Thank you for the insightful comments. We will make sure to address all requests for elucidation in the revision.

R1 and R2: Single image evaluation: Our main motivation is that of single sample video generation. However, as discussed in Sec. 4.2, our method also applies to images. We follow a similar evaluation protocol to SinGAN: quality (SIFID), diversity (Fig. 5 RHS) and a user study (Tab. 3 RHS). Using $M = 3$ (see L.245-252), we outperform with all these metrics. **Additional applications:** In Fig. (a-c) below, we consider the applications of inpainting, harmonization and editing, as used in SinGAN. We also consider the following application: produce random videos of higher 1024×256px resolution, compared to the 256×192px resolution of training video. Two consecutive frames are shown for two videos (d and e). Full videos will be provided. **Low resolution, FPS and length:** We produce videos of the same FPS as the training video (24), and of a standard 256px resolution. For visualization purposes, GIFs produced in the SM are played at a slightly lower FPS. By using an input of higher FPS, videos of higher FPS can be produced. Our method produces 13 frames as opposed to baselines' 16, but of higher resolution. This is a result of GPU memory limit and sampling technique, and not a limitation of the method. See Sec. 3.2 and SM Sec. 2.1 for additional details.

R1: Many differences to SinGAN: We introduce a novel formulation of a VAE that operates on patches of a single sample. Previously, only patch-GAN existed [25, 24, 26, 27]. While attempts were made to combine GANs and VAEs [17, 18, 19,20], we are the first to do so within an hierarchical structure. We believe that both of these formulations could be used in other generation tasks. Moreover, we consider the challenging task of diverse video generation from a single sample. SinGAN completely fails on this task, and collapses, even when modified accordingly. **Study of r:** Following the review, we conducted an additional experiment with $r = 31$. As noted in L.120-121, setting r too large, results in a small number of samples K. In our experiment this resulted in failure to generate realistic samples. **Fig 6, row 3:** Fig 5 shows that when using the unoptimal setting of $M = 9$ (only VAE), SVFID score is high, indicating low quality, yet diversity is high. In Fig. 6, row 3, the sky is scattered completely differently from row 1. The appearance of only a few balloons is specific to this randomly generated sample. In the SM, for example, many balloons appear.

R2: Realism: While video quality can be improved, our method significantly outperforms baselines (see Fig. 5 and Tabs. 1-3). **Complex motion:** Note first SM example in 'Longer Training Videos' depicting hot-air balloons moving in different directions and first SM example in 'Randomly Generated Videos' of skydivers depicting significant camera motion. Following the review, we considered an input video of a savanna, with cheetahs and an impala moving at different directions, and with a moving camera. Fig. (f) shows a randomly generated output. Random videos capture the same camera motion of the input video, while generating realistic objects (cheetahs and impalas) moving in novel directions. **L.122-126:** Applying SinGAN on multiple samples would require reconstructing each sample, at the the coarsest level, from the same fixed noize z^*, resulting in bad reconstructions. As our method uses a VAE formulation, it does not have this limitation. **Prior Work:** Random samples were generated using the public implementation of baseline methods. We note that MoCoGAN [31] did not consider the UCF101 dataset and hence resulted in the worst generation quality. Regarding variability, we do not claim to generate variable content from multiple videos, as may be the case for multiple-sample baselines. However, our method is superior to baselines in generating diverse content from the same internal statistics of the input video (see Tab. 3 and Fig. 5). We will clarify this in the next revision. **Eq. 10:** We thank the reviewer for noting this typo, which we will correct. A WGAN-GP loss with gradient penalty is used. **User study videos** are shown for unlimited time. **Latent representation dimensionality**: As discussed in L.100-104, the encoder E can be seen as creating a distribution $q(z|x)$ of of r-sized patches. Our default setting uses $r = 11$, and $c = 128$, and so the dimensionality of a patch of size 3x11x11x11 in the input video is reduced to 128. **Change of G_0**: Unlike $G_1, \dots, G_n$, G_0 has a special role of decoding a video from noize. G_0 adapts to the distribution $q(z|x)$ of the latent encodings (which may change when moving to a new level of the hierarchy) produced by the encoder. The distribution of lower resolution inputs to $G_1, \dots, G_n$ is fixed, and so $G_1, \dots, G_n$ do not continue training (L.157-159).

R3 and R4: Comparison: We will gladly add a discussion of [1] and of [a], both of which operate in the multiple sample setting. We did our best to compare with methods for which code was made public. For [1, 32] this is not the case. Regarding [31], we thank R3 for pointing to the implementation. We will make an effort to add this comparison.

R3: Encoder input: As our encoder E and decoder G_0 are fully convolutional, G_0's output is of the same size as E's input, which must match x_0's size. Instead, to enrich the latent space using x_n's finer details, E is still updated through x_n's reconstruction loss. **Multiple sample baseline:** We verified that for the subset of videos where the 2nd NN is from a completely different video, SVFID is indeed somewhat higher, but still superior by a sizable gap to baselines.

R4: Limitations: Our method is trained in an unsupervised manner on a single input video. As a result, it has no semantic understanding or notion of "scenes". While all the local elements are preserved (people walking, car moving), the global structure may be unnatural: For example, in the SM, example 13 of airplanes, the plane trail appears separated from the plane in the first random example. We will add a discussion of the limitations in the next revision, specifically discussing the inability to extract complex spatial relations as those in Cityscapes from only one video.