We thank the reviewers for the time and effort spent assessing our work. We are grateful for the positive feedback and the suggestions which are helpful to further improve our paper. Please find our responses to requests below.

[Reviewer 1] **Presentation clarity can be improved:** Thanks for the detailed suggestions. These will be incorporated. **Not radically novel in the methodological contributions:** We do believe our method is quite novel and makes a strong methodological contribution not found in previous work. Its simplicity is a key strength which will hopefully help with wider adoption. **Discussion of and comparison with multimodal image-to-image translation:** Thank you for the suggestion. We now discuss these techniques in the related work section. **Interpretation of sample diversity metric:** Here, sample diversity is not a metric of quality but an indicator of how different samples are from each other. To measure how closely we match the expert distribution, we present the generalised energy distance. We will clarify this. **Artefacts at the stitched patch borders:** During training, the model is not capturing inter-patch correlations. However, during inference, the distribution is built over the whole image before sampling, preventing artefacts from appearing between patches. If we sample and then stitch (as tried in initial experiments), artefacts would appear.

[Reviewer 3] **Mode-collapse:** This can only occur if the covariance becomes zero/negligible, in which case the samples revert to the mean. This can happen even for similar pixels (see toy example). We believe this is a feature, not a bug. The integral of eq 1 and 6 penalises predicting a single mode unless it is 100% accurate, in which case there is no uncertainty. **Intuition about the mechanism of calculating the covariance matrix:** The inner product is not taken on the output pixels but on the covariance factor coming from its separate set of conv-filters. Thus the covariance does not amount to the cosine similarity between output pixels. It would be the cosine similarity between features of the covariance factor if the inner-product was normalised and no diagonal component was added. **Covariance being spatial and not between classes for each logit:** Actually, the covariance is spatial *and* between classes for each logit. We will add some clarification on this. **LIDC evaluation:** The baselines were retrained with new random splits. During development, we contacted the authors of [9], but they were unable to provide their splits. In our experiments, we found [8] and [9] to perform almost equal to the experiments reported in [9]. However, note that in our paper, we report DSC in a different (arguably more correct) manner, see Appendix A.2. We will make this important difference clear in the text. **Literature concurrent to reference 9:** Thank you, we have missed this work and will include it in the updated discussion of related work. **Statement regarding prior work requiring a full forward pass for each task doesn't quite hold:** Thanks for pointing this out. We agree and will change the statement accordingly. **Figures could be improved:** Thanks for the helpful comments on how to improve the figures. **Figure 2 (right), expect a block-diagonal covariance matrix:** The matrix is block-diagonal. However, the first and last blocks have effectively zero variance (label never changes) and hence are not visible. **Missing training details for LIDC:** Thanks for pointing this out, we will include these details in the appendix as suggested. **Mean of the logit map for prediction versus averaging samples:** We kept the baseline experiments as close as possible to the reported state-of-the-art in [9], which used the sample average. For the baselines, the expected value of the output needs to be computed using a sample average due to the neural network in the middle. In our method, the mean of the distribution already represents the expected value of the logit map.

[Reviewer 4] **Training with a higher rank and reducing rank post-training:** Thanks, this is a very interesting suggestion that we hadn't considered yet. **More elaborate distributions or an implicit probabilistic model to improve the predictive performance:** While we did not compare with a mixture, what we have shown is that a simple distribution can improve over the implicit probabilistic models (baselines) while having lower complexity. Nevertheless, comparing with a mixture is a good suggestion for further work.

[Reviewer 5] **Better motivation for the specific weak independence assumption:** Thanks for pointing this out, we will clarify the motivation behind our choices. The multivariate normal is the simplest distribution that can model correlations between pixels. The low-rank parameterisation is motivated by computational constraints and as a way of controlling the expressiveness of the distribution (see point below).

[R3+R4+R5] **Influence of the rank on predictive performance:** We agree that studying the effect of varying the rank would be insightful. Therefore, we are preparing an ablation study to be added to the appendix showing how performance metrics (including sample diversity) vary with this parameter. Intuitively, the rank controls the number of independent clusters of pixels that are controlled together.

[R4+R5] **Computational overhead of SSNs and impact of rank on training and inference time:** The computational overhead is minimal. The overall cost is dominated by the forward pass of the underlying network. The overhead is (1) predicting three maps instead of one at very the end of the network (2) Sampling from the low-rank normal distribution to compute the loss. The cost of sampling is linear with the rank ($\mathcal{O}(rank)$). We will add this to the updated paper.

[R3+R4] **Other application domains:** Medical imaging is among the most critical applications for uncertainty estimation. We hope the methods and results are relevant for the wider NeurIPS community but agree that in future work, other domains could be explored.