We thank all reviewers for constructive and valuable feedback. It will be used to improve the presentation. Reviewers note that the paper to unifies discrete autoregressive models and flows (R1,R2,R4,R5). This interpretation allow us to study the dequantization gap, which all reviewers note as a significant contribution (R1,R2,R4,R5). Reviewers also note that the paper is well-written and clear (R1,R2,R5). We address individual concerns below.

Reviewer 1:

- *Relabeling existing work:* We would like to emphasize that subset flows are mostly introduced as a vehicle to explicitly connect flows and discrete autoregressive models. After expressing PixelCNN models as subset flows, we are able to treat these models as building blocks in flows (which is experimented with in Sec. 6.3). Moreover, it allows us to 1) perform latent space interpolations in PixelCNN models, which was not previously possible and 2) explicitly quantify the effects of dequantization during training and evaluation of flow models.
- *Quadratic spline extension:* The reviewer is right that the mixture of logistics (MOL) transformation may, for certain data, have better inductive bias than quadratic splines. One benefit of quadratic splines compared to MOL is that it has an analytic inverse, rather than requiring the iterative bisection method.
- *Multilayer subset flows:* We again emphasize that subset flows are mostly introduced to make the explicit connection between flows and discrete autoregressive models. As you note, extending subset flows to multiple layers is not straightforward. We study a restricted version of multilayer subset flows in Appendix F.
- *Experiments:* The interpolation experiments are meant to demonstrate the fact that, when expressed as flows, PixelCNN models possess a latent space. As latent space interpolation experiments have commonly been performed for VAE-type and flow-type of models, we here demonstrate the possibility of doing this for PixelCNN-type models, which was previously not possible. The experiments combining PixelCNN models and coupling flows are meant to demonstrate the use of PixelCNN models as building blocks in flows. By replacing the uniform base distribution with more flow layers, a more expressive model is obtained.

Reviewer 2:

- *Computational cost:* The reviewer is correct. The training of PixelCNN models is costly. In our experiments (depending on the type of GPU used) we trained a stock PixelCNN in about 30 hours and a stock PixelCNN++ in about 10 days. We will include a comment on this in the paper.
- *Performance of PixelFlow:* You are correct that both PixelCNN++ and PixelFlow++ obtains 2.92 bpd. However, PixelCNN++ obtains $= 2.924$, while PixelFlow++ obtains ≤ 2.917. PixelFlow++ thus performs better than PixelCNN++, but due to the dequantization gap and rounding, the stated number is 2.92 in both cases.
- *Rotation:* Rotations refer to rotating the feature maps 90°.
- *Tiling:* The tiling of $\mathcal{Y}$ automatically leads to a tiling of $\mathcal{Z}$ since they are related using a bijective map.

Reviewer 4:

- *Mischaracterization of discrete flows:* In the statement "Flow models...are naturally continuous and therefore require dequantization to be applied to discrete data" we are implicitly referring to continuous flows (because the vast majority of work on flows is continuous, since, as you correctly note, discrete flows are severely limited in their expressiveness). We agree that the statement is not true for discrete flows (as we explicitly addressed in the related work). We will clarify that we are indeed referring to continuous flows in the statement.
- *Subset Flows as RAD Flow Mixtures:* We respectfully disagree that subset flows are better expressed as flow mixtures. Flow mixtures involve a finite/infinite collection of flows where during sampling, the component to use is selected before performing the transformation. RAD, for example, tiles the space $\mathcal{Y}$ and maps to all of $\mathcal{Z}$ (not a subset of it). The transformation is therefore not invertible and during generation requires sampling of an index that decides which region to map to. Subset flows, on the other hand, are themselves plain flows. Subset flows tile both spaces and thus preserve invertibility. However, we are happy to include references to these papers to discuss how this work relates to ours.

Reviewer 5:

- *Follow-up on insights:* This is a good point. We would suggest that there is no "discrete autoregressive models vs. flows", but rather "how well can you reduce the dequantization gap for your flow?". Consequently, this suggests development of methods for reducing the dequantization gap as an avenue for further research.
- *Specific to PixelCNNs:* We would not expect that the magnitude of the dequantization gaps are specific to PixelCNN-based flows, although it is difficult to quantify the exact gap for other flow models.

All: We will use all comments and questions to improve the presentation. Thanks again for your work.