We thank all the reviewers for their feedback. We are encouraged by the praise received for both our motivation for quantized hierarchies (R2,R4) and empirical results (R1,R2,R3,R4). We are also pleased to see a recognition of significant impact for the generative compression sub-community (R4). All reviewers noted the effectiveness of HQA in the extreme low bitrate regime, without needing autoregressive components - this was a key contribution we wanted to emphasise. We will incorporate all feedback in a future revision but address key themes below.

Praise & Criticism Of Section 4 (Lossy Compression Using Quantized Hierarchies) A key contribution of our work is motivating why, under an information bottleneck, *quantized hierarchies with stochastic quantization* show a fundamentally different modelling behaviour from other likelihood-based systems. We believe we are the first to present this form of motivation. We were glad to see R4 recognise this section as being a "very good explanation" of how quantized hierarchies remedy common pathologies of likelihood-based image compression and R2 notes "a fundamental improvement over the VQ-VAE approach". However, as R2 comments, the "structure and writing of this section is confusing, especially Section 4.3". Noted - this section is important but the reviewers have convinced us that the exposition needs a significant rewrite. To summarize, we show how quantized hierarchies exchange mode-covering behaviour in the input space for mode-covering behaviour in the latent space. Stochastic quantization then allows for high quality reconstructions without sacrificing diversity.

R1 gives several excellent suggestions for how the clarity of this section can be improved, which we think will lead to a much stronger paper. R1: "I don't get how it's fair to compare the toy 2-layer HQA with a simple VQ-VAE with 2-code latents" - when comparing two compression schemes it is only the number of codes in the top layer that is controlled for as this ultimately determines the model's compression rate.

Experiments We are glad to see that all authors found our empirical results impressive. However, R1 and R2 would have found our experimental section more convincing if we included results on higher resolution images, such as FFHQ and ImageNet. We completely agree - the criticism is entirely valid. In particular, we thank R2 for their helpful analysis of the relationship between FID and image resolution. We are saddened to say that we simply don't have the resources to pursue these higher resolutions. VQ-VAE, VQ-VAE2 and HAMs were all developed by the same research lab that has access to several orders of magnitude more compute than us. We stress that it is common for well received work in generative modelling and generative compression to only quote on 64x64 images [1, 2, 3]. However, we agree that higher resolution evaluations would have significantly improved the paper - we will provide 128x128 reconstructions and make a best effort to obtain the compute budget needed to train a 256x256 system.

The other points raised regarding our experimental section are more easily addressed. As R2 notes, our experiments were designed around the generative compression setting, as opposed to generative modelling - we stand by this evaluation protocol and leave sampling for future work. R2 also notes that the two faces we show in Table 2 are not quite enough to convince them of a strong qualitative argument; we have plenty more comparisons which we will add into the appendix to solidify our qualitative argument. R4 notes a comparison to [3] would have been useful - noted - we actually did produce an internal comparison. We found the visual quality of HQA to be preferable but the FID scores of [3] to be superior. We believe this is in part due to [3] being adversarially trained. We made the decision not to include these results as we wanted to present a fair investigation comparing *likelihood-based* approaches, as we discuss briefly in Section 6.1. We will include these results in future supplementary material.

After including the above results, we believe our work will be much stronger and will sufficiently address the scepticism around our experimental analysis.

Novelty Clarifications R1 and R3 expressed concerns about the novelty of our work. This may be linked to the misgivings R1 and R3 have regarding Section 4. The analysis of quantized hierarchies in Section 4 is a key contribution of our work. As we mention above, we strongly believe that an improved exposition in Section 4 will lead to both R1 and R3 deeming our paper more novel. We also wish to highlight the strength of the HQA scheme lies in our novel prediction target, namely reconstructing z_e using MSE and not z_q using cross-entropy (as done in HAMs). This small change has **dramatic** consequences, as shown by both Table 2 and Row 5 in Table 3. We believe this novel prediction target, and scheme taken as a whole, have large implications for those working on quantized hierarchies and set our work apart from others. (We would also wish to highlight to R3 that the novel prediction target is separate from the probabilistic commitment and codebook losses, which may address some of their confusion. We will clarify this in a future revision). We also appreciate R1's feedback on the introduction of our contributions and will rewrite according to their suggestions.

[1] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.

[2] A. Genevay, G. Peyre, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences.

[3] M. Tschannen, E. Agustsson, and M. Lucic. Deep generative models for distribution-preserving lossy compression.