VGG19 model and the ResNet50, a global average pool is applied across all spatial dimensions prior to the final linear output layer. As a result, changes in input dimension do not require changes in model architecture. We will add a paragraph clarifying this point within section 3.3 in the final paper.

Global pruning scaling (R2): We have not evaluated rescaling the weights based on the pruning fraction. However, the relevant comparison here is between winning tickets and random tickets neither of which is rescaled. We therefore consider it unlikely that rescaling would change our core results since we have no reason to expect that rescaling would preferentially benefit winning or random tickets.

Clarification of target/source in lines 150-155 (R2): In order to evaluate the generality of winning tickets, winning tickets were generated by iteratively training and pruning a model on one dataset/optimizer ("source" configuration, as defined in lines 144-145), and evaluated on a second dataset/optimizer ("target" configuration). For standard lottery ticket experiments in which the source and target dataset are identical, each iteration of training represents the winning ticket performance for the model at the current pruning fraction. However, because the source and target dataset/optimizer are different for our experiments and because we primarily care about performance on the target dataset for this study, we must re-evaluate each winning ticket's performance on the target dataset, adding an additional training run for each pruning iteration. In the final paper, we will expand this section to better clarify this point.

Inconsistent colors (R1, R3): Thank you for pointing out the inconsistency in colors between plots! We agree that this makes the plots difficult to read and will correct the colors to be consistent across plots and figures in the final paper.

Inclusion of preserved mask (R2): While we feel the use of globally random masks is important to properly evaluate the performance of random tickets (because masks can leak substantial amounts of information from the final trained model to the initialization, as discussed in Section 3.1), we will add additional plots with the preserved mask case to the appendix for our primary results in the final paper.

Generality claims (R2): We agree with the reviewer that this claim is a little too strong since, indeed, some winning tickets generated on small datasets do not generalize to larger datasets. We will soften this claim in the final paper.