We thank the reviewers for their suggestions. We have revised the text for clarity, added experiment details so that the paper is self-contained, and added the following comparisons requested by reviewers: 1) a comparison of SNIS, HIS, and LARS on the Continuous MNIST, Fashion MNIST, and CelebA datasets, 2) a comparison of performance as K and T are varied, and 3) experiments which use HIS and SNIS as the prior for a convolutional hierarchical VAE (ConvHVAE) as well as using the ConvHVAE as a proposal for SNIS. These experiments confirm that SNIS and HIS outperform or perform comparably to LARS while optimizing a proper lower bound and being simpler to implement. We include a subset of these results in Tables 1 and 2. We have also added samples from each model to the Appendix to allow for qualitative comparisons.

Contributions. The paper characterizes the bound gap for Monte Carlo Objectives, explicitly reveals the connection with auxiliary variable variational inference, and derives a novel class of models that balance tractability with the inductive biases of energy-based models. Furthermore, as **R1** notes, we draw links between disparate techniques and unify many existing approaches in a common framework.

Method	Static	Dynamic	Fashion
ConvHVAE	-82.43	-81.14	-226.39
ConvHVAE w/ SNIS prior	$\mathbf{-81.51}$	$\mathbf{-80.19}$	$\mathbf{-225.83}$
ConvHVAE w/ HIS prior	-81.89	-80.52	-226.15
ConvHVAE w/ LARS prior	-81.70	-80.30	-225.92
SNIS w/ ConvHVAE prop.	$\mathbf{-81.65}$	$\mathbf{-79.93}$	$\mathbf{-225.53}$
SNIS w/ VAE prop.	-87.65	-83.43	-227.63
LARS w/ VAE prop.	—	-83.63	—

Table 1: Discrete MNIST datasets.

Method	MNIST	Fashion
SNIS w/Small VAE prop.	-1254.77	-2462.60
HIS w/Small VAE prop.	$\mathbf{-1207.88}$	$\mathbf{-2449.48}$
LARS w/Small VAE prop.	-1256.83	-2464.65
SNIS w/VAE prop.	-1000.83	$\mathbf{-2244.06}$
HIS w/VAE prop.	$\mathbf{-996.32}$	-2250.19
LARS w/ VAE prop.	-1005.20	-2263.46

Table 2: Continuous datasets.

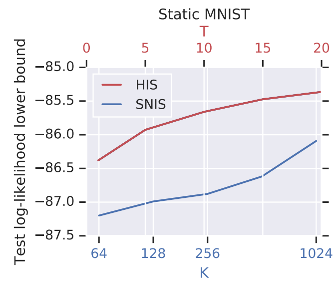

Figure 1: Varying T & K.

R1: Compare with LARS on the continuous datasets. We have added this comparison and find that LARS underperforms SNIS and HIS (Table 2 and similar results on CelebA).

R1: Varying T and K. We have added this comparison and find that increasing K or T improves performance at the cost of more computation (Fig. 1).

R1: Can existing lower bounds can be improved and a more general analysis of bound tightness? We agree that it would be interesting to experimentally examine this; however, due to space constraints, we chose to focus our experiments on the proposed models. Theoretically, we generically characterize the bound gap in terms of KL divergences (Eqs. 1 & 2). While the bound gap was known for specific cases (e.g., IWAE), we can use the general result to characterize the bound gap of VSMC, the Hamiltonian VAE, and semi-implicit VI. To first order, the bound gap is related to the variance of the partition function estimator, which motivates using lower variance estimators. We now explain this in the main text.

R2, R3: Clarity. We have rewritten the introduction to focus on the key concepts that are used later and to bridge the two halves of the paper. Section 3.2 was intended as a complex example of a method falling into our framework. We have rewritten the section to more clearly connect it back to the central story. We have reworked the experimental sections so that the setup is clear without having to read LARS.

R2: Why does HIS perform worse on synthetic data? Our implementation of HIS implicitly uses $K = 1$ and can be extended to $K > 1$ by drawing additional samples, reweighting, and sampling. Increasing K or T improves the performance of HIS on the synthetic data. To verify our claims about density mismatch, we reran the synthetic experiments where the proposal distribution has smaller variance and found that HIS outperforms the other methods. We now include these comparisons in the Appendix.

R2: HIS vs HVAE vs HIS+HVAE. Our experiments show that they provide complementary improvements. SNIS and HIS improve the performance of both the VAE and ConvHVAE when used as the prior distribution. Because the latent space is small (50-dimensional), the additional computation cost of SNIS or HIS is small.

R2: With stronger proposals, do HIS/SNIS still provide benefits? Yes, we now include experiments with a ConvHVAE as the proposal and show that SNIS continues to improve performance (Table 1).

R3: The formulation of energy functions is not mentioned. We have moved the details from Appendix D to the main text.

R3: Benefit of the energy function formulation? Asymptotically, neither is superior, however, in practice, the energy function exploits different inductive biases than the VAE. Instead of directly specifying a generative distribution, it determines the distribution by scoring images. As we show in the experiments, the VAE and the energy function formulation are complementary and combining them produces the best results. We have added this intuition to the main text.