We thank all reviewers for their thorough reviews. Given their already positive comments, we hope our responses below will help increase the reviewers' confidence further and resolve any remaining doubts.

We'd like to remind reviewers that our proposals significantly improved the baseline and advanced the state-of-the-art on the extremely competitive task of image compression. Building a successful data compression method on the progress in generative modeling is nontrivial, and knowledge transfer between the two fields has only begun recently. Both our annealed optimization method for integer representations and lossy bitsback are significant novel ideas in this direction.

R1: *"Can SGA be used at training time as well?"* $\rightarrow$ It can in principle, and should increase performance further (see concurrent work [arXiv:2006.09952]) but a naive implementation would slow down training considerably. Mitigating this by genrealizing ideas from [Kim et al., 2018] or [Marino et al., 2018] to SGA would be interesting followup work.

R1: *[timing comparison]* $\rightarrow$ Yes, that's a good point. We will provide detailed results in the final version of our paper. Unfortunately, the rebuttal period was too short this year to generate a full analysis in time. In preliminary results, we see a slowdown for *encoding* (i.e., compressing) of about 100x in our non-optimized code, which is similar to what has been reported in [Campos et al., 2019]. Please note that our proposed standalone variant [M1] changes only compression and does not affect *de*compression speed (which is more relevant, e.g., for images on a website with many visitors).

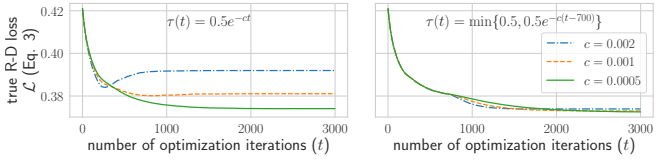
R2 & R4: *[limited novelty]* $\rightarrow$ We respectfully disagree. Our paper proposes two significant novel inventions: (i) a novel inference method over an *infinite discrete* set, which significantly improves compression performance, and (ii) the first lossy bitsback coding algorithm. We believe that each of these two would already be a significant contribution on its own. However, we decided to combine both contributions into a single paper since, empirically, they strongly complement each other (compare model [M1] to ablation [A3] in Section 4).

R2 & R3: *"improvements from bitsback coding are relatively marginal"* $\rightarrow$ We would like to clarify that these improvements ([M2] in Section 4) are *on top of* an already novel method [M1] proposed in our paper, which already improves performance significantly over the previous state of the art on this very competitive lossy image compression benchmark. Further, we would like to point out that generalizing bitsback coding to lossy compression is nontrivial (this has also recently been confirmed to us in private conversations with leading industry researchers working on this topic). To the best of our knowledge, our work is the first empirically successful lossy variant of the 30-year-old bitsback algorithm. We expect it to enable research on much more powerful hierarchical prior models for neural compression.

R2 & R4: *"implement the proposed method based on [M4] (context+hyperprior)"*, *"why the latent structure is restricted to a 2-layer structure"* $\rightarrow$ Our paper focuses on *inference* rather than model architectures. While the proposed inference methods can also be applied to other models, [M4] faces computational difficulties due to its inherently serial nature (and is thus also excluded by [Johnston et. al, 2019]), and models from the broader VAE literature are often not good for compression (e.g., compression models usually need much larger latent spaces). We deliberately used a model that is common in the neural compression literature so that we could study the effect of improving inference in isolation from improving the model architecture. We find such separation of concerns essential for generating scientific insights.

R2: *[comparison to arXiv:2003.11282]* $\rightarrow$ Thank you for the bringing this work to our attention, we will cite it in our paper. The idea of optimizing the encoder parameters indeed seems related to our approach. By contrast, our approach directly optimizes the output of the encoder, and is thus not limited by the expressivity of an encoder architecture

R3: *[analysis of temperature annealing]* $\rightarrow$ Good point! We will add the below curves to Figure 2 of the paper. The left plot shows the true R-D objective of SGA using a naive temperature schedule $\tau(t) = 0.5e^{-ct}$, for various decay factors c. As can be seen, too fast annealing with this naive schedule can lead to suboptimal solutions. The right plot shows that we can overcome the suboptimality from fast annealing by fixing the temperature to τ_0 for some initial steps (until the R-D objective roughly converges) before annealing; we used $\tau_0 = 0.5$ as it approximates soft quantization As shown, our resulting method is robust to different choices of the annealing factor c.



R3: *"How the entropy coding is implemented [...] how the side information is designed?"* $\rightarrow$ Like the original (lossless) bitsback algorithm, the proposed lossy bitsback algorithm builds *on top of* entropy coding and is agnostic to both the specific entropy coder used and the origin of the side information. We will provide a simple ANS entropy coder in our public code repository. Our results for the proposed bitsback method [M2] report expected net bitrate for a *random bitstring* of side information (this is a worst-case scenario since a random bitstring cannot be further compressed).

R3: *[SGA & bitsback in a non-hierarchical VAE]* $\rightarrow$ Thank you for pointing this out. Indeed, SGA does not strictly require a hierarchical VAE, but hierarchical VAEs have proved to lead to superior compression performance in the literature. The proposed lossy bitsback coding algorithm also builds on a hierarchical model; exploiting the increased expressivity of a hierarchical model without paying the price of the marginalization gap is precisely its strength.