General response ($R1$, $R2$, $R3$)

Dear Reviewers, we thank you for taking the time to provide valuable feedback. We will correct the final manuscript to fix issues related to typos and missing citations. We strongly believe in the significance of our work as a first deep probabilistic modeling approach towards end-to-end video compression. Below we address the main issues raised.

Simplicity of datasets and scalability. First and foremost, we address comments ($R1$, $R2$, $R3$) regarding the simplicity of the datasets, being short, low-res videos. Our approach to performing video compression by entropy coding the video according to dynamic probabilities from a deep generative model (i.e. not block motion based video compression) is novel. Its performance depends on our ability to predict the distribution over future frames with low entropy. Currently all papers on deep video generation (upon which our approach is based) consider data sets comparable to ours and face scalability difficulties. Scalability limitations of the considered architecture can be can be partially ameliorated by using convolutional architectures instead of fully connected ones. We stress that our contribution is not only a specific sequential VAE architecture, but the full concept of using a temporally-conditioned, learned prior for entropy coding sequences, and gives rise to various extensions (multi-modal sequential priors, implicit distributions, hierarchical latent sequence models with multi-scale dynamics, etc.). We will emphasize these aspects more in a revised version.

Relationship to existing image compression work. While all reviewers agree that our approach extends existing work, we here clarify the similarities and differences to to Balle's approach to image compression, addressing in particular $R1$'s concern. While giving full credits to this related work for adopting its idea of discretization and entropy coding, we stress that our video compression model is quite different from Balle‘s image compression model, and that the sequential VAE setup that we analyzed posed new technical challenges that we overcame. In contrast to a stationary learned prior, our approach builds on sequential VAE architectures for high dimensional time series and in employs RNNs to model dynamics in the latent space. Further contributions include separating static from dynamic information with local/global variables as well as entropy-coding the sequence according to the non-stationary, learned prior. We compare with a basic temporal predictive model using Kalman filtering to show the advantages of such tailored design.

$R1$ **Response:**

>...how the model could be scaled to longer videos and reported the results of this preliminary investigation...

Note that the video length is not an issue if one weakens the assumption of a global state besides the local one. We have recently found some efficient architectures that can be incorporated into our encoder/decoder part to improve our model. We are currently working on this and hope to include our results in the final revised version. Research progress towards better deep generative models for high-resolution, diverse video is required to push this idea further to high resolution. In particular, a generative model that outperforms next-frame block motion prediction on high-res video is needed.

$R2$ **Response:**

>The use of a global encoding of the entire sequence might limit applicability of the approach...

Live videos can be divided into chunks of T frames, where every chunk can be encoded separately with independent global states. The only limitation is that the video has to be encoded in chunks of T frames, and during decoding a new reference global state must be used every T-frames. This is not too different from encoding/decoding chunks of video between key frames in classical codecs. The key frame stores information shared among the frames and is used as a reference to decode the subsequent frames. One advantage of the global encoding is that as a reference it has access to all of the frames in the chunk to better store shared information.

Note that the global state is not strictly required—our approach would also work with a purely local state. Our paper actually contains experiments with such architectures (see Fig. 3 LSTMP-L) as well, and is formulated more broadly.

>The relation of Eq 3 wih Eq 4 is not obvious...

The sequential prior is not specified in Eq. 4 as opposed to Eq. 3 (but is meant to be the same). We will fix this.

> (Uncited) recent work...

Thanks for pointing this out. We will cite and discuss it in our final version. We stress that this approach is different.

$R3$ **Response:**

>... whether a deep probabilistic model for video compression is in fact better than a deterministic neural-network ...

At inference time, our method is deterministic as required by entropy coding. Our method is probabilistic in the sense that the probability distribution over all the next possible frames is used to remove temporal redundancy, and can be successively improved by better video prediction. Classical video codecs (or related neural network approaches) remove temporal redundancy by subtracting the most-likely estimate for the next frame from the actual next frame.