Implicit Variational Rejection Sampling

Jian Xu[*1,2] **Shigui Li**[3] **Wei Chen**[3] **Jiacheng Li**[3] **Zhiqi Lin**[3]
Delu Zeng[†3] **Xinghao Ding**[4] **John Paisley**[5] **Qibin Zhao**[‡2]

[1]RIKEN iTHEMS
[2]RIKEN AIP
[3]South China University of Technology
[4]Xiamen University
[5]Columbia University

Abstract

Variational Inference (VI) is a fundamental inference technique in Bayesian machine learning for approximating complex posterior distributions. Traditional VI often relies on the mean-field factorization, which can inadequately capture true posterior complexity. Recent advancements have leveraged neural networks to model implicit distributions, offering increased flexibility. However, the practical constraints of neural network architectures still produces inaccuracies. In this paper, we propose a method called Implicit Variational Rejection Sampling (IVRS), which integrates implicit distributions with rejection sampling to improve the posterior approximation. Our method uses neural networks to construct implicit proposal distributions, and rejection sampling with a discriminator network that estimates the density ratio between the implicit proposal and the true posterior for refining the approximation. Towards this end, we introduce the Implicit Resampling Evidence Lower Bound (IR-ELBO) as a metric to characterize the resampled distribution's quality and derive a tighter variational lower bound. Experimental results demonstrate that our method outperforms traditional variational inference techniques.

1 INTRODUCTION

Variational Inference (VI) has emerged as a fundamental technique in Bayesian machine learning for approximating complex posterior distributions [Jordan et al., 1999, Hoffman et al., 2013]. Traditional VI methods frequently rely on a mean-field assumption [Blei et al., 2017], which trades off posterior expressiveness for computational tractability. To address this limitation, implicit distributions, which are typically modeled using neural networks, have been proposed to leverage their flexibility when approximating complex posterior distributions [Mescheder et al., 2017, Huszár, 2017, Titsias and Ruiz, 2019, Shi et al., 2018]; such implicit and diffusion-based constructions have also been extended to richer Bayesian models such as deep Gaussian processes [Xu et al., 2024c, 2026b]. While neural networks are highly expressive in theory [Hornik et al., 1989, Krizhevsky et al., 2012, LeCun et al., 2015], they may still struggle to match complex true posteriors in practice, especially under limited capacity, poor initialization, or optimization difficulty [Arora et al., 2017]. As a result, the posterior approximations with neural networks are not robust as an off-the-shelf method.

To improve neural network posterior approximations, we propose Implicit Variational Rejection Sampling (IVRS), which leverages rejection sampling [Gilks and Wild, 1992] to better exploit the strengths of implicit distributions. We first use neural networks to construct implicit distributions that serve as proposals. We then design an acceptance probability function related to the density ratio between the proposal distribution and the true posterior, and apply rejection sampling to generate resampled samples. A discriminator network is used to approximate the density ratio, thereby refining the proposal distribution into a more accurate posterior approximation. By incorporating adversarial training techniques, this approach enables us to construct an Implicitly Resampled Evidence Lower Bound (IR-ELBO).

We summarize our contributions below:

1) We introduce Implicit Variational Rejection Sampling (IVRS) to combine implicit distributions with rejection sampling for more accurate variational inference with neural networks.
2) By incorporating adversarial training techniques, we construct an Implicit Resampling Evidence Lower Bound (IR-ELBO) and analyze the resampled distri-

[*]Email: jian.xu@riken.jp
[†]Corresponding author. Email: dlzeng@scut.edu.cn
[‡]Corresponding author. Email: qibin.zhao@riken.jp

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7651–7669.

bution, particularly its reduced KL divergence from the true posterior, providing theoretical support for improved accuracy.

3) We demonstrate through experiments that IVRS can outperform traditional variational inference methods in terms of accuracy and efficiency.

2 MODEL FRAMEWORK

2.1 BAYESIAN GENERATIVE MODELS

Consider an unsupervised generative model for a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ that has latent variables $\mathbf{z}$ and model parameters θ. The joint distribution of the model has the form

$$p(\mathbf{x}, \mathbf{z}|\theta) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z}, \theta), \tag{1}$$

where $p(\mathbf{z})$ is the prior distribution of $\mathbf{z}$ and $p(\mathbf{x}|\mathbf{z}, \theta)$ is a parametric generative model. In more traditional models, such as the Gaussian Mixture Model (GMM), this distribution is often specified through manual design. With the advance of deep learning, and generative models in particular as exemplified by Variational Autoencoders (VAEs), this distribution is frequently parameterized using a neural network. While this framework can be extended to supervised learning scenarios, we develop our method in the unsupervised learning regime.

2.2 VARIATIONAL INFERENCE

The goal of inference is to model the posterior distribution of the latent variables in Equation 1. For non-conjugate models, such as those involving deep learning architectures, the posterior distribution is highly complex and requires approximate methods. Variational inference (VI) provides a KL divergence approximation to the posterior distribution $p(\mathbf{z}|\mathbf{x})$ using a predefined variational distribution $q(\mathbf{z}|\phi)$ by maximizing the Evidence Lower Bound (ELBO),

$$\mathcal{L}(\mathbf{x}, \theta, \phi) = \mathbb{E}_{q(\mathbf{z}|\phi)} \left[\log p(\mathbf{x}, \mathbf{z}|\theta) - \log q(\mathbf{z}|\phi)\right]. \tag{2}$$

The traditional mean-field approximation assumes a factorized form for the variational distribution,

$$q(\mathbf{z}|\mathbf{x}, \phi) = \textstyle\prod_{i=1}^m q(\mathbf{z}_i|\mathbf{x}, \phi_i), \tag{3}$$

where m represents the number of factors in the decomposition, and each $q(\mathbf{z}_i|\mathbf{x}, \phi_i)$ is often a simple parametric distribution.

The mean field approximation sacrifices accuracy for tractability. To address this limitation, the implicit variational inference method employs a neural network parameterization of the variational distribution. These methods aim to capture more accurate, complex posterior structures by leveraging the expressive power of neural networks. They take the form

$$\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}) \quad \longrightarrow \quad \mathbf{z} = f_\phi(\mathbf{x}, \epsilon), \;\; \epsilon \sim p(\epsilon). \tag{4}$$

Here, ϕ represents the parameters of a neural network, while ϵ is drawn independently from a simple distribution such as a Gaussian. Recently, various algorithms have been proposed to effectively train such models, including Adversarial Variational Bayes [Mescheder et al., 2017] and Semi-implicit Variational Inference [Yin and Zhou, 2018].

Despite their flexibility, traditional neural networks face practical limitations, and their structural design often relies on empirical heuristics. Rejection sampling [Naesseth et al., 2017, Jankowiak and Phan, 2024] offers a flexible approach to more robust implicit distribution learning without requiring additional model capacity or architectural changes. We therefore view rejection sampling as a complementary mechanism for improving implicit VI, particularly when the variational family lacks sufficient support. In the following section, we introduce a method that incorporates rejection sampling to address these challenges.

3 PROPOSED METHOD

3.1 REJECTION SAMPLING

Rejection sampling is a standard statistical technique for generating samples from a target distribution via a proposal distribution. Given a target distribution $p_{\text{tar}}(\mathbf{z})$ and a proposal distribution $q_{\text{pro}}(\mathbf{z})$, rejection sampling accepts a sample $\mathbf{z} \sim q_{\text{pro}}(\mathbf{z})$ with probability defined by an acceptance probability function $a(\mathbf{z})$ that is proportional to the ratio of the target density to the proposal density,

$$a(\mathbf{z}) = p_{\text{tar}}(\mathbf{z})/M q_{\text{pro}}(\mathbf{z}), \tag{5}$$

where $M > 0$ and the choice of M ensures that the acceptance probability is less than or equal to 1. In our model, the target distribution $p_{\text{tar}}(\mathbf{z}) = p_\theta(\mathbf{z}|\mathbf{x})$, being the true posterior of the latent variables of model structure Equation 1. We define the proposal distribution to be the implicit distribution in Equation 5, $q_{\text{pro}}(\mathbf{z}) = q_\phi(\mathbf{z}|\mathbf{x})$. Therefore, we write the acceptance rate function as $a(\mathbf{z}; \mathbf{x}, \theta, \phi)$.

However, unlike traditional rejection sampling, calculating the acceptance rate function $a(\mathbf{z}; \mathbf{x}, \theta, \phi)$ in our model is not analytical. The primary challenges are two-fold: i) The target distribution $p_\theta(\mathbf{z}|\mathbf{x})$ is the true posterior, typically only representable in the unnormalized joint likelihood form of Bayes rule in Equation 1; ii) The proposal distribution $q_\phi(\mathbf{z}|\mathbf{x})$ is an implicit distribution, often generated by a structure that makes it difficult to determine its probability density function. We next turn to our proposal for addressing these two issues.

3.2 THE ACCEPTANCE PROBABILITY FUNCTION

To address these two challenges, we first express the target distribution $p_\theta(\mathbf{z}|\mathbf{x})$ using Bayes rule,

$$p_\theta(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})/p_\theta(\mathbf{x}), \tag{6}$$

where $p_\theta(\mathbf{x}|\mathbf{z})$ is the likelihood, $p(\mathbf{z})$ is the prior, and $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,d\mathbf{z}$ is the evidence. To ensure the acceptance probability is within the range $[0, 1]$, we can ignore the evidence term provided we choose an appropriate scaling factor M such that

$$a(\mathbf{z}; \mathbf{x}, \theta, \phi) = \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{Mq_\phi(\mathbf{z}|\mathbf{x})} \leq 1. \tag{7}$$

To ensure this constraint in practice, the acceptance rate function is usually constructed as

$$a(\mathbf{z}; \mathbf{x}, \theta, \phi) = \min\left[\frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{Mq_\phi(\mathbf{z}|\mathbf{x})}, 1\right]. \tag{8}$$

However, the min function makes gradient-based optimization for variational posterior parameters challenging. To address this, we adopt the fully differentiable approximation of Grover et al. [2018],

$$a(\mathbf{z}; \mathbf{x}, \theta, \phi) = \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z}) + Mq_\phi(\mathbf{z}|\mathbf{x})} \in (0, 1) \tag{9}$$

This approximation addresses the first challenge.

For the second challenge, where the proposal distribution $q_\phi(\mathbf{z}|\mathbf{x})$ is implicitly modeled by a neural network ϕ, we estimate it using adversarial training. Specifically, we address the challenge of computing the term $\log p(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})$ by introducing an additional discriminative network $T(\mathbf{x}, \mathbf{z})$, which distinguishes between pairs $(\mathbf{x}, \mathbf{z})$ sampled from the true joint distribution $p(\mathbf{x}, \mathbf{z})$, and pairs $(\mathbf{x}, \mathbf{z})$ sampled using the implicit proposal distribution $q_\phi(\mathbf{z}|\mathbf{x})$. The objective $D(T)$ for this discriminator $T(\mathbf{x}, \mathbf{z})$ is

$$\begin{aligned} D(T) \; &= \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \sigma(T(\mathbf{x}, \mathbf{z}))\right] \\ &+ \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{p(\mathbf{z})}\left[\log(1 - \sigma(T(\mathbf{x}, \mathbf{z})))\right], \end{aligned} \tag{10}$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$ denotes the sigmoid function. By Goodfellow et al. [2014] and Mescheder et al. [2017], the optimal discriminator $T^*(\mathbf{x}, \mathbf{z})$ is

$$T^*(\mathbf{x}, \mathbf{z}) = \log q_\phi(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}). \tag{11}$$

We see that T^* can be directly substituted into Equation (9) to compute the implicit proposal distribution,

$$a(\mathbf{z}; \mathbf{x}, \theta, \phi) = \frac{p_\theta(\mathbf{x}|\mathbf{z})}{p_\theta(\mathbf{x}|\mathbf{z}) + M\exp\left(T^*(\mathbf{x}, \mathbf{z})\right)}. \tag{12}$$

As a result, we can effectively perform rejection sampling even in the absence of explicit analytical forms for the proposal distribution. Numerous methods are available for estimating density ratios of non-analytical distributions. We employ adversarial training here [Goodfellow et al., 2014, Mescheder et al., 2017], although alternative estimators, such as recent diffusion- and Schrödinger-bridge-based density-ratio estimation [Chen et al., 2025], are equally compatible with our framework. Additionally, the expectation in Equation (10) is on the outermost layer, so Monte Carlo estimation remains unbiased and is suitable for mini-batch algorithms.

Algorithm 1 Sampler for $r_{\theta,\phi}(\mathbf{z}|\mathbf{x})$

Require: $a_{\theta,\phi}(\mathbf{z}; \theta, \phi)$, $q_\phi(\mathbf{z}|\mathbf{x})$
Ensure: $\mathbf{z} \sim r_{\theta,\phi}(\mathbf{z}|\mathbf{x})$
1: Perform gradient ascent on $D(T_\eta)$ in Equation (10) with respect to η to obtain T^*_η
2: **while** True **do**
3: $\quad\mathbf{z} \leftarrow$ sample from implict proposal $q_\phi(\mathbf{z}|\mathbf{x})$ by Equation (4)
4: $\quad$Compute acceptance probability $a(\mathbf{z}; \mathbf{x}, \theta, \phi)$ by Equation (12)
5: $\quad$Sample uniform $u \sim \mathcal{U}[0, 1]$
6: $\quad$**if** $u < a(\mathbf{z}; \mathbf{x}, \theta, \phi)$ **then**
7: $\quad\quad$Output sample $\mathbf{z}$
8: $\quad$**end if**
9: **end while**

3.3 IMPLICIT RESAMPLING ELBO

Unlike previous implicit variational inference, we us as our variational approximation the distribution resampled via rejection sampling, denoted as

$$r_{\theta,\phi}(\mathbf{z}|\mathbf{x}) = \frac{q_\phi(\mathbf{z}|\mathbf{x})a(\mathbf{z}; \mathbf{x}, \theta, \phi)}{Z_{\theta,\phi}(\mathbf{x})}, \tag{13}$$

where $Z_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[a(\mathbf{z}; \mathbf{x}, \theta, \phi)]$.

To sample from $r_{\theta,\phi}$, we follow the procedure defined in Algorithm 1. First, we use a neural network parameterized by η to represent the discriminative network $T_\eta(\mathbf{x}, \mathbf{z})$. By using gradient-based optimization, we obtain a local optimal value $T^*_\eta(\mathbf{x}, \mathbf{z})$. Then, through an accept-reject step, we resample from the implicit proposal distribution $r_{\theta,\phi}$. We then define the *Implicit Resampling Evidence Lower Bound* (IR-ELBO) on the marginal log-likelihood of $\mathbf{x}$. This involves using the implicit distribution as the proposal distribution and the resampled distribution $r_{\theta,\phi}$ as the variational distribution. By Jensen's inequality, we have that,

$$\begin{aligned} \log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{r_{\theta,\phi}(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{r_{\theta,\phi}(\mathbf{z}|\mathbf{x})}\right] \\ &= \mathbb{E}_{r_{\theta,\phi}(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})Z_{\theta,\phi}(\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})a(\mathbf{z}; \mathbf{x}, \theta, \phi)}\right]. \end{aligned} \tag{14}$$

In this equation, we can use the discriminative network $T^*_\eta(\mathbf{x}, \mathbf{z})$ described in Algorithm 1 to compute the probability density function of the implicit distribution. Using

Algorithm 2 Implicit Variational Rejection Sampling

Require: Data $\mathbf{x}$, model parameters θ, neural network parameters ϕ and η

Ensure: Optimized parameters θ^* and ϕ^*

1: **Sample Generation:** Generate samples $\{\mathbf{z}_i\}_{i=1}^{Q}$ from implicit proposal $q_\phi(\mathbf{z}|\mathbf{x})$.
2: **Density Ratio:** Use discriminator network $T_\eta(\mathbf{x}, \mathbf{z})$ to estimate density ratio for each $\mathbf{z}_i$.
3: **Rejection Sampling:** Accept/reject $\mathbf{z}_i$ based on acceptance function $a(\mathbf{z}_i; \mathbf{x}_i, \theta, \phi)$. (Alg. 1)
4: **ELBO Optimization:** Compute the IR-ELBO by Equation (15) and Equation (16).
5: Update $\theta \leftarrow \theta + \alpha \nabla_\theta$IR-ELBO.
6: Update $\phi \leftarrow \phi + \beta \nabla_\phi$IR-ELBO.
7: Repeat steps 1 to 6 until convergence.

Equations (9) and (11), we therefore have that

$$\mathbb{E}_{r_{\theta,\phi}}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z}) Z_{\theta,\phi}(\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x}) a(\mathbf{z};\mathbf{x},\theta,\phi)}\right] = \mathbb{E}_{r_{\theta,\phi}}\left[\log\left(\frac{p_\theta(\mathbf{x}|\mathbf{z})}{\exp\left(T_\eta^*(\mathbf{x},\mathbf{z})\right)} + M\right)\right] + \log Z_{\theta,\phi}(\mathbf{x}). \tag{15}$$

For the term $\log Z_{\theta,\phi}(\mathbf{x})$, we can again use Jensen's inequality to define a lower bound,

$$\begin{aligned}\log Z_{\theta,\phi}(\mathbf{x}) &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log a(\mathbf{z};\theta,\phi)] \\ &= \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})}{p_\theta(\mathbf{x}|\mathbf{z}) + M\exp\left(T_\eta^*(\mathbf{x},\mathbf{z})\right)}\right].\end{aligned} \tag{16}$$

Substituting the lower bound for $\log Z_{\theta,\phi}(\mathbf{x})$ from Equation (16) into Equation (15) yields the final loss function, which we call the IR-ELBO. Similar to Equation (10), since the expectation is applied to the outermost layer, the Monte Carlo approximation of this objective function remains unbiased and is appropriate for mini-batch algorithms. Samples from the implicit distribution can be directly obtained, and by adjusting the parameter M they can be resampled.

3.4 IMPLICIT VARIATIONAL REJECTION SAMPLING

Using the above derivations, we propose a new inference method for generative models called Implicit Variational Rejection Sampling (IVRS), which combines the strengths of implicit distributions and rejection sampling to achieve a more accurate posterior approximation. The algorithm is summarized in Algorithm 2. Optimization of the discriminator network $T_\eta(\mathbf{x}, \mathbf{z})$ is reflected in both Algorithm 1 and Algorithm 2 — these steps can be merged in practice.

Discussion of IVRS. We briefly analyze the properties of the resampling distribution $r_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ and show that it is indeed a better approximation compared to the implicit proposal distribution $q_\phi(\mathbf{z}|\mathbf{x})$. To that end, we directly compute the KL divergence between $r_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ and the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$,

$$\mathrm{KL}\left(r_{\theta,\phi}(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right) = \int \frac{q_\phi(\mathbf{z}|\mathbf{x}) a(\mathbf{z};\mathbf{x},\theta,\phi)}{Z_{\theta,\phi}(\mathbf{x})} \log \frac{q_\phi(\mathbf{z}|\mathbf{x}) a(\mathbf{z};\mathbf{x},\theta,\phi)}{Z_{\theta,\phi}(\mathbf{x}) p_\theta(\mathbf{z}|\mathbf{x})} d\mathbf{z}. \tag{17}$$

We directly substitute Equation (9) into the above KL divergence. The RHS can then be rewritten as

$$\int \frac{p_\theta(\mathbf{z}|\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x})}{\Lambda} \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{\Lambda} d\mathbf{z}, \tag{18}$$

where $\Lambda = Z_{\theta,\phi}(\mathbf{x})\left(p_\theta(\mathbf{z}|\mathbf{x}) + M q_\phi(\mathbf{z}|\mathbf{x})\right)$.

We now characterize how this divergence depends on the scaling parameter M.

Proposition 1. *Consider the idealized setting in which the discriminator is optimal, so that the acceptance probability in Equation (9) uses the exact density ratio. Then* $\mathrm{KL}\left(r_{\theta,\phi}(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})\right)$ *is a monotonically non-increasing function of* M*. Moreover, as* $M \to 0$ *the acceptance rate approaches* 1 *and* $r_{\theta,\phi}$ *approaches* $q_\phi(\mathbf{z}|\mathbf{x})$*, whereas as* $M \to \infty$ *the acceptance rate approaches* 0 *and* $r_{\theta,\phi}$ *approaches the true posterior* $p_\theta(\mathbf{z}|\mathbf{x})$*.*

The key step is to write the acceptance probability in the reduced form $a_M(\mathbf{z}) = 1/(1 + M\, w(\mathbf{z}))$, with density ratio $w(\mathbf{z}) = q_\phi(\mathbf{z}|\mathbf{x})/p_\theta(\mathbf{z}|\mathbf{x})$. Differentiating the divergence then gives $\frac{d}{dM}\mathrm{KL}(r_{\theta,\phi} \parallel p_\theta) = -\mathrm{Cov}_{r_{\theta,\phi}}\big(c_M, \log c_M\big) \leq 0$, where $c_M = w/(1 + Mw)$ and the inequality follows because $\mathrm{Cov}(X, f(X)) \geq 0$ for any monotone non-decreasing f. A complete derivation is provided in Appendix A. It follows that, in our algorithm, M contributes positively to model improvement by driving $r_{\theta,\phi}$ to provide a more accurate approximation compared to the implicit proposal distribution $q_\phi(\mathbf{z}|\mathbf{x})$, thereby achieving a tighter variational lower bound. In our experiments we empirically determine M by cross-validation. When the discriminator is only approximately optimal, the acceptance probability and the induced resampled distribution become approximate, so the monotonicity above is best understood as an idealized analytical property rather than a strict guarantee.

4 EXPERIMENTS

We compare IVRS with other ELBO-based variational inference methods, including the Adversarial Variational Bayes (AVB) [Mescheder et al., 2017] and Semi-Implicit Variational Inference (SIVI) [Yin and Zhou, 2018], across a range of unsupervised learning tasks. These include some illustrative studies on several toy examples, followed by a comparison one a regression tast with Bayesian Neural Networks (BNNs) and an autoencoder task. We also consider

variants of SIVI, such as UIVI [Titsias and Ruiz, 2019]. We use the Adam optimizer [Kingma and Ba, 2015] and empirically select M for training via cross-validation. All experiments were conducted on an RTX 4090. The code will be available with a final draft of the paper.

Distribution	$-\mathbb{E}_{q_\phi}[\log p_{\text{tar}}]$	w/ reject samp	M
Gaussian	-0.095 ± 0.001	-0.102 ± 0.001	0.1
Laplace	-0.107 ± 0.001	-0.112 ± 0.001	0.1
GMM-1D	-0.101 ± 0.000	-0.104 ± 0.000	0.1
Banana	1.235 ± 0.005	1.189 ± 0.003	500
X-shape	0.744 ± 0.008	0.682 ± 0.005	500
GMM-2D	1.292 ± 0.006	1.214 ± 0.004	500

Table 1: A comparison of cross-entropy values before and after rejection sampling for different targets using implicit distributions as proposal distributions. We observe that rejection sampling provides improved approximate posterior samples. (Lower is better.)

4.1 DENSITY ESTIMATION OF TOY DATASETS

We IVRS to approximate six low dimensional synthetic distributions: A 1D Gaussian distribution, a 1D Laplace distribution, a 1D bimodal Gaussian Mixture Model (GMM), a 2D banana-shaped distribution, a 2D X-shaped mixture of Gaussians, and a 2D bimodal GMM. The six densities are pictured in Figure 1 and listed in Table 1 along with the performance results.

The dimension of ϵ was set to 10, and the network approximation f_ϕ was parameterized by a 4-layer MLP with layer widths $[20, 40, 20, 2]$. The output of f_ϕ was then combined with Gaussian noise. Since this task is straightforward, we adopt a Monte Carlo estimator to estimate q_ϕ and $\mathrm{KL}(q_\phi||p_{\text{tar}})$ for gradient-based optimization. The key difference in our method is that the trained neural network was used as the proposal distribution, followed by rejection sampling using the acceptance probability function defined in Equation 9. Following Yin and Zhou [2018], we used 100 iterations for each inner-loop of Monte Carlo sampling to estimate the entropy of the implicit distribution. All methods were trained with 50,000 parameter updates.

Figure 1 shows the contour plots of the synthetic distributions, along with kernel density estimates from samples drawn from the trained implicit distributions q_ϕ. Additionally, Figure 1 compares the distributions before and after applying rejection sampling. The results show the neural network's slight misalignment with the target distribution contour. However, after applying rejection sampling, our method demonstrates improved approximation with better alignment to the target distribution as expected. Due to the challenges in normalizing the distribution after rejection sampling, we instead report the cross-entropy between the target distributions and the approximate distributions in Table 1. We conducted 10 runs and report the mean and standard deviation of the cross-entropy values. As shown, our method outperforms across all toy target distributions by reducing the cross-entropy, further validating its effectiveness.

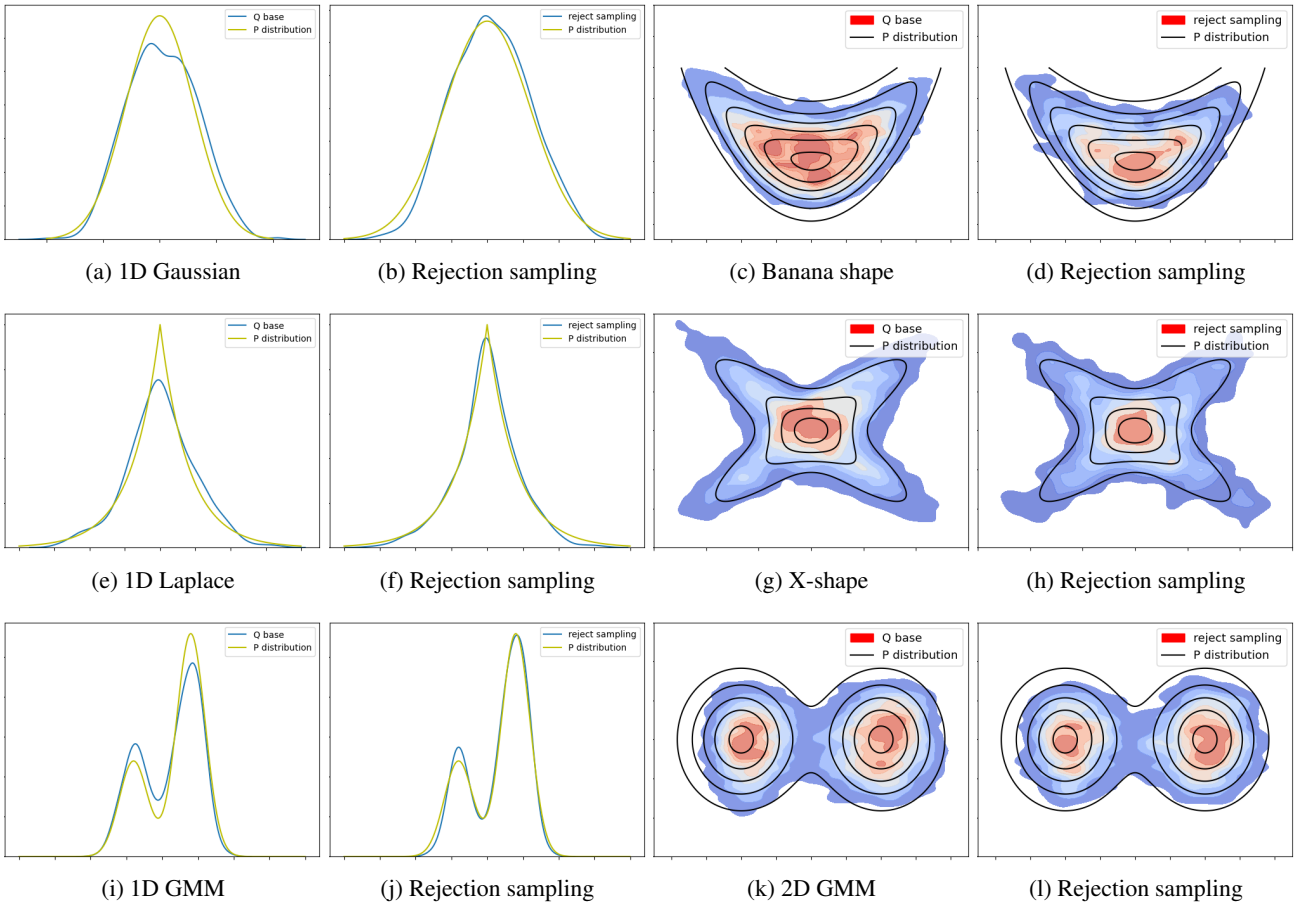
4.2 BAYESIAN NEURAL NETWORKS

We next consider the problem of sampling from the posterior of a Bayesian neural network (BNN). In this scenario, the latent model variables $\mathbf{z}$ correspond to the BNN weights. We utilize a two-layer network with 50 hidden units and ReLU activation functions. We compare our method with AVB, SIVI, and several SIVI variants, including UIVI, SIVI-SM [Yu and Zhang, 2023], and KSIVI [Cheng et al., 2024]. We use six common UCI datasets to perform these experiments. Each dataset is randomly partitioned, with 90% used for training and 10% for testing. Both the proposal distribution ϕ and the discriminator η are modeled using four-layer fully connected neural networks. The results are averaged over 10 random trials.

Table 2 presents the average test root mean squared error (RMSE) and negative log-likelihood (NLL) along with their standard deviations. The six UCI datasets considered are indicated by their names. As is evident, IVRS achieves competitive or superior performance compared to the baselines on nearly all problems, indicating that rejection sampling provides a more accurate representation of the BNN model variables. Additional ablations—fixed-architecture comparisons, discriminator-optimization sensitivity, latent-dimensionality scaling, and a decomposition of the gains—are provided in Appendix C.

4.3 SENSITIVITY TO THE REJECTION HYPERPARAMETER M

The hyperparameter M controls the acceptance probability and directly shapes the resampled variational distribution $r_{\theta,\phi}(\mathbf{z} \mid \mathbf{x})$: increasing M tightens the approximation by pushing $r_{\theta,\phi}$ toward the true posterior, at the cost of a lower acceptance rate and higher computational overhead. We evaluate IVRS with $M \in \{0.1, 1, 10, 100, 500\}$ on the three UCI BNN benchmarks (*Boston*, *Concrete*, *Protein*), keeping all other settings fixed and averaging over 10 runs. As shown in Table 3, larger M consistently lowers both NLL and RMSE—empirically confirming the monotonicity established in Proposition 1—while the acceptance rate decreases and the relative cost grows only moderately (within 2–3$\times$ even at $M = 500$). The main results in Table 2 correspond to $M = 100$; moderate values $M \in [10, 100]$ already capture most of the gains at a reasonable cost, which is why we select M in this range via lightweight cross-validation.



(a) 1D Gaussian (b) Rejection sampling (c) Banana shape (d) Rejection sampling

(e) 1D Laplace (f) Rejection sampling (g) X-shape (h) Rejection sampling

(i) 1D GMM (j) Rejection sampling (k) 2D GMM (l) Rejection sampling

Figure 1: Density estimation tasks for 1D and 2D toy datasets. The subplots sequentially show the kernel density curves estimated using the original implicit distribution Q base for the target distribution P, as well as the curves after applying rejection sampling. The color gradient represents the density of the generated samples, with red indicating higher concentration and blue indicating sparser regions. Table 1 shows quantitative values.

4.4 VARIATIONAL AUTOENCODER TASK ON MNIST

Variational Autoencoders (VAEs) [Kingma and Welling, 2013] are a popular method for unsupervised feature learning and dimensionality reduction that learn encoder ϕ and decoder θ parameters by maximizing the Evidence Lower Bound (ELBO) as defined in Equation 2. Typical VAE implementations utilize Gaussian distributions and amortized inference to approximate a complex posterior distribution q_ϕ. To improve this approximation various approaches such as Adversarial Variational Bayes (AVB) and Semi-Implicit Variational Inference (SIVI) have been proposed, leveraging adversarial training and semi-implicit hierarchical structures, respectively.

To evaluate our proposed Implicit Variational Rejection Sampling (IVRS) on the VAE inference problem, we conducted experiments on the MNIST dataset and compare with AVB and other methods. We trained our model on 60,000 training samples and evaluated performance on 10,000 test samples. The encoder is designed as a two-layer convolutional neural network (CNN) to map to the latent space, while the decoder consists of a four-layer transposed convolution network for image reconstruction. Figure 2 shows examples sampled from the test set and our model, demonstrating that our method generates images with a closer resemblance to the ground truth compared to the baseline AVB method, owing to the improved sampling quality introduced by rejection sampling.

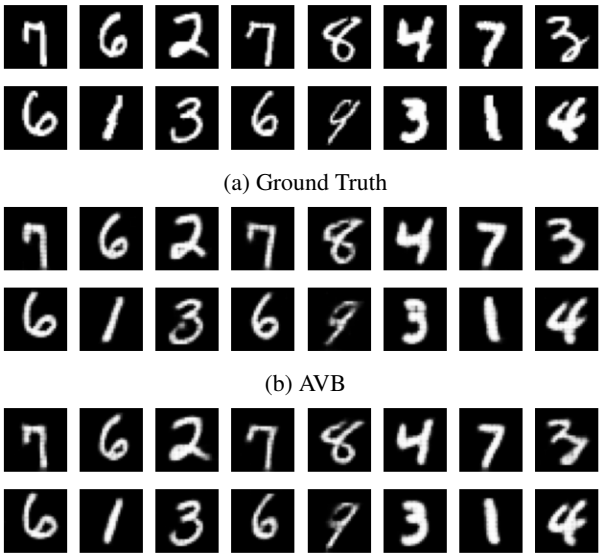
To evaluate the efficiency of our model, we conducted an empirical time analysis comparing IVRS to the baseline AVB on MNIST. The results are shown in Table 4. Using a batch size of 64 during training, we observed that our model incurs only a slight increase in computational cost, showing the advantage of our GPU-accelerated parallel sampling implementation.

We also quantitatively benchmarked our model against several well-established autoencoding methods, including Im-

	Boston		Concrete		Protein	
	NLL (↓)	RMSE (↓)	NLL (↓)	RMSE (↓)	NLL (↓)	RMSE (↓)
AVB	2.489 ± 0.02	2.685 ± 0.03	3.406 ± 0.01	7.091 ± 0.03	2.969 ± 0.05	4.670 ± 0.04
SIVI	2.481 ± 0.00	2.621 ± 0.02	3.337 ± 0.00	6.932 ± 0.02	2.967 ± 0.03	4.669 ± 0.02
UIVI	2.490 ± 0.02	2.617 ± 0.03	3.331 ± 0.01	6.806 ± 0.02	2.973 ± 0.03	4.671 ± 0.02
SIVI-SM	2.542 ± 0.01	2.785 ± 0.03	3.229 ± 0.01	5.973 ± 0.04	3.047 ± 0.00	5.087 ± 0.01
KSIVI	2.506 ± 0.01	2.555 ± 0.02	3.309 ± 0.01	5.750 ± 0.03	3.034 ± 0.04	5.027 ± 0.01
IVRS	$\mathbf{2.365 \pm 0.03}$	$\mathbf{2.421 \pm 0.03}$	$\mathbf{2.964 \pm 0.01}$	$\mathbf{5.68 \pm 0.04}$	$\mathbf{2.794 \pm 0.04}$	$\mathbf{4.601 \pm 0.03}$

	Power		Wine		Yacht	
	NLL (↓)	RMSE (↓)	NLL (↓)	RMSE (↓)	NLL (↓)	RMSE (↓)
AVB	2.795 ± 0.02	3.865 ± 0.02	0.905 ± 0.01	0.609 ± 0.00	1.751 ± 0.06	1.567 ± 0.04
SIVI	2.791 ± 0.00	3.861 ± 0.01	0.904 ± 0.00	0.597 ± 0.00	1.721 ± 0.03	1.505 ± 0.07
UIVI	2.794 ± 0.00	3.863 ± 0.02	0.907 ± 0.00	0.613 ± 0.00	1.808 ± 0.03	1.569 ± 0.05
SIVI-SM	2.822 ± 0.00	4.009 ± 0.00	0.916 ± 0.00	0.615 ± 0.00	1.432 ± 0.01	$\mathbf{0.884 \pm 0.01}$
KSIVI	2.797 ± 0.00	3.868 ± 0.01	0.901 ± 0.00	0.595 ± 0.00	1.752 ± 0.03	1.237 ± 0.05
IVRS	$\mathbf{2.670 \pm 0.01}$	$\mathbf{3.684 \pm 0.03}$	$\mathbf{0.900 \pm 0.00}$	$\mathbf{0.591 \pm 0.00}$	$\mathbf{1.421 \pm 0.03}$	1.065 ± 0.04

Table 2: Quantitative results for six UCI regression tasks. More accurate posterior sampling allows IVRS to outperform several other VI approximations for learning the same Bayesian neural network.



(a) Ground Truth

(b) AVB

(c) IVRS (ours)

Figure 2: Examples of 16 MNIST images from the test set using the VAE learned by AVB and IVRS. Our method generates images with a closer resemblance to the ground truth compared to the baseline due to the rejection sampling.

portance Weighted Autoencoders (IWAE) [Burda et al., 2015], Hamiltonian Variational Inference (HVI) [Salimans et al., 2015], and Normalizing Flows (NF) [Rezende and Mohamed, 2015]. Many approaches have also adopted deep architectures for the VAE encoder to achieve more effective feature extraction. Therefore, we also report comparisons with recent deep architecture-based VAE improvements such as NVAE [Vahdat and Kautz, 2020], CR-NVAE [Sinha and Dieng, 2021], and Efficient-VDVAE [Hazami et al., 2022].

As shown in Table 5, our vanilla IVRS method achieves a Negative Log Evidence (NLE) score of 81.78, surpassing traditional variational inference methods like SIVI and AVB. Additionally, we considered integrating IVRS within the NVAE architecture, which further boosts performance to an NLE of 77.36. This also represents an improvement over the original NVAE at 78.01. Although CR-NVAE [Sinha and Dieng, 2021] is the one approach to demonstrate slightly better results, that approach relies on additional data augmentation techniques. While data augmentation is highly effective in preventing overfitting, our focus is on demonstrating the competitiveness of rejection sampling in enhancing implicit variational inference, but we include those results for completeness. We further compare against the more recent DVP-VAE [Kuzina and Tomczak, 2024], a hierarchical VAE with a diffusion-based VampPrior, which reports an NLE of 77.10; this is directly comparable to our IVRS+NVAE result of 77.36, while IVRS additionally offers a general inference-refinement mechanism that is orthogonal to the choice of prior.

4.5 RESULTS ON CIFAR-10 AND IMAGENET

To further evaluate IVRS on higher-dimensional image data, we conducted VAE learning experiments on both the CIFAR-10 and ImageNet datasets and compare against baseline inference methods.

The CIFAR-10 dataset consists of 50,000 training images

Table 3: Impact of the rejection hyperparameter M on performance and efficiency. Relative cost is computed as the inverse of the acceptance rate. Lower NLL and RMSE indicate better performance.

Dataset	M	NLL (↓)	RMSE (↓)	Accept Rate	Rel. Cost
Boston	0.1	2.478	2.648	92.3%	1.08×
	1	2.412	2.502	83.7%	1.19×
	10	2.385	2.451	71.2%	1.40×
	100	2.365	2.421	54.8%	1.82×
	500	2.361	2.418	41.5%	2.41×
Concrete	0.1	3.285	6.845	88.5%	1.13×
	1	3.142	6.254	76.8%	1.30×
	10	3.021	5.892	63.4%	1.58×
	100	2.964	5.680	48.2%	2.07×
	500	2.951	5.652	35.6%	2.81×
Protein	0.1	2.928	4.658	94.1%	1.06×
	1	2.857	4.632	85.2%	1.17×
	10	2.812	4.614	73.8%	1.35×
	100	2.794	4.601	61.5%	1.63×
	500	2.789	4.598	48.7%	2.05×

Method	batch	iter	per epoch	per iter
AVB	64	938	19.11s	0.020s
IVRS (Ours)	64	938	24.72s	0.026s

Table 4: An empirical time analysis comparing the AVB method and our proposed IVRS on the MNIST dataset. GPU-accelerated parallelization is able to reduce the impact of additional computation required by our method.

and 10,000 test images, each consisting of 32×32 pixels and 3 color channels (RGB) across 10 classes. The ImageNet dataset [Deng et al., 2009], a large-scale dataset commonly used for image classification and generative modeling tasks, includes over 14 million images spanning 1,000 object categories. For our experiments, we utilized a subset of ImageNet consisting of 64×64 pixel color images, which are more manageable for generative models like VAEs. This subset introduces greater inference complexity due to the diversity of object categories and higher resolution compared to CIFAR-10. Given the increased complexity and higher dimensionality of these datasets compared to MNIST, we adapted the VAE architecture by directly employing the encoder from the deep architecture NVAE [Vahdat and Kautz, 2020].

For quantitative evaluation, we calculate bits per dimension (bpd), a standard metric for high-dimensional image datasets. Table 6 reports the bpd scores for various variational inference methods on CIFAR-10, while Table 7 presents the corresponding results on ImageNet. As can be seen, the proposed IVRS consistently achieves lower bpd values than a range of classic VAE-based approaches, indicating improved density estimation quality. Notably, while CR-NVAE [Sinha and Dieng, 2021] attains strong performance by leveraging additional data augmentation, our controlled experiments under the same augmentation protocol show that incorporating IVRS further improves performance. In particular, on CIFAR-10, CR-NVAE achieves 2.51 bpd with augmentation, and adding IVRS (with $M = 1$) reduces this to 2.43 bpd. This demonstrates that IVRS provides a non-trivial gain even on top of strong augmentation-based baselines, suggesting that the proposed inference refinement is orthogonal and complementary to data augmentation. For a more recent point of comparison, the DVP-VAE [Kuzina and Tomczak, 2024] reports 2.73 bpd on CIFAR-10, which is on par with our IVRS+NVAE result of 2.76 bpd, while IVRS remains complementary as an inference-refinement step. Moreover, despite the additional rejection sampling step, the overall computational overhead remains modest, indicating that IVRS is practical for large-scale VAE training scenarios.

5 CONCLUSION

We have introduced Implicit Variational Rejection Sampling, a novel posterior approximation method that enhances variational inference by integrating the flexibility of implicit distributions with the precision of rejection sampling. By optimizing the derived Implicit Resampled Evidence Lower Bound (IR-ELBO), IVRS provides a mechanism for enhancing posterior approximation over traditional VI methods. Our experimental results demonstrate the effectiveness of IVRS across various tasks, including regression problems and VAE-based image modeling, achieving superior quantitative performance over related inference approaches. The supplement below contains more practical details about implementation and limitations, along with an in depth description of related work and ablation study.

Methods	NLE
Results from Burda et al. [2015]	
VAE + IWAE	86.76
IWAE + IWAE	84.78
Results from Salimans et al. [2015]	
VAE + HVI (1 leapfrog step)	88.08
VAE + HVI (4 leapfrog steps)	86.40
VAE + HVI (8 leapfrog steps)	85.51
Results from Rezende and Mohamed [2015]	
VAE + NICE [Dinh et al., 2014] (k = 80)	87.2
VAE + NF (k = 40)	85.7
VAE + NF (k = 80)	85.1
Results from Gregor et al. [2015]	
NADE	88.33
DBM 2hl	84.62
DBN 2hl	84.55
EoNADE-5 2hl (128 orderings)	84.68
DARN 1hl	84.13
Results from Sønderby et al. [2016]	
Auxiliary VAE ($L = 1$, IW $= 1$)	84.59
Results from Mescheder et al. [2017]	
VAE + IAF Kingma et al. [2016]	84.9
AVB	83.7
Results from Yin and Zhou [2018]	
SIVI (3 stochastic layers) + IW($\tilde{K} = 10$)	83.25
Results from Hazami et al. [2022]	
PixelVAE++ [Sadeghi et al., 2019]	78.00
Local Mask PixelCNN [Jain et al., 2020]	77.58
NVAE [Vahdat and Kautz, 2020]	78.01
CR-NVAE [Sinha and Dieng, 2021]	**76.93***
Efficient-VDVAE [Hazami et al., 2022]	79.09
Results from Kuzina and Tomczak [2024]	
DVP-VAE [Kuzina and Tomczak, 2024]	77.10
IVRS	81.78
IVRS+NVAE [Vahdat and Kautz, 2020]	77.36

Table 5: Comparison of reported Negative Log Evidence (NLE) values across different algorithms on the MNIST dataset. An asterisk ($\star$) indicates results obtained with data augmentation. The **best**, second-best, and third-best results are highlighted.

Methods	bpd
VAE + IAF [Sønderby et al., 2016]	3.11
BIVA [Maaløe et al., 2019]	3.08
DVAE [Vahdat et al., 2018]	3.38
δ-VAE [Razavi et al., 2019]	2.83
PixelVAE++ [Sadeghi et al., 2019]	2.90
Local Mask PixelCNN [Jain et al., 2020]	2.89
MAE [Ma et al., 2019b]	2.95
NVAE [Vahdat and Kautz, 2020]	2.91
VDVAE [Child, 2021]	2.87
Efficient-VDVAE [Hazami et al., 2022]	2.87
DVP-VAE [Kuzina and Tomczak, 2024]	2.73
CR-NVAE [Sinha and Dieng, 2021]	2.51*
IVRS+NVAE	2.76
IVRS+CR-NVAE	**2.43***

Table 6: Comparison of reported Bits per Dimension (bpd) values among various algorithms for the CIFAR-10 dataset. An asterisk ($\star$) indicates results obtained using data augmentation. The **best**, second-best, and third-best results are highlighted.

Methods	bpd
SPN [Menick and Kalchbrenner, 2019]	3.52
MaCow [Ma et al., 2019a]	3.69
Sparse Transformer [Child et al., 2019]	3.44
Local Mask PixelCNN [Jain et al., 2020]	3.64
NVAE [Vahdat and Kautz, 2020]	3.58
VDVAE [Child, 2021]	3.52
CR-NVAE [Sinha and Dieng, 2021]	**3.30***
IVRS+NVAE	3.38

Table 7: Comparison of reported Bits per Dimension (bpd) values among various algorithms for the Imagenet 64 $\times$64 dataset. An asterisk ($\star$) indicates results obtained using data augmentation. The **best**, second-best, and third-best results are highlighted.

References

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International conference on machine learning*, pages 224–232. PMLR, 2017.

Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, and Augustus Odena. Discriminator rejection sampling. In *International Conference on Learning Representations*, 2018.

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 66–75. PMLR, 2019.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Wei Chen, Shigui Li, Jiacheng Li, Junmei Yang, John Paisley, and Delu Zeng. Dequantified diffusion-schrödinger bridge for density ratio estimation. In *International Conference on Machine Learning*, pages 8427–8452. PMLR, 2025.

Zhichao Chen, Haoxuan Li, Fangyikang Wang, Odin Zhang, Hu Xu, Xiaoyu Jiang, Zhihuan Song, and Hao Wang. Re-

thinking the diffusion models for missing data imputation: A gradient flow perspective. *Advances in Neural Information Processing Systems*, 37:112050–112103, 2024a.

Zhichao Chen, Hao Wang, Guofei Chen, Yiran Ma, Le Yao, Zhiqiang Ge, and Zhihuan Song. Analyzing and improving supervised nonlinear dynamical probabilistic latent variable model for inferential sensors. *IEEE Transactions on Industrial Informatics*, 20(11):13296–13307, 2024b.

Ziheng Cheng, Longlin Yu, Tianyu Xie, Shiyue Zhang, and Cheng Zhang. Kernel semi-implicit variational inference. In *International Conference on Machine Learning*, pages 8248–8269. PMLR, 2024.

Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=RLRXCV6DbEJ.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pages 1462–1471. PMLR, 2015.

Aditya Grover, Ramki Gummadi, Miguel Lazaro-Gredilla, Dale Schuurmans, and Stefano Ermon. Variational rejection sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 823–832. PMLR, 2018.

Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficientvdvae: Less is more. *arXiv preprint arXiv:2203.13751*, 2022.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Alexandra Hotti, Oskar Kviman, Ricky Molén, Víctor Elvira, and Jens Lagergren. Efficient mixture learning in black-box variational inference. In *Proceedings of the 41st International Conference on Machine Learning*, pages 18972–18991. PMLR, 2024.

Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

Ajay Jain, Pieter Abbeel, and Deepak Pathak. Locally masked convolution for autoregressive models. In *Conference on Uncertainty in Artificial Intelligence*, pages 1358–1367. PMLR, 2020.

Martin Jankowiak and Du Phan. Reparameterized variational rejection sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 739–747. PMLR, 2024.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37: 183–233, 1999.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in Neural Information Processing Systems*, 29, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Anna Kuzina and Jakub M. Tomczak. Hierarchical VAE with a diffusion-based vampprior. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=NUkEoZ7Toa.

Oskar Kviman, Ricky Molén, Alexandra Hotti, Semih Kurt, Vıctor Elvira, and Jens Lagergren. Cooperation in the latent space: The benefits of adding mixture components in variational autoencoders. In *International Conference on Machine Learning*, pages 18008–18022. PMLR, 2023.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Shigui Li, Wei Chen, and Delu Zeng. EVODiff: Entropy-aware variance optimized diffusion inference. *Advances in Neural Information Processing Systems*, 38:148134–148181, 2026.

Jen Ning Lim and Adam M Johansen. Particle semi-implicit variational inference. *Advances in Neural Information Processing Systems*, 37:123954–123990, 2024.

Xuezhe Ma, Xiang Kong, Shanghang Zhang, and Eduard Hovy. Macow: Masked convolutional generative flow. *Advances in Neural Information Processing Systems*, 32, 2019a.

Xuezhe Ma, Chunting Zhou, and Eduard Hovy. MAE: Mutual posterior-divergence regularization for variational autoencoders. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=Hke4l2AcKQ.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems*, 32, 2019.

Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HylzTiC5Km.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, pages 2391–2400. PMLR, 2017.

Vincent Moens, Hang Ren, Alexandre Maraval, Rasul Tutunov, Jun Wang, and Haitham Ammar. Efficient semi-implicit variational inference. *arXiv preprint arXiv:2101.06070*, 2021.

Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit variational inference. In *International Conference on Artificial Intelligence and Statistics*, pages 2593–2602. PMLR, 2019.

Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498. PMLR, 2017.

Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-VAEs. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJe0Gn0cY7.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

Hossein Sadeghi, Evgeny Andriyash, Walter Vinci, Lorenzo Buffoni, and Mohammad H Amin. Pixelvae++: Improved pixelvae with discrete prior. *arXiv preprint arXiv:1908.09948*, 2019.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226. PMLR, 2015.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference. In *International Conference on Learning Representations*, 2018.

Samarth Sinha and Adji Bousso Dieng. Consistency regularization for variational auto-encoders. *Advances in Neural Information Processing Systems*, 34:12943–12954, 2021.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in Neural Information Processing Systems*, 29, 2016.

Vincent Stimper, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Resampling base distributions of normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pages 4915–4936. PMLR, 2022.

Michalis K Titsias and Francisco Ruiz. Unbiased implicit variational inference. In *International Conference on Artificial Intelligence and Statistics*, pages 167–176. PMLR, 2019.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Arash Vahdat, William Macready, Zhengbing Bian, Amir Khoshaman, and Evgeny Andriyash. Dvae++: Discrete variational autoencoders with overlapping transformations. In *International conference on machine learning*, pages 5035–5044. PMLR, 2018.

Alexandre Verine, Muni Sreenivas Pydi, Benjamin Negrevergne, and Yann Chevaleyre. Optimal budgeted rejection sampling for generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 3367–3375. PMLR, 2024.

Jian Xu and Delu Zeng. Sparse variational student-t processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16156–16163, 2024.

Jian Xu, Shian Du, Junmei Yang, Qianli Ma, and Delu Zeng. Neural operator variational inference based on regularized stein discrepancy for deep gaussian processes. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6723–6737, 2024a.

Jian Xu, Zhiqi Lin, Shigui Li, Min Chen, Junmei Yang, Delu Zeng, and John Paisley. Flexible bayesian last layer models using implicit priors and diffusion posterior sampling. *arXiv preprint arXiv:2408.03746*, 2024b.

Jian Xu, Delu Zeng, and John Paisley. Sparse inducing points in deep gaussian processes: Enhancing modeling with denoising diffusion variational inference. In *International Conference on Machine Learning*, pages 55490–55500. PMLR, 2024c.

Jian Xu, Shian Du, Junmei Yang, Xinghao Ding, Delu Zeng, and John Paisley. Bayesian gaussian process ODEs via double normalizing flows. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025a.

Jian Xu, Shian Du, Junmei Yang, Qianli Ma, Delu Zeng, and John Paisley. Variational learning of gaussian process latent variable models through stochastic gradient annealed importance sampling. In *Conference on Uncertainty in Artificial Intelligence*, pages 4663–4680. PMLR, 2025b.

Jian Xu, Zhiqi Lin, Min Chen, Junmei Yang, Delu Zeng, and John Paisley. Fully bayesian differential gaussian processes through stochastic differential equations. *Knowledge-Based Systems*, 314:113187, 2025c.

Jian Xu, Delu Zeng, and John Paisley. Sparse variational student-t processes for heavy-tailed modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 2026a.

Jian Xu, Delu Zeng, Qibin Zhao, and John Paisley. Diffusion bridge variational inference for deep gaussian processes. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL `https://openreview.net/forum?id=zyRmy0Ch9a`.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *International Conference on Machine Learning*, pages 5660–5669. PMLR, 2018.

Longlin Yu and Cheng Zhang. Semi-implicit variational inference via score matching. In *The Eleventh International Conference on Learning Representations*, 2023.

A PROOF OF THE MONOTONICITY OF THE KL DIVERGENCE IN M

Here we give a rigorous proof of Proposition 1: in the idealized setting with an optimal discriminator, the divergence $\mathrm{KL}(r_{\theta,\phi}(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x}))$ is monotonically non-increasing in M. We first record an elementary covariance inequality.

Lemma 1. *Let X be a real-valued random variable and let f be a monotone non-decreasing function such that the expectations below exist. Then* $\mathrm{Cov}(X, f(X)) \geq 0$.

Proof. Let X' be an independent copy of X. Then

$$\mathrm{Cov}(X, f(X)) = \tfrac{1}{2}\,\mathbb{E}\big[(X - X')\big(f(X) - f(X')\big)\big]\,.$$

Since f is monotone non-decreasing, $(X - X')\big(f(X) - f(X')\big) \geq 0$ for every pair (X, X'). Taking expectations gives $\mathrm{Cov}(X, f(X)) \geq 0$. □

Proof of Proposition 1. Write $\pi(\mathbf{z}) = p_\theta(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})/p_\theta(\mathbf{x})$ for the true posterior, and let the acceptance probability in Equation (9) be

$$a_M(\mathbf{z}) = \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z}) + Mq_\phi(\mathbf{z}|\mathbf{x})}.$$

Using $p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = p_\theta(\mathbf{x})\pi(\mathbf{z})$, this can be rewritten as

$$a_M(\mathbf{z}) = \frac{p_\theta(\mathbf{x})\pi(\mathbf{z})}{p_\theta(\mathbf{x})\pi(\mathbf{z}) + Mq_\phi(\mathbf{z}|\mathbf{x})}.$$

Defining the density ratio and rescaled parameter

$$w(\mathbf{z}) = \frac{q_\phi(\mathbf{z}|\mathbf{x})}{\pi(\mathbf{z})}, \qquad \tilde{M} = \frac{M}{p_\theta(\mathbf{x})},$$

gives $a_M(\mathbf{z}) = 1/(1 + \tilde{M}w(\mathbf{z}))$. Since $p_\theta(\mathbf{x})$ is a positive constant independent of $\mathbf{z}$, we may absorb it into the scalar parameter and equivalently write $a_M(\mathbf{z}) = 1/(1 + Mw(\mathbf{z}))$.

The induced resampled distribution is

$$r_M(\mathbf{z}) = \frac{q_\phi(\mathbf{z}|\mathbf{x})\, a_M(\mathbf{z})}{Z_M}, \qquad Z_M = \mathbb{E}_{q_\phi}[a_M(\mathbf{z})].$$

Using $q_\phi(\mathbf{z}|\mathbf{x}) = w(\mathbf{z})\pi(\mathbf{z})$, we obtain

$$r_M(\mathbf{z}) = \frac{\pi(\mathbf{z})\, c_M(\mathbf{z})}{\tilde{Z}_M}, \quad c_M(\mathbf{z}) = \frac{w(\mathbf{z})}{1 + Mw(\mathbf{z})}, \quad \tilde{Z}_M = \mathbb{E}_\pi[c_M(\mathbf{z})].$$

Hence the divergence to the true posterior is

$$\mathrm{KL}(r_M \parallel \pi) = \frac{1}{\tilde{Z}_M}\mathbb{E}_\pi[c_M \log c_M] - \log \tilde{Z}_M.$$

Differentiating with respect to M and using

$$\frac{dc_M}{dM} = -c_M^2, \quad \frac{d\tilde{Z}_M}{dM} = -\mathbb{E}_\pi[c_M^2], \quad \frac{d(c_M \log c_M)}{dM} = -c_M^2(1 + \log c_M),$$

the two terms involving $\mathbb{E}_\pi[c_M^2]/\tilde{Z}_M$ cancel, leaving

$$\frac{d}{dM}\mathrm{KL}(r_M \parallel \pi) = -\frac{1}{\tilde{Z}_M}\mathbb{E}_\pi[c_M^2 \log c_M] + \frac{\mathbb{E}_\pi[c_M^2]}{\tilde{Z}_M^2}\mathbb{E}_\pi[c_M \log c_M].$$

Rewriting the expectations under r_M via $\mathbb{E}_{r_M}[g] = \mathbb{E}_\pi[c_M g]/\tilde{Z}_M$ yields

$$\frac{d}{dM}\mathrm{KL}(r_M \parallel \pi) = -\mathbb{E}_{r_M}[c_M \log c_M] + \mathbb{E}_{r_M}[c_M]\,\mathbb{E}_{r_M}[\log c_M] = -\mathrm{Cov}_{r_M}(c_M, \log c_M).$$

Applying Lemma 1 with $X = c_M$ and the monotone non-decreasing function $f(X) = \log X$ gives $\mathrm{Cov}_{r_M}(c_M, \log c_M) \geq 0$. Therefore

$$\frac{d}{dM}\mathrm{KL}(r_M \parallel p_\theta(\mathbf{z}|\mathbf{x})) \leq 0,$$

so the divergence is monotonically non-increasing in M in the idealized exact-ratio setting. When the discriminator is imperfect, the acceptance probability and the induced resampled distribution are only approximate, so this monotonicity should be interpreted as an idealized analytical property rather than a strict guarantee. □

B RELATED WORK

The main idea of implicit variational inference is to transform a simple base distribution into a more expressive one using a deep neural network [Mescheder et al., 2017, Huszár, 2017, Titsias and Ruiz, 2019, Shi et al., 2018]. To avoid density ratio estimation, semi-implicit variational inference has been proposed, where the variational distributions are formed through a semi-implicit hierarchical construction, and surrogate ELBOs (asymptotically unbiased) are employed for training [Yin and Zhou, 2018, Molchanov et al., 2019, Moens et al., 2021, Lim and Johansen, 2024, Yu and Zhang, 2023, Cheng et al., 2024]. These methods have made improvements to implicit variational inference or semi-implicit variational inference at various levels. However, these methods do not address the limited expressiveness of neural networks, which may still fall short in certain scenarios. Our work aims to address this issue by proposing an implicit rejection sampling algorithm.

Beyond standard latent-variable models, implicit and diffusion-based variational inference has been increasingly applied to richer Bayesian models, including deep Gaussian processes and their dynamical extensions [Xu et al., 2024c, 2026b, 2024a, 2025b,a,c], Bayesian last-layer models [Xu et al., 2024b], heavy-tailed Student-t processes [Xu and Zeng, 2024, Xu et al., 2026a], and nonlinear probabilistic latent-variable models for industrial sensing [Chen et al., 2024b]. In parallel, diffusion-based generative models have motivated new tools for posterior sampling and density-ratio estimation [Chen et al., 2025, Li et al., 2026, Chen et al., 2024a], which are complementary to the discriminator-based density-ratio estimation used in IVRS.

On the other hand, rejection sampling is a classical method to generate samples from a distribution using samples drawn from a different distribution [Gilks and Wild, 1992, Grover et al., 2018, Azadi et al., 2018, Stimper et al., 2022, Verine et al., 2024]. Specifically, Grover et al. [2018] and Stimper et al. [2022] have embedded latent rejection sampling within their training processes, applying it within a variational inference and a normalizing flow framework, respectively. We acknowledge their contributions; however, in this work, we address a different problem—leveraging implicit distributions as proposal distributions for rejection sampling and variational inference, creating a novel variational inference algorithm.

The prior in Bauer and Mnih [2019] is reformulated as a resampled distribution, whereas our method explicitly derives a new evidence lower bound (IR-ELBO) based on rejection sampling. While both approaches use density ratio estimation, our work focuses on leveraging rejection sampling to construct a tighter variational objective by directly approximating the posterior through implicit distributions.

Equation (2) in Jankowiak and Phan [2024] and our Equation (13) share a similar mathematical form, but a key difference lies in the nature of the variational distribution: in Jankowiak and Phan [2024], the distribution is explicit, while in our work, it is implicit. This distinction is significant as it aligns with our objective of improving implicit variational inference by utilizing rejection sampling to create more flexible posterior approximations.

Our method can also be extended to importance sampling Burda et al. [2015], and could be adapted for use with mixtures of variational distributions. Using mixtures could enhance the expressivity of the approximate posterior and align well with the results from Hotti et al. [2024] and Kviman et al. [2023]. Combining the importance-weighted ELBO (IW-ELBO) Burda et al. [2015] with Eq. (16) could provide a tighter bound. This would involve integrating importance weighting within the rejection sampling framework, potentially amplifying the advantages of both techniques. This presents an exciting direction for future work.

C ABLATION STUDY

In this section, we provide additional ablation experiments that complement the sensitivity analysis of the rejection hyperparameter M in the main text. We focus on three representative UCI Bayesian neural network regression benchmarks: *Boston Housing*, *Concrete*, and *Protein*. These datasets allow us to systematically examine the trade-off between posterior accuracy and computational efficiency induced by rejection sampling.

Table 8: Ablation study with fixed neural network architecture. All models use a 4-layer MLP; [k] denotes hidden width per layer. Lower NLL and RMSE indicate better performance.

Dataset	Method	NLL (↓)	RMSE (↓)	Architecture
Boston	Implicit VI	2.489	2.685	4-layer MLP [50]
	+ Rejection Sampling (IVRS)	**2.365**	**2.421**	Same
	+ 2× Network Capacity	2.461	2.638	4-layer MLP [100]
	+ 3× Network Capacity	2.445	2.612	4-layer MLP [150]
Concrete	Implicit VI	3.406	7.091	4-layer MLP [50]
	+ Rejection Sampling (IVRS)	**2.964**	**5.680**	Same
	+ 2× Network Capacity	3.318	6.824	4-layer MLP [100]
	+ 3× Network Capacity	3.275	6.705	4-layer MLP [150]
Protein	Implicit VI	2.969	4.670	4-layer MLP [50]
	+ Rejection Sampling (IVRS)	**2.794**	**4.601**	Same
	+ 2× Network Capacity	2.928	4.645	4-layer MLP [100]
	+ 3× Network Capacity	2.905	4.628	4-layer MLP [150]

C.1 FIXED ARCHITECTURE COMPARISON

To isolate the contribution of rejection sampling from increased network capacity, we conduct controlled experiments where all methods share identical neural network architectures. This allows us to explicitly disentangle the effect of posterior refinement via rejection sampling from that of simply adding more model parameters.

Specifically, we compare the following settings:

1. Baseline implicit variational inference (Implicit VI) with a fixed 4-layer MLP.
2. IVRS applied to the same architecture (no increase in network capacity).
3. Implicit VI with increased network capacity (2× and 3× hidden width).

All models are trained and evaluated under identical optimization settings.

Table 8 reports the results on three representative UCI BNN regression datasets. Several important observations emerge:

- **Rejection sampling outperforms capacity increase.** Applying IVRS with the same base architecture consistently yields larger improvements in NLL and RMSE than doubling or tripling the network width. For example, on the Boston dataset, IVRS achieves a 4.9% NLL improvement over the baseline, whereas doubling network capacity yields only a 1.1% improvement.
- **Capacity alone cannot replicate IVRS gains.** Even with $3\times$ network capacity, implicit VI underperforms IVRS using the original architecture across all datasets. This indicates that the performance gains of IVRS do not stem from increased representational capacity.
- **Complementary mechanisms.** These results suggest that rejection sampling improves posterior approximation in a manner fundamentally different from architectural scaling. IVRS refines the variational distribution through sample-level selection rather than function-level expressiveness.

Overall, this ablation demonstrates that IVRS provides a principled and parameter-efficient alternative to increasing network capacity, reinforcing the claim that rejection sampling contributes meaningfully beyond standard neural network design choices.

C.2 COMPLETE HYPERPARAMETER SETTINGS FOR M

For completeness and reproducibility, we summarize the values of the rejection hyperparameter M used across all experiments.

Toy Density Estimation (Table 1). For low-dimensional toy distributions, we set:

- 1D distributions (Gaussian, Laplace, GMM): $M = 0.1$
- 2D distributions (Banana, X-shape, GMM): $M = 500$

Higher-dimensional targets require larger M to sufficiently refine the proposal distribution, consistent with the analysis in Section 3.4.

UCI Regression (BNN Experiments, Table 2). For Bayesian neural network regression, we selected M via cross-validation on held-out validation sets:

- Boston, Concrete, Protein, Power: $M = 100$
- Yacht, Wine: $M = 10$

These BNN posteriors are relatively low-dimensional (on the order of hundreds of parameters). At $M = 100$, acceptance rates remain practical (approximately 50–60%), yielding substantial accuracy improvements (4–13% NLL reduction) with moderate computational overhead (1.6–2.1×).

VAE Experiments. For large-scale VAE training, we fix $M = 1$ for all datasets:

- MNIST
- CIFAR-10
- ImageNet 64×64

Unlike BNN tasks, VAE training involves significantly larger datasets (e.g., 60K for MNIST, 50K for CIFAR-10, and over 1M samples for ImageNet). To maintain reasonable training time, we use a small M, which yields high acceptance rates (approximately 80%) while still demonstrating the benefit of rejection sampling. These experiments primarily serve to validate the general applicability of IVRS across different model classes.

C.3 TRAINING AND INFERENCE WITH REJECTION SAMPLING

We clarify the role of rejection sampling during both training and inference in IVRS, as this point may be a source of potential confusion.

Training Phase. During training, rejection sampling is an integral component of IVRS. The variational distribution is defined as the resampled distribution $r_{\theta,\phi}(z \mid x)$, and all expectations in the proposed IR-ELBO are taken with respect to this distribution. Consequently, rejection sampling is explicitly applied during training to generate samples from $r_{\theta,\phi}(z \mid x)$, as described in Algorithm 1 and Algorithm 2.

Inference Phase. At test time, posterior samples are also drawn from the same resampled variational distribution $r_{\theta,\phi}(z \mid x)$ using the identical rejection sampling procedure. This ensures consistency between training and inference, and allows the improved posterior approximation obtained by IVRS to directly translate into better predictive performance.

Practical Considerations. In practice, the acceptance rate is controlled by the hyperparameter M and remains sufficiently high in all reported experiments (see Appendix C.2). For large-scale settings such as VAE training, we adopt a smaller M to balance computational efficiency and posterior refinement. Importantly, no additional approximation is introduced at test time beyond the rejection sampling mechanism already used during training.

C.4 SENSITIVITY TO DISCRIMINATOR OPTIMIZATION

The quality of the discriminator is a critical part of the method. In our framework the discriminator is not a peripheral component: it is used to approximate the density-ratio quantity that enters the acceptance probability, the induced resampled posterior, and the practical training objective. If the discriminator is inaccurate, the resulting error does not remain local but propagates through the entire procedure. The idealized analysis of Proposition 1 should therefore be understood as relying on a sufficiently accurate—in the strongest case optimal—discriminator. In practical optimization, however, the discriminator is only approximately trained, so the acceptance function, the resampled posterior, and the optimized objective

are all approximate; we therefore do not claim a general theoretical guarantee that approximation errors in the discriminator cannot degrade the posterior approximation. What is optimized in practice is better viewed as a discriminator-parameterized surrogate objective, which is also why Algorithm 2 updates the discriminator and the variational model in alternation rather than fully solving the discriminator subproblem at each outer step.

Our experiments show that the method is sensitive to discriminator optimization, but not in the trivial sense that stronger discriminator training always helps—overly aggressive updates can degrade both acceptance behavior and downstream posterior quality. On the Concrete dataset, under a fixed evaluation protocol with 100 accepted posterior samples, we observe the behavior in Table 9: a moderate setting is stable, whereas making the discriminator much stronger (more inner steps) or updating it too frequently drives the acceptance rate toward zero and substantially worsens RMSE/NLL. This indicates that the main failure mode is not underfitting of the discriminator, but *imbalance* in the alternating optimization between the discriminator and the variational model; the sensitivity is primarily a training-dynamics issue rather than a purely approximation-theoretic one.

Discriminator setting	RMSE	NLL
moderate (`every=5, steps=2`)	5.6974	3.0267
strong (`steps=25`)	11.3274	3.9247
frequent (`every=1`)	8.8113	3.6260

Table 9: Effect of discriminator optimization on IVRS (Concrete, 100 accepted samples). The "strong" setting drives the acceptance rate down to approximately 0.0002.

C.5 EFFECT OF LATENT DIMENSIONALITY

Although our approach uses a smooth rejection-inspired mechanism rather than classical exact rejection sampling, it still inherits a related scalability issue: as the effective latent dimensionality grows, acceptance behavior can deteriorate and tuning becomes more delicate, so the method is not immune to a curse-of-dimensionality effect. On MNIST, varying the latent dimensionality from 16 to 392 under a fixed evaluation protocol gives the results in Table 10: the best NLE is achieved in a moderate range (around 32–64 dimensions), while larger latent spaces gradually reduce the calibrated acceptance rate and worsen NLE. Once the latent space becomes too large, the proposal is harder to refine effectively by rejection-style resampling. Overall, IVRS is most suitable when one has a moderately expressive proposal, manageable latent dimensionality, and sufficient room for posterior refinement through selective resampling.

Latent dim	16	32	64	128	256	392
NLE	83.65	81.48	81.58	82.10	82.72	82.83
Accept. rate	0.1563	0.1319	0.1295	0.1255	0.1177	0.1127

Table 10: Effect of latent dimensionality on IVRS (MNIST). "Accept. rate" is the calibrated acceptance rate.

C.6 DECOMPOSING THE SOURCE OF THE GAINS

To understand where the improvements come from, we perform controlled decomposition ablations on the Concrete dataset using the same backbone, training protocol, and evaluation budget, while isolating four cases: `baseline` (implicit train + implicit test), `reweight_only` (IVRS-style train + implicit test), `selection_only` (implicit train + IVRS-style test), and `full_ivrs` (IVRS-style train + IVRS-style test). Table 11 reports RMSE/NLL under two discriminator settings (M=1 and M=0.1, both with `disc-update-every=5`, `disc-steps=1`). The gain cannot be attributed to density-ratio reweighting alone, and "implicit regularization" from the training objective by itself is not consistently beneficial; rather, the improvement comes from the *interaction* between reweighting and selection rather than from either component in isolation.

D LIMITATIONS AND ADDITIONAL DISCUSSION

Despite the clear advantages of IVRS, which combines the accuracy of rejection sampling with the flexibility of implicit distributions, there are some limitations. One key limitation is the need for empirical hand-tuning of the hyperparameter M

Configuration	M=1 (RMSE/NLL)	M=0.1 (RMSE/NLL)
`baseline`	7.068 / 3.4322	7.068 / 3.4322
`reweight_only`	7.0230 / 3.4567	9.2461 / 3.3719
`selection_only`	9.4048 / 3.5626	8.0144 / 3.6206
`full_ivrs`	**6.2992 / 3.1451**	**6.7390 / 3.2773**

Table 11: Controlled decomposition ablation on Concrete. Only the full combination of reweighting (train) and selection (test) consistently improves over the baseline.

in the acceptance probability function. Additionally, while our experiments have shown improved bpd scores on standard datasets like CIFAR-10, further testing on more complex and higher-dimensional datasets is necessary to fully assess the robustness of our approach.

D.1 ROBUSTNESS OF THE METHOD IN HIGH-DIMENSIONAL SETTINGS

We acknowledge the challenges in high-dimensional scenarios due to the curse of dimensionality. To mitigate this, our approach leverages implicit proposal distributions parameterized by neural networks, which are designed to closely approximate the target distribution. This reduces the rejection rate and improves efficiency even in higher dimensions.

In our experiments, we have evaluated the proposed method on datasets with moderately high-dimensional posterior distributions (e.g., several hundred dimensions). Results demonstrate that the acceptance rates remain manageable, and the method performs favorably compared to baseline variational inference (VI) approaches. We recognize, however, that testing on more extreme high-dimensional cases (e.g., thousands of dimensions) could provide additional insights. This is a valuable direction for future work to implement more scalable architectures to further explore this aspect.

D.2 COMPUTATIONAL EFFICIENCY COMPARED TO OTHER METHODS

Computational overhead is a critical factor in assessing the practicality of our approach. The addition of the discriminator network introduces some computational cost, particularly for estimating the density ratio and acceptance probability. To address this issue:

- **Comparison with Baselines:** In our experiments, we observed that the computational cost of training the discriminator is offset by the reduction in bias due to tighter approximation of the Evidence Lower Bound (ELBO). Compared to standard VI methods, our approach exhibits a trade-off where the marginal improvement in accuracy justifies the additional computation.
- **Efficiency in Large-Scale Settings:** To improve scalability, we used a lightweight architecture for the discriminator and optimized its training through mini-batch techniques. However, we acknowledge that in large-scale datasets or extremely high-dimensional problems, further optimization (e.g., parallelization or approximate methods) may be necessary.
- **Maximum Iteration Limit:** We introduce a maximum iteration limit for the rejection sampling process to prevent it from getting stuck in an infinite loop

D.3 COST AND DIFFICULTY OF TRAINING THE DISCRIMINATOR (EQ. 10)

In our framework, the discriminator approximates the density ratio between the proposal distribution $q_\phi(z|x)$ and the true posterior $p(z|x)$, which determines the acceptance probability in the rejection sampling step. It is parameterized using the same architecture as in AVB and trained with a standard binary cross-entropy loss. As a result, the computational cost remains comparable to AVB, and no additional model complexity is introduced.

The discriminator directly influences the sampling process by defining the acceptance probability $a(z; x, \theta, \phi)$, making its accuracy crucial to overall performance. However, we found that full convergence of the discriminator at each update step is not necessary in practice. Instead, we adopt an alternating optimization strategy, where the discriminator is updated for a few epochs (typically 5–10) before updating the variational parameters, similar to the alternating training scheme used in Generative Adversarial Networks (GANs).

D.4 MANUAL TUNING OF THE HYPERPARAMETER M

The hyperparameter M controls the acceptance threshold in rejection sampling and directly governs the trade-off between posterior accuracy and computational efficiency. In our experiments, M is selected via cross-validation on held-out validation sets.

As demonstrated in the ablation studies in Appendix C, increasing M consistently improves posterior approximation quality, as reflected by lower NLL and RMSE, while simultaneously reducing the acceptance rate and increasing computational cost. This behavior aligns with the theoretical analysis in Section 3.4, which shows that the KL divergence between the resampled variational distribution and the true posterior decreases monotonically with M.

Importantly, we find that IVRS is not overly sensitive to the exact choice of M within a reasonable range. Moderate values of M already provide substantial accuracy gains over the implicit proposal distribution, while maintaining practical acceptance rates. This allows M to be chosen using lightweight cross-validation without extensive tuning.

While automatic or adaptive selection of M is an interesting direction for future work, our results indicate that the current strategy provides a reliable and interpretable balance between accuracy and efficiency across a wide range of tasks.