Appendix

Autoencoder model details: Our compression model had a similar architecture to [1], but we substituted the 2D convolution with 3D convolutions. Due to the high memory consumption, we removed all attention layers from the compression model. We used 64 base channels, with a channel multiplier of [1,2,2,2] and 2 residual blocks per level. Our latent space had a dimensionality of $20 \times 28 \times 20$ with 3 latent channels. We trained our model over 50 epochs with minibatch of 8, with an Adam optimizer and a base learning rate of 0.00005. We used a patch-based discriminator in our adversarial loss with 64 base channels and a learning rate of 0.0001.

Diffusion model details Our diffusion model uses the U-net architecture from [1], with 256 base channels, a channel multiplier of [1,2,3] and 2 residual blocks per level. We used a base learning rate of 2.5e10-5 and an Adam optimizer. We used a L2 loss parametrized to predict the epsilon of the diffusion process. We used a Markov chain with 1000 timesteps, with a linear variance schedule, from 0.0015 to 0.0205. We used a hybrid approach for conditioning with a spatial transformer with a depth of 1 and 1 attention heads.

References

1. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)