Dear reviewers, thank you for a thorough review of our paper. We provide a point-by-point response to each reviewer below.

Reviewer 1

1. Using TRIP as a variational distribution is an interesting direction of further research, although we will not be able to apply a reparameterization trick for a TRIP proposal in a way it is used in Gaussian proposals. We will have to use REINFORCE, which may lead to a high gradient variance and, hence, unstable learning.
2. Corrected.
3. For GAN-GMM and GAN-TRIP, we used baselines to reduce REINFORCE's gradient variance (see Eq. 10). A prior of GAN-$\mathcal{N}(0, I)$ is not trainable and hence does not require a baseline. We will add clarification about using baselines to train GAN-GMM to the paper.
4. We thank the reviewer for suggesting to use $128 * 10$ components in the GMM baseline. 1000 components stated in the paper is a typo, the actual number of components was indeed 1280, see the source code file `train_gans.py` from supplementary materials, line 103. We will fix the typo in the paper.
5. We chose the core size to balance computational complexity and empirical performance (see Table 1 below). For $m_k = 20$ the model converged after around one day of training, while for $m_k = 50$ training takes around a week, since it requires more epochs to converge.

Table 1: Time and memory consumption of operations with prior (per batch). m_k is a core size, latent space dimension $d = 100$, number of Gaussians per dimension $N = 10$, batch size $b = 128$. Other parameters are the same as used in the paper. We performed the experiments on Tesla K80.

m_k	$\mathcal{O}$-NOTATION	1	10	20	50	100
LOG-LIKELIHOOD, MS	$O(b \cdot d \cdot (m_k^3 + m_k^2 N + N))$	126 ± 7	137 ± 4	193 ± 15	200 ± 20	308 ± 12
SAMPLING, MS		201 ± 21	232 ± 13	312 ± 18	360 ± 17	882 ± 15
MEMORY, MB	$O(d \cdot (m_k^2 + N))$	0.023	0.77	3.1	19.5	78.1

The reviewer also asked to test the TRIP model for a posterior collapse. For a multimodal prior, a posterior collapse is indeed unlikely, since we cannot approximate a multimodal distribution with a single mode; the only failure mode is when prior collapses to a unimodal distribution along some axis. For our VAE-TRIP model, the number of active units (AU) was $100/100$. We will also add an experiment on MNIST and StackedMINST to a camera-ready version.

Reviewer 2

1. The reviewer suggested benchmarking the models with TRIP, GMM, and Gaussian priors with the same number of parameters. We present the result of this experiment in Table 2 below, supporting the conclusions we got from the original experiment.

Table 2: VAEs with different priors and a comparable number of parameters

	$\mathcal{N}(0,1)$	GMM	TRIP	$\mathcal{N}(0,I)$-FLOW	GMM-FLOW	TRIP-FLOW
PARAMETERS (MODEL)	11,4M	11,1M	10,7M	11.3M	10.7M	10.4M
PARAMETERS (PRIOR)	0	0,2M	0,6M	0.3M	0.5M	0.7M
PARAMETERS (TOTAL)	11,4M	11,3M	11,1M	11.5M	11.2M	11.1M
ELBO	-192.6	-190.05	-189.1	-185.3	-186.0	-184.7

Reviewer 3

1. The proposed TRIP model has many useful properties such as conditioning on a subset of attributes (Sec. 4)—a property that other priors (including flow-based models) do not have. For a fair comparison, we incorporated TRIP as an initial distribution of a flow-based RealNVP prior and show in Table 2 that such model outperforms a standard RealNVP prior. We will add a section on incorporating neural priors to the updated paper, including VAMP and IAF priors.
2. The computational costs of TRIP depend on the number of dimensions d and core size m_k (usually constant for all k). We report asymptotic complexities, time, and memory measurements in Table 1, showing that TRIP is practical for moderate core sizes.