We thank the reviewers for their thoughtful comments and suggestions. We are happy that reviewers see this line of work as useful for both neuroscientists and ML researchers. Here we paraphrase and respond to the main criticisms.

The results could reflect failures of optimization(R1, R4) R1 had questions about the criterion for deciding that a stimulus was a model metamer. R4 supposed that the optimization is harder for deeper layers and wondered whether this could explain the variation in human-recognizability with depth.

As described in section 3.1, once we ran the optimization procedure for 15,000 iterations, the following two conditions had to be true for a model metamer to be included in our experiments: (1) The network predicted the same label for the synthetic metamer and the paired natural image. This is the same classification test we apply to humans and other networks. (2) The spearman ρ fell outside of the null distribution of activations measured between two randomly chosen image pairs (Supplement Figs 1-8). We will clarify the motivation for this approach in the revision. We believe that the comparison to a null distribution from random inputs is better than applying a strict threshold, because the expected correlation varies with network layer. Setting hard cutoffs could potentially call samples metameric which are no more matched than chance. Empirically, we found this approach crucial for random networks (Supplement Fig 3).

Regarding the potential confound described by R4: the quality of optimization is in general not worse in deeper layers. Indeed, the final layer of logits is among the easiest to match via optimization (all with median spearman ρ above 0.99, Supplement Tables 3 and 4) and all tested models have model metamers generated from the logits that are unrecognizable. Further, for metamers generated at other layers of the network, the corresponding logits are highly correlated with those of the original input (median spearman ρ>0.99), i.e., the model is nearly equally confident in its prediction for the natural and the synthetic stimulus. We will include these median logit correlations as another column in Supplemental Tables 3-4 in the final manuscript. We agree that additional optimization improvements are of interest for future work (the linear gradient ReLUs used here contribute to this effort). However, because we required the model to predict the original class label for each metamer, the substantial lack of recognition by humans point to significant discrepancies between the model and human representations even if the metamer representations are not "exactly" the same. We will rephase lines 39-43 to capture this important point. We also note that similar past work [18, 41] did not quantify the degree to which the optimization succeeded, and we believe our metrics are a step forward.

We need a more nuanced take on model success and failure (R2) We largely agree. We will rework the introduction and discussion to emphasize that what we take as model failures may not apply to all lines of work. But it seems likely that model metamers could guide discovery of neural networks that more closely resemble human perception.

Failure of the metamer test could reflect training on a single task (R2) We agree models will no doubt be limited by training on a single task. However, we contend that a model intended to replicate the basis of human performance of a particular task, such as speech recognition, should have metamers that are at a minimum recognizable to humans, even if they do not sound/look exactly the same as the original (potentially due to the single-task training of the model). We examined the effect of training task in Figure 4, and argue that more human-relevant tasks can improve models, but we will clarify this issue in the final paper.

Transfer of metamers across random initializations (R2) For the audio-trained networks, metamers generally transferred across initializations. ImageNet random seed results were not included in our manuscript because we used publicly available checkpoints that only had one training run. We have since begun training ImageNet architectures with two different random seeds to ask precisely this question, and will include those data in Figure 5 of the final paper.

The relation to Jacobsen et al. [41] is unclear (R2) [41] focuses on how "excessive invariance" (model metamers that are not classified the same by humans) relate to adversarial examples, and propose a modification to the cross entropy loss to reduce invariance. The general conclusions are similar to ours, but they exclusively study the final classification layer, and in qualitative terms. We instead examine all layers and explicitly perform human and network-network experiments, investigating how task and architecture can shape the space of network invariance. Further, [41] is focused on reducing adversarial vulnerability, while our motivation is to introduce metamers as a generic model comparison tool. We will include another sentence in the discussion explicitly addressing the relationship to [41].

The relation to visual crowding literature is unclear (R4) We will add a related work section further elaborating on this literature in the final paper. As R4 notes, this literature uses pooling in the periphery of visual models to test how well their features align with a particular aspect of human perception. We see our work as a more general instantiation of this approach, applicable to domains outside of peripheral vision where invariances arise in the service of recognition, rather than as a consequence of pooling. Moreover, our work introduces metamers for artificial model comparison, which is highly relevant to the ML community.

Work on metamers that are especially different from the original (R4) We agree that this approach could be useful. However, our paper shows that even without adversarial constraints, model metamers are often unrecognizable to humans. We believe this is a more generic model failure than those obtained from stimuli that are explicitly distinct.