A Proofs

Theorem 1. *Let $p(x, z)$ be a distribution where $z \in \mathcal{Z}$. Consider a joint proposal distribution $q(z_{0:k})$ over $\mathcal{Z}^k$. Let $v(i) \subset \{0, \dots, k\} \setminus \{i\}$ for all i, and π be a categorical distribution over $\{0, \dots, k\}$. The following construction, which we denote the Generalized IWAE Bound, is a valid lower bound of the log-marginal-likelihood*

$$\mathbb{E}_{q(z_{0:k})} \ln \sum_{i=0}^{k} \pi_i \frac{p(x, z_i)}{q(z_i \mid z_{v(i)})} \leq \ln p(x), \tag{10}$$

Proof. To show the validity of this lower bound, note that

$$\begin{aligned}
\mathbb{E}_{q(z_0,\dots,z_k)} \left[\sum_i \pi_i \frac{p_\theta(x, z_i)}{q(z_i \mid z_{v(i)})} \right] &= \sum_i \pi_i \mathbb{E}_{q(z_0,\dots,z_k)} \frac{p_\theta(x, z_i)}{q(z_i \mid z_{v(i)})} && (26)\\
&= \sum_i \pi_i \mathbb{E}_{q(z_{v(i)})} \mathbb{E}_{q(z_i \mid z_{v(i)})} \frac{p_\theta(x, z_i)}{q(z_i \mid z_{v(i)})} && (27)\\
&= \sum_i \pi_i \mathbb{E}_{q(z_{v(i)})} p_\theta(x) && (28)\\
&= p_\theta(x). && (29)
\end{aligned}$$

Applying Jensen's inequality shows that the lower bound in the theorem is valid. □

Lemma 1. *The BSVI gradient with θ is*

$$\nabla_\theta BSVI(x) = \mathbb{E}_{q_{\text{sir}}(z \mid x)} \nabla_\theta \ln p_\theta(x, z), \tag{16}$$

where q_{sir} is a sampling-importance-resampling procedure defined by the generative process

$$\begin{aligned}
z_{0:k} &\sim q(z_{0:k} \mid x) && (17)\\
i &\sim r(i \mid z_{0:k}) && (18)\\
z &\leftarrow z_i, && (19)
\end{aligned}$$

and $r(i \mid z_{0:k}) = (\pi_i w_i)/(\sum_j \pi_j w_j)$ is a probability mass function over $\{0, \dots, k\}$.

Proof.

$$\begin{aligned}
\nabla_\theta \text{BSVI}(x) &= \mathbb{E}_{q(z_{0:k} \mid x)} \nabla_\theta \ln \sum_{i=0}^{k} \pi_i w_i && (30)\\
&= \mathbb{E}_{q(z_{0:k} \mid x)} \mathbb{E}_{r(i \mid z_{0:k})} \nabla_\theta \ln p_\theta(x, z_i), && (31)
\end{aligned}$$

The double-expectation can now be reinterpreted as the *sampling-importance-resampling* distribution q_{sir}. □

Theorem 2. *When $q(z_{0:k}) = \prod_i q_i(z_i)$, the implicit distribution $q_{\text{sir}}(z)$ admits the inequality*

$$\mathbb{E}_{q_{\text{sir}}(z)} \ln \frac{p_\theta(x,z)}{q_{\text{sir}}(z)} \geq \mathbb{E}_{q(z_{0:k})} \ln \sum_{i=0}^{k} \pi_i w_i \tag{20}$$

$$= \mathbb{E}_{q(z_{0:k})} \ln \sum_{i=0}^{k} \pi_i \frac{p_\theta(x,z)}{q_i(z_i)}. \tag{21}$$

Proof. Recall that the q_{sir} is defined by the following sampling procedure

$$(z_0, \ldots, z_k) \sim q(z_0, \ldots, z_k) \tag{32}$$

$$i \sim r(i \mid z_{0:k}) \tag{33}$$

$$z \leftarrow z_i, \tag{34}$$

where

$$r(i \mid z_{0:k}) = \frac{\pi_i w_i}{\sum_j \pi_j w_j} = \frac{\pi_i \frac{p(x,z_i)}{q(z_i \mid z_{<i})}}{\sum_j \pi_j \frac{p(x,z_i)}{q(z_j \mid z_{<j})}} \tag{35}$$

We first note that, for any distribution $r(z)$

$$r(z) = \int_a r(a)\delta_z(a)\,\mathrm{d}a = \mathbb{E}_{r(a)}\delta_z(a). \tag{36}$$

This provides an intuitive way of constructing the probability density function by reframing it as a sampling process (the expectation w.r.t. $r(a)$) paired with a filtering procedure (the dirac-delta $\delta_z(a)$). Thus, the density under q_{sir} is thus

$$q_{\text{sir}}(z) = \mathbb{E}_{q(z_{0:k})}\mathbb{E}_{r(i|z_{0:k})}\delta_z(z_i). \tag{37}$$

Additionally, we shall introduce the following terms

$$\tilde{p}(z) = p_\theta(x,z) \tag{38}$$

$$\bar{w}_i = \frac{\pi_i w_i}{\sum_j \pi_j w_j} \tag{39}$$

$$\bar{v}_i = \frac{w_i}{\sum_j \pi_j w_j} \tag{40}$$

$$\bar{v}_i(z) = \frac{w(z)}{\pi_i w(z) + \sum_{-i} \pi_j w_j}. \tag{41}$$

for notational simplicity. Note that the density function q_{sir} can be re-expressed as

$$q_{\text{sir}}(z) = \mathbb{E}_{q(z_{0:k})}\mathbb{E}_{r(i|z_{0:k})}\delta_z(z_i) \tag{42}$$

$$= \mathbb{E}_{q(z_{0:k})} \sum_i \pi_i \bar{v}_i \delta_z(z_i) \tag{43}$$

$$= \mathbb{E}_{\pi(i)}\mathbb{E}_{q_{z_{-i}}}\mathbb{E}_{q_i(z_i|z_{-i})}\bar{v}_i\delta_z(z_i) \tag{44}$$

$$= \mathbb{E}_{\pi(i)}\mathbb{E}_{q_{z_{-i}}}\bar{v}_i(z)q_i(z \mid z_{-i}). \tag{45}$$

We now begin with the ELBO under $q_{\text{sir}}(z)$ and proceed from there via

$$\mathbb{E}_{q_{\text{sir}}(z)} \ln \frac{\tilde{p}(z)}{q_{\text{sir}}(z)} = -\tilde{D}(q_{\text{sir}}(z) \parallel \tilde{p}(z)) \tag{46}$$

$$= -\tilde{D}(\mathbb{E}_{\pi(i)}\mathbb{E}_{q_{z_{-i}}}\bar{v}_i(z)q_i(z \mid z_{-i}) \parallel \tilde{p}(z)) \tag{47}$$

$$\geq -\mathbb{E}_{\pi(i)}\mathbb{E}_{q_{z_{-i}}}\tilde{D}(\bar{v}_i(z)q_i(z \mid z_{-i}) \parallel \tilde{p}(z)), \tag{48}$$

where we use Jensen's Inequality to exploit the convexity of the unnormalized Kullback-Leibler divergence $\tilde{D}(\cdot \,\|\, \cdot)$. We now do a small change of notation when rewriting the unnormalized KL as an integral to keep the notation simple

$$\mathbb{E}_{q_{\text{sir}}}(z) \ln \frac{\tilde{p}(z)}{q_{\text{sir}}(z)} \geq \mathbb{E}_{\pi(i)} \mathbb{E}_{q_{z_{-i}}} \int_{z_i} \bar{v}_i q(z_i \mid z_{-i}) \ln \frac{\tilde{p}(z_i)}{\bar{v}_i q(z_i \mid z_{-i})} \tag{49}$$

$$= \mathbb{E}_{\pi(i)} \mathbb{E}_{q_{z_{-i}}} \mathbb{E}_{q(z_i \mid z_{-i})} \bar{v}_i \ln \frac{\tilde{p}(z_i)}{\bar{v}_i q(z_i \mid z_{-i})} \tag{50}$$

$$= \mathbb{E}_{q(z_{0:k})} \sum_i \bar{w}_i \ln \frac{\tilde{p}(z_i)}{\bar{v}_i q(z_i \mid z_{-i})} \tag{51}$$

$$= \mathbb{E}_{q(z_{0:k})} \sum_i \bar{w}_i \ln \left(\sum_j \pi_j w_j \cdot \frac{q(z_i \mid z_{<i})}{q(z_i \mid z_{-i})} \right) \tag{52}$$

$$= \mathbb{E}_{q(z_{0:k})} \sum_i \bar{w}_i \left[\ln \left(\sum_j \pi_j w_j \right) + \ln \left(\frac{q(z_i \mid z_{<i})}{q(z_i \mid z_{-i})} \right) \right] \tag{53}$$

$$\tag{54}$$

If $z_{0:k}$ are independent, then it follows that $q(z_i \mid z_{<i}) = q(z_i \mid z_{-i}) = q(z_i)$. Thus,

$$\mathbb{E}_{q_{\text{sir}}}(z) \ln \frac{\tilde{p}(z)}{q_{\text{sir}}(z)} \geq \mathbb{E}_{q(z_{0:k})} \sum_i \bar{w}_i \left[\ln \left(\sum_j \pi_j w_j \right) + \ln \left(\frac{q(z_i \mid z_{<i})}{q(z_i \mid z_{-i})} \right) \right] \tag{55}$$

$$= \mathbb{E}_{q(z_{0:k})} \sum_i \bar{w}_i \ln \left(\sum_j \pi_j w_j \right) \tag{56}$$

$$= \mathbb{E}_{q(z_{0:k})} \ln \left(\sum_j \pi_j w_j \right). \tag{57}$$

□

B Model Performance on Test and Training Data

Here we report various performance metrics for each type of model trained on the training set for both Omniglot and SVHN. As stated earlier, log-likelihood is estimated using BSVI-500, and ELBO* refers to the lower bound achieved by SVI-500 (i.e. $z \sim q_{500}$). KL* and Reconstruction* are the rate and distortion terms for ELBO*, respectively.

$$\text{Log-likelihood} = \mathbb{E}_{q(z_{0:500}|x)}\left[\ln \sum_{i=0}^{500} \pi_i \frac{p_\theta(x, z_i)}{q(z_i \mid z_{<i}, x)}\right] \tag{58}$$

$$\text{ELBO*} = \underbrace{\mathbb{E}_{q_{500}}\left[\ln p_\theta(x \mid z)\right]}_{\text{Reconstruction*}} + \underbrace{\tilde{D}(q_{500}(z) \parallel p_\theta(z))}_{\text{KL*}} \tag{59}$$

Table 4: Test set performance on the Omniglot dataset. Note that $k = 9$ and $k' = 10$ (see Section 6.1). We approximate the log-likelihood with BSVI-500 bound (Appendix C). We additionally report the SVI-500 bound (denoted ELBO*) along with its KL and reconstruction decomposition.

Model	Log-likelihood	ELBO*	KL*	Reconstruction*
VAE	-89.83 ± 0.03	-89.88 ± 0.02	0.97 ± 0.13	88.91 ± 0.15
IWAE-k'	-89.02 ± 0.05	-89.89 ± 0.06	4.02 ± 0.18	85.87 ± 0.15
SVI-k'	-89.65 ± 0.06	**-89.73 ± 0.05**	1.37 ± 0.15	88.36 ± 0.20
BSVI-k-DS	-88.93 ± 0.02	-90.13 ± 0.04	8.13 ± 0.17	81.99 ± 0.14
BSVI-k	-88.98 ± 0.03	-90.19 ± 0.06	8.29 ± 0.25	81.89 ± 0.20
BSVI-k-π	-88.95 ± 0.02	-90.18 ± 0.05	8.48 ± 0.22	**81.70 ± 0.18**
BSVI-k-SIR	**-88.80 ± 0.03**	-90.24 ± 0.06	7.52 ± 0.21	82.72 ± 0.22
BSVI-k-SIR-π	-88.84 ± 0.05	-90.22 ± 0.02	7.44 ± 0.04	82.78 ± 0.05

Table 5: Test set performance on the grayscale SVHN dataset.

Model	Log-likelihood	ELBO*	KL*	Reconstruction*
VAE	-2202.90 ± 14.95	-2203.01 ± 14.96	0.40 ± 0.07	2202.62 ± 14.96
IWAE-k'	-2148.67 ± 10.11	-2153.69 ± 10.94	2.03 ± 0.08	2151.66 ± 10.86
SVI-k'	-2074.43 ± 10.46	-2079.26 ± 9.99	45.28 ± 5.01	2033.98 ± 13.38
BSVI-k-DS	**-2054.48 ± 7.78**	**-2060.21 ± 7.89**	48.82 ± 4.66	2011.39 ± 9.35
BSVI-k	-2054.75 ± 8.22	-2061.11 ± 8.33	51.12 ± 3.80	**2009.99 ± 8.52**
BSVI-k-π	-2060.01 ± 5.00	-2065.45 ± 5.88	47.24 ± 4.62	2018.21 ± 1.64
BSVI-k-SIR	-2059.62 ± 3.54	-2066.12 ± 3.63	51.24 ± 5.03	2014.88 ± 5.30
BSVI-k-SIR-π	-2057.53 ± 4.91	-2063.45 ± 4.34	49.14 ± 5.62	2014.31 ± 8.25

Table 6: Test set performance on the FashionMNIST dataset.

Model	Log-likelihood	ELBO*	KL*	Reconstruction*
VAE	-1733.86 ± 0.84	-1736.49 ± 0.73	11.62 ± 1.01	1724.87 ± 1.70
IWAE-k'	-1705.28 ± 0.66	-1710.11 ± 0.72	33.04 ± 0.36	1677.08 ± 0.70
SVI-k'	-1710.15 ± 2.51	-1718.39 ± 2.13	26.05 ± 1.90	1692.34 ± 4.03
BSVI-k-DS	-1699.14 ± 0.18	-1706.92 ± 0.11	41.73 ± 0.18	1665.19 ± 0.26
BSVI-k	**-1699.01 ± 0.33**	**-1706.62 ± 0.35**	41.48 ± 0.16	**1665.14 ± 0.39**
BSVI-k-π	-1699.24 ± 0.36	-1706.92 ± 0.37	41.60 ± 0.49	1665.32 ± 0.31
BSVI-k-SIR	-1699.44 ± 0.45	-1707.00 ± 0.49	41.48 ± 0.12	1665.52 ± 0.41
BSVI-k-SIR-π	-1699.09 ± 0.28	-1706.68 ± 0.26	41.18 ± 0.19	1665.50 ± 0.31

Table 7: Training set performance on the Omniglot dataset. Note that $k = 9$ and $k' = 10$ (see Section 6.1). We approximate the log-likelihood with BSVI-500 bound (Appendix C). We additionally report the SVI-500 bound (denoted ELBO*) along with its KL and reconstruction decomposition.

Model	Log-likelihood	ELBO*	KL*	Reconstruction*
VAE	-88.60 ± 0.18	-88.66 ± 0.18	1.00 ± 0.13	87.66 ± 0.19
IWAE-k'	-87.09 ± 0.12	**-87.88 ± 0.12**	4.18 ± 0.19	83.70 ± 0.29
SVI-k'	-88.09 ± 0.16	-88.18 ± 0.15	1.38 ± 0.14	86.80 ± 0.27
BSVI-k-SIR	-87.24 ± 0.22	-88.57 ± 0.25	7.67 ± 0.22	80.89 ± 0.44
BSVI-k-DS	**-87.00 ± 0.11**	-88.13 ± 0.10	8.30 ± 0.18	79.83 ± 0.23
BSVI-k	-87.11 ± 0.11	-88.23 ± 0.10	8.45 ± 0.22	79.77 ± 0.28
BSVI-k-π	-87.10 ± 0.11	-88.24 ± 0.10	8.67 ± 0.27	**79.57 ± 0.31**
BSVI-k-SIR-π	-87.17 ± 0.10	-88.45 ± 0.11	7.63 ± 0.04	80.83 ± 0.13

Table 8: Training set performance on the grayscale SVHN dataset.

Model	Log-likelihood	ELBO*	KL*	Reconstruction*
VAE	-2384 ± 13.58	-2384 ± 13.59	0.5 ± 0.09	2384 ± 13.58
IWAE-k'	-2345 ± 8.77	-2350 ± 9.58	2.19 ± 0.03	2348 ± 9.55
SVI-k'	-2274 ± 8.87	-2280 ± 8.34	56 ± 5.75	2224 ± 12.20
BSVI-k-SIR	-2260 ± 2.73	-2268 ± 3.01	62.17 ± 5.51	2206 ± 4.86
BSVI-k-DS	**-2255.28 ± 7.38**	**-2262 ± 7.51**	59 ± 5.34	2203 ± 9.28
BSVI-k	-2255.47 ± 7.31	-2263 ± 7.47	62.20 ± 4.27	**2201 ± 8.26**
BSVI-k-π	-2261 ± 5.09	-2268 ± 6.13	58 ± 5.10	2210 ± 1.43
BSVI-k-SIR-π	-2258 ± 3.89	-2265 ± 3.30	60 ± 6.36	2206 ± 7.79

Table 9: Training set performance on the FashionMNIST dataset.

Model	Log-likelihood	ELBO*	KL*	Reconstruction*
VAE	-1686.11 ± 2.40	-1688.84 ± 2.27	11.34 ± 1.01	1677.50 ± 3.16
IWAE-k'	-1659.12 ± 0.59	-1663.80 ± 0.53	33.62 ± 0.31	1630.18 ± 0.56
SVI-k'	-1666.89 ± 2.47	-1675.15 ± 2.11	25.34 ± 1.80	1649.81 ± 3.90
BSVI-k-SIR	-1653.34 ± 1.36	-1660.79 ± 1.35	41.91 ± 0.15	1618.88 ± 1.51
BSVI-k-DS	-1653.47 ± 0.85	-1661.15 ± 0.82	42.13 ± 0.23	1619.02 ± 1.05
BSVI-k	**-1652.87 ± 0.87**	**-1660.27 ± 0.89**	41.85 ± 0.12	**1618.43 ± 0.86**
BSVI-k-π	-1654.35 ± 0.74	-1661.99 ± 0.68	42.00 ± 0.48	1619.99 ± 1.09
BSVI-k-SIR-π	-1654.75 ± 1.19	-1662.18 ± 1.25	41.58 ± 0.22	1620.60 ± 1.44

C Log-likelihood Estimation Using BSVI and IWAE

A popular way to approximate the true log-likelihood is to use the IWAE-k bound with a sufficiently large k during evaluation time [21, 2, 25]. Here we compare log-likelihood estimates of BSVI and IWAE in Tables 10 and 11 and empirically show that BSVI bounds are as tight as IWAE bounds in all of our experiments. This justifies the use of BSVI-500 for estimating log-likelihood in our reports.

Table 10: Log-likelihood estimates using BSVI-k and IWAE-k on the Omniglot test set. The tightest estimate is bolded for each model unless there is a tie. Note that k is fixed to 500, and for IWAE we use five different numbers of particles: $k, 2k, 3k, 4k, 5k$.

Model	**BSVI-k**	**IWAE-k**	**IWAE-$2k$**	**IWAE-$3k$**	**IWAE-$4k$**	**IWAE-$5k$**
VAE	-89.83	-89.83	-89.83	-89.83	-89.83	-89.83
SVI-k'	-89.65	-89.65	-89.65	-89.65	-89.65	-89.65
IWAE-k'	**-89.02**	-89.05	-89.04	-89.03	-89.03	-89.03
BSVI-k-DS	**-88.93**	-89.05	-89.00	-88.99	-88.98	-88.97
BSVI-k	**-88.98**	-89.10	-89.06	-89.04	-89.03	-89.02
BSVI-k-SIR	**-88.80**	-88.92	-88.88	-88.86	-88.85	-88.84
BSVI-k-π	**-88.95**	-89.07	-89.03	-89.01	-89.00	-88.99
BSVI-k-SIR-π	**-88.84**	-88.95	-88.91	-88.89	-88.88	-88.87

Table 11: Log-likelihood estimates using BSVI-k vs. IWAE-k on the SVHN test set. The tightest estimate is bolded for each model unless there is a tie. Note that k is fixed to 500, and for IWAE we use five different numbers of particles: $k, 2k, 3k, 4k, 5k$.

Model	**BSVI-k**	**IWAE-k**	**IWAE-$2k$**	**IWAE-$3k$**	**IWAE-$4k$**	**IWAE-$5k$**
VAE	-2203	-2203	-2203	-2203	-2203	-2203
SVI-k'	**-2074**	-2096	-2095	-2094	-2094	-2093
IWAE-k'	-2149	-2149	-2149	-2149	-2149	-2149
BSVI-k-DS	**-2054**	-2079	-2078	-2077	-2077	-2077
BSVI-k	**-2055**	-2081	-2080	-2080	-2079	-2079
BSVI-k-SIR	**-2060**	-2087	-2086	-2085	-2085	-2084
BSVI-k-π	**-2060**	-2085	-2083	-2083	-2082	-2082
BSVI-k-SIR-π	**-2058**	-2083	-2082	-2081	-2081	-2080

Table 12: Log-likelihood estimates using BSVI-k vs. IWAE-k on the FashionMNIST test set. The tightest estimate is bolded for each model unless there is a tie. Note that k is fixed to 500, and for IWAE we use five different numbers of particles: $k, 2k, 3k, 4k, 5k$.

Model	**BSVI-k**	**IWAE-k**	**IWAE-$2k$**	**IWAE-$3k$**	**IWAE-$4k$**	**IWAE-$5k$**
VAE	**-1733.86**	-1737.76	-1737.49	-1737.35	-1737.25	-1737.18
SVI-k'	**-1705.28**	-1727.30	-1726.26	-1725.72	-1725.35	-1725.07
IWAE-k'	**-1710.15**	-1721.01	-1720.23	-1719.80	-1719.51	-1719.29
BSVI-k-DS	**-1699.14**	-1727.55	-1726.37	-1725.71	-1725.25	-1724.93
BSVI-k	**-1699.01**	-1727.38	-1726.19	-1725.53	-1725.09	-1724.75
BSVI-k-SIR	**-1699.24**	-1727.48	-1726.28	-1725.63	-1725.19	-1724.86
BSVI-k-π	**-1699.44**	-1728.05	-1726.88	-1726.23	-1725.77	-1725.44
BSVI-k-SIR-π	**-1699.09**	-1727.03	-1725.86	-1725.20	-1724.77	-1724.44

D Experiment Setup

Here we describe our detailed experiment setup. For both Omniglot and SVHN experiments, we used a ResNet with three hidden layers of size 64 as the encoder and a 12-layer gated PixelCNN with the constant layer size of 32 as the decoder. Network parameters (ϕ, θ) were trained with the AMSGrad optimizer [24]. For SVI, we followed the experimental setup of [18] and optimized local variational parameters $\lambda_{0:k}$ with SGD with momentum with learning rate 1.0 and momentum 0.5. To stabilize training, we applied gradient clipping to both network parameters and local variational parameters. Each model was trained for 200k steps with early-stopping based on validation loss. The best-performing models on the validation set were then evaluated on the test set. All experiments were performed four times, and we reported the mean and standard deviation of relevant metrics.

Omniglot. We used 2000 randomly-selected training images as the validation set. Each digit was dynamically binarized at training time based on the pixel intensity. We used 32-dimensional latent variable with unit Gaussian prior. Each pixel value was modeled as a Bernoulli random variable where the output of the decoder was interpreted as log probabilities. We also followed the training procedure in [18] and annealed the KL term multiplier [2, 26] from 0.1 to 1.0 during the first 5000 iterations of training.

SVHN. We merged "train" and "extra" data in the original SVHN dataset to create our training set. We again reserved 2000 randomly-selected images as the validation set. To reuse the network architecture for the Omniglot dataset with minimal modifications, we gray-scaled all images and rescaled the pixel intensities to be in $[0, 1]$. The only differences from Omniglot experiments are: increased latent variable dimensions (64), larger image size (32×32), and the use of discretized logistic distribution by [30] with a global scale parameter for each pixel. Similar to [31], we lower-bound the scale parameter by a small positive value.

FashionMNIST. Similar to above, we used 2000 randomly-selected training images as the validation set. The network architecture and hyperparameters were identical to those of SVHN dataset, except we used 32-dimensional latent variables and did not employ KL term annealing.

Below is the list hyperparameters used in our experiments. Since we have two stochastic optimization processes (one for the model and one for SVI), we employed separate gradient clipping norms.

Table 13: Hyperparameters used for our experiments.

Hyperparameter	**Omniglot**	**SVHN**	**FashionMNIST**
Learning rate	0.001	0.001	0.001
SVI learning rate	1.0	1.0	1.0
SVI momentum	0.5	0.5	0.5
Batch size	50	50	50
KL-cost annealing steps	5000	0	0
Max gradient norm (ϕ, θ)	5.0	5.0	5.0
Max gradient norm (SVI)	1.0	1.0	1.0
Latent variable dimension	32	64	32
Observation model	Bernoulli	Discretized Logistic	Discretized Logistic
Scale parameter lower bound	N/A	0.001	0.001