We thank all reviewers for their constructive and helpful comments.

Reviewer 1: We will be sure to provide a more accurate and nuanced discussion of the downsides of our auxiliary bits requirements in a revision. The particular example we had in mind for the sentence on video was the case of compressing an hour-long video: at 30 frames per second, this is 100 thousand frames, after which we expect the auxiliary bits to be negligible. The auxiliary bits are of course not negligible for shorter videos, and we will change the text to plainly describe the downside of large auxiliary bits requirements.

Reviewer 1: Regarding runtime evaluation, what we called the "wall clock time" is the sum of the GPU time and the CPU time, and the reported time to "run the neural net on its own" is the GPU time. We will correct and clarify our terminology regarding timing here.

Reviewer 1: Regarding hardware and software differences for encoder and decoder: indeed it is crucial for all computations, especially those in the flow model, to be exactly reproducible on both sides of the communication channel, otherwise we indeed do run into catastrophic error propagation. In our implementation, we set flags to force usage of deterministic CUDA kernels, and we use the same hardware for both encoding and decoding. As with other methods that rely on exactly reproducible probability models, these issues can be addressed with careful engineering.

Reviewers 2 and 3: Comparing the most modern instantiation of bits-back coding with hierarchical VAEs (Bit-Swap), our algorithm and models have better net bitrates, at the expense of a large number of auxiliary bits. Specifically, on 32x32 Imagenet, we attain a net codelength of 3.88 bits/dim at the expense of approximately 40 bits/dim of auxiliary bits (depending on hyperparmeter settings). In contrast, the models in the Bit-Swap paper attain a net codelength of 4.48 bits/dim, with only approximately 2.5 bits/dim of auxiliary bits. Indeed, as Reviewer 1 mentioned, our auxiliary bits requirement is a downside of our method for short sequences, which do not have enough timesteps to amortize out this requirement. We will revise our paper to include this discussion.

Reviewer 3: One of the motivations of our work is theoretical interest: likelihood-based models are known to optimize lossless compression rate, but it is not always clear for any one given likelihood-based model how to achieve this rate in practice. We have filled in this gap in the literature for flow models. Another motivation is practical: when running our algorithm on RealNVP-type models, encoding/decoding passes are fast and parallelizable, unlike arithmetic coding or ANS for autoregressive models, which are slow for decoding. We also attain good net codelengths, which are currently better than those in the VAE compression literature (though at the expense of a worse auxiliary bits requirement, as just discussed).

Reviewer 3: Regarding clarification of Section 3.5: coding integer-valued data at high precision is a waste of bits, because it is not necessary to specify the data at a resolution smaller than that of a unit hypercube. In the one-dimensional case, doing so is akin to storing integers with a nonzero amount of precision after the decimal point – this is a waste of bits, because those digits after the decimal point will always be zero. The image datasets we use consist of integer data, so we wish to avoid coding bins of volume less than 1.