We thank all reviewers for their valuable and positive feedback, and share their excitement that MDEQ validates implicit models for large-scale vision applications. We address a few comments below, which we will incorporate into the paper.

Review #1. We first want to clarify a misunderstanding here: each resolution's representation is **not** treated independently. While they have their own residual blocks (which Eq. (5) describes), the features of each resolution are all subsequently mixed in the multi-resolution fusion step (Fig. 1). Both components (residual block and multiscale fusion) are in the f_θ that MDEQ iterates. Hence importantly, instead of having "independent losses and solvers", the whole MDEQ system has *only one solver*, and the equilibria are solved *simultaneously* (see L187-204 and Appendix C), with loss induced on resolution i can flow to the parameters of resolution $j \neq i$ via implicit differentiation. In other words, the multi-resolution interaction itself is a part of the transformation driven to the equilibrium. Indeed, to answer the reviewer's question, this is one of the primary contributions of MDEQ, and we hope this clarifies the confusion.

As a result, MDEQ is different from the original DEQ models and the Neural ODE models (which are both single-stream (see Fig. 3)). However, we found the particular choice of downsampling not important. For instance, we didn't downsample CIFAR-10 images at all, and downsampled for Cityscapes only to small degree by 2 convolutions (see Table 4 for "# of Downsamplings before Equilibrium Solver"). In contrast, NODEs typically downsample MNIST/CIFAR images by $4\times$ before ODE solvers, because the high-dim ODEs can be challenging to solve. The scales we used and the splitting of parameters can both be found in Appendix A (under "Hyperparameters" and Table 4); the parameters across models were kept equal by adjusting the "Width Expansion" factor in the residual block.

Finally, there is no difference in performance (empirically) for injecting the input $\mathbf{x}$ to 1) the highest resolution *only*; or 2) to *all resolutions*. In fact, we started off with the latter design but then simplified the model to become the #1 case that we use now. Besides this minimal empirical influence, injecting only to the highest resolution both *reduces* the need for extra parameters (i.e., one needs to downsample the signals to lower resolutions first, before injecting them to these resolutions) and *simplifies* the model by removing redundant (explicit) pre-processing of the input.

Review #2. The original purpose of Figure 4(a) was only to compare implicit models of the same size ($\sim$170K parameters, which was used by most ODE-based models but rarely by explicit networks like ResNet-18). But we are happy to add the curves for larger MDEQ and larger explicit models to Figure 4(a) as suggested by the reviewer. In terms of the hyperparameters shown in Table 4, we mostly followed the hyperparameters used by prior works such as HRNet [54] (e.g., weight decay, optimizer, etc.), while the threshold of Broyden iterations were adjusted according to the size of the original input – larger images typically need more quasi-Newton iterations to converge (see Figure 5).

Review #3. We absolutely agree and acknowledge that in practice, rather than "a layer", what MDEQ used to iterate in the implicit solver is actually more of a *block*, and will definitely clarify this language. As the reviewer identified, we found that at the finest level of MDEQ, such mini explicit transformation is still very helpful in terms of the representation driven to equilibrium. Regarding the downsampling of the original raw image, as mentioned above, its primary utility is to reduce the runtime of the model, rather than increase the accuracy. However, we do agree with R3 that the usage of explicit structure is still helpful in both *the design* and *the training* (e.g., shallow warmup) of MDEQ, and we will clarify this further in the paper.

Per the suggestion of the reviewer, we also performed an additional ablation study using a weight-tied (unrolled) version of an MDEQ(-large) layer on ImageNet (we used 5 layers, which is the max our GPU can fit). This model eventually achieved 75.8% accuracy, as compared to MDEQ-large's 77.5% (which we believe validates the benefit of modeling an "infinite layer"). Regarding the original DEQ's improvement on NODE/ANODE, we generally found that the usage of initial downsampling not critical for the final results on CIFAR-10. Empirically, even with the same $4\times$ smaller resolution, we are still able to achieve 87% accuracy with MDEQ-small and 93.5% with MDEQ, which is the same level of performance as currently reported in Table 1. When using full-resolution, we found that ODE-based models are significantly slower (especially the original NODE), and full resolution ANODE still reached only 60.3%.

Review #4. The rationale behind keeping all resolutions throughout were: 1) Having access to all resolutions is the key factor enabling MDEQ to be pretrained (e.g., on ImageNet) and subsequently finetuned (e.g., on Cityscapes) on very different resolution "interfaces", a common setting for many pattern recognition problems. 2) Having multiple scales side-by-side is actually one of the ways for us to reduce latency (rather than increase it), which we discuss in depth in Appendix C (and in Figure 6). Driving multiple streams to the equilibrium stabilizes the convergence to equilibrium, making the Broyden convergences typically faster.

To further investigate the effect of multi-resolution (i.e., multiple stream) itself, we have also compared the MDEQ we used with single-stream DEQ models in the paper. As can be seen in Table 1 and 2, the single-stream DEQ (which is essentially MDEQ with 1 resolution) performs substantially worse than when modeling multiple resolutions (though still much better that NODE/ANODE models). Figure 6 also shows that the convergence to the equilibrium is also much slower at the highest resolution. We can run further ablations investigating the effect of the number of resolutions.