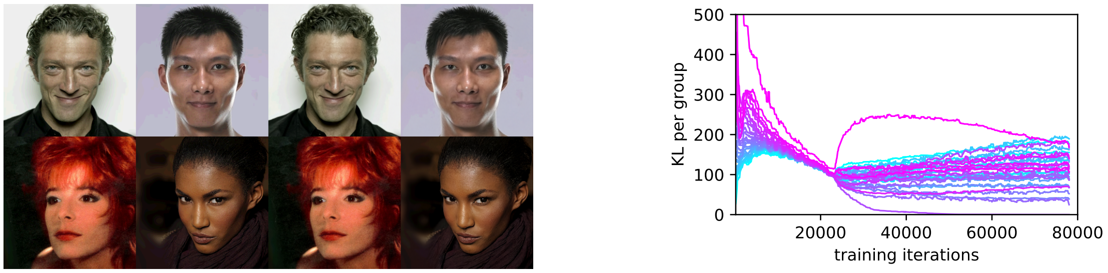

(a) Reconstruction results (best seen when zoomed in).

(b) Average KL per group.

Figure 1: (a) Input on the left and reconstructed image on the right for CelebA HQ 256. (b) KL per group on CIFAR10.

We would like to thank all the reviewers for positive and constructive feedback.

R1 ---

Reconstruction: The reconstructed images in NVAE are indistinguishable from the training images (see Fig. 1(a)).

Discussing additional SOTA hybrid models: Thanks for pointing this out. We will include this in our final version

Non-dimension-reduction and non-disentangling design: Since our goal is to maximize (a lower bound on) the marginal data log-likelihood, hierarchical dependencies help us build expressive approximate posteriors, which often result in better generative performance. However, when the training goal shifts towards disentangled representation learning, we can sacrifice generative performance for representation learning (See β-VAE, "Fixing a Broken ELBO").

Non-face datasets such as ImageNet: We haven't explored the ImageNet dataset beyond the 32x32 version. However, our hypothesis is that ImageNet may correspond to a high-entropy distribution (compared to face images) which may require even bigger VAE models. GANs are perhaps less prone to this, as they may drop modes without being penalized.

R2 ---

Is the data conditioned on all $\boldsymbol{z}$'s: Yes, because the representation at the bottom of the top-down model (Fig. 2(b)) is a function of all $\boldsymbol{z}$'s.

Parametrization for σ: We parameterize σ's in their log space, and we limit $\log \sigma$ to be in [-5, 5] which is much larger.

Training curves: Fig. 10 in the supplementary material demonstrates training stability with spectral regularization.

Posterior collapse: Since we are using more latent variables than the data dimensionality, it is natural for VAE objective to turn off many latent variables. However, our KL balancing mechanism (Sec. A in the appendix) stops the hierarchical groups from turning off. In Fig. 1(b), you can see KL per group in CIFAR10 (for 30 groups). Note how most groups obtain a similar KL on average, and only one group is turned off. We apply KL balancing mechanism only during KL warm-up (the first $\sim$ 25000 iterations). In the remaining, we are using ELBO without any KL balancing (Eq. 1).

Importance weighted (IW) AE: We do not train our model with the IW bound. The comparison to the previous work is fair in this regard as we also use ELBO for training (see Eq. 1).

Autoregressive decoder + NVAE: Autoregressive decoders 1) are often slow to sample from and 2) tend to turn off latent variables. A careful study is required in this space.

Sampling time compared to pixelCNN++: On CIFAR10 using a Titan V GPU and batch size of 16, NVAE takes 5.1 ms per image vs. 8,898.75 ms per image for pixelCNN++. NVAE is $\sim$1700x faster than pixelCNN++.

R3 ---

Reproducibility: Our current top priority is to release the code publicly with clear instructions to reproduce the results.

Test likelihood after BN adjustment: Thanks for suggesting this. Since the generative model changes after re-adjusting the BN layers, this requires encoder retraining. We'll add a careful examination of this to the final version.

R4 ---

Lack of novelty: NVAE offers expressive networks and it enables stable training of deep hierarchical VAEs for large images. We believe NVAE has the potential of becoming the backbone of future VAEs with complex statistical models.

Efficacy of depthwise conv: Depthwise conv has less computational complexity, however, it is still slow on current hardware because of memory bottleneck. We agree with you that depthwise conv seems less efficient, but in our early experiments, the generative quality of regular conv could not match it, even with more channels or longer training.

VQ-VAE: Thanks for pointing this out. We will update our description of VQ-VAE to indicate that it was originally motivated by deriving a lower bound on log-likelihood.