Thank you to the reviewers for their comments, we are happy that all reviewers agreed that the Fourier analysis we introduce provides novel and useful insights towards understanding model robustness. We will focus this response on addressing concerns raised by Reviewer 2.

R2: "Are malicious adversaries considered?"

As stated in our abstract and introduction, our work is centered on the problem of distribution shift and not specifically adversarial robustness. We are primarily interested in understanding how a low (or high) frequency bias of image models affects robustness to low (or high) frequency corruptions. We consider the primary contribution of our work to be the Fourier analysis that we introduce, which we believe will prove useful for future research on robustness.

R2: "Section 4.2 needs more clarity. It's not clear if the Fog noise model is a good model for fog corruption..."

The purpose of this section was to demonstrate an asymmetry that occurs when augmenting in the low vs high frequency domain. Figures [3,4,5,6] all demonstrate that Gaussian data augmentation biases the model towards low frequency statistics in the image. This results generally in improved robustness to high frequency corruptions, while reducing robustness to low frequency corruptions. A natural question is what happens if we augment with low frequency noise, do we then bias the model towards high frequency statistics and improve robustness to low frequency corruptions? This appears not to hold, as demonstrated by the fog noise experiments. We will rework this section to be more clear as to the motivation for the considered experiment.

Note the goal is not to specifically design a data augmentation that is a good model for fog. Our problem setting is distribution shift, here fog is used as an example of a domain shift that is unknown at training time. Performing well on fog is trivial if the model is trained specifically on fog (e.g. see Figure 4 in another related work *MNIST-C: A Robustness Benchmark for Computer Vision*).

R2: "... a strategy for choosing this diverse set may be defined in the Fourier space but this is never investigated"

The experiments in Section 4.2 are specifically exploring data augmentation in the Fourier space. However, because augmenting on low frequency Fourier noise does not transfer to other low frequency corruptions, we decided to investigate more sophisticated augmentation strategies (e.g. AutoAugment). Note as well that Gaussian data augmentation is mathematically equivalent to adding all Fourier basis vectors with i.i.d Gaussian coefficients.

R2: "The motivation for Section 4.3 to use a more varied set of augmentations – this is reasonable but not informed (non-trivially) by the study."

A large motivation for exploring AutoAugment was specifically the experiments in 4.2. Furthermore, the fact that we discovered that the method achieves SOTA on Imagenet-C is certainly relevant to our work and worth including. Additionally, we applied our Fourier analysis to this model and discovered that while it appears to improve robustness on the Imagenet-C corruptions, there exist Fourier frequencies for which performance is degraded relative to the naturally trained model (see the high frequencies in Figure 4). This demonstrates that the Fourier analysis we introduce can identify blind spots in models for which the Imagenet-C benchmark misses. We believe this discovery can lead to additional corruptions to add to this benchmark in order to achieve a more complete perspective on model robustness.

R2: "It's not clear if generic statements can be drawn about adversarial training based on a single adversarial attack model."

We believe there was some confusion over what we were arguing in section 4.4. Reviewer 2 is absolutely correct that one should be careful about drawing conclusions about adversarial examples by analyzing a single attack, this in fact is exactly what we were discussing in paragraph 3. As discussed, the statistics of adversarial examples will ultimately depend on how the adversary generates the perturbation. We will update this section to make it more clear that what Figure 7 is demonstrating is that the statistics of the model gradients (as measured by applying a few steps of PGD) confirm our overall hypothesis about how adversarial training biases the model towards lower frequency statistics in the image. Note, the experiments in Figures [3,5,6] also confirm this hypothesis. Taken together, the experiments in the whole paper lend support to the statement that adversarial training biases the model towards low frequency statistics in the data. We will be more clear that it would be incorrect to conclude that all possible adversarial perturbations must have these properties. In fact, Figure 3 demonstrates that errors exist in the worst-case at all frequencies.

R3: "Asking for more experiments than just CIFAR-10"

Please note that the AutoAugment results in Table 2 are on Imagenet-C. Additionally, the Fourier analysis in Figure 4 is for models trained on Imagenet. We are also updating the paper to include new experiments on Imagenet akin to Figure 1, where the models are trained and tested with extreme high and low pass filtering applied. What we find on Imagenet is interesting, models can achieve well over 50% accuracy using only high frequency features which are typically invisible to the human eye. We also demonstrate a rescaling technique that can be used to visualize these features.