Denoising Implicit Variational Inference

Zitong Wang* **Zekun Wu*** **Longlin Yu** **Cheng Zhang†**

School of Mathematical Sciences
Peking University

Abstract

Semi-implicit variational inference (SIVI) expands the representational capacity of variational families, but optimizing the Evidence Lower Bound (ELBO) remains challenging because the marginal score of the semi-implicit distribution is generally intractable. Unbiased Implicit Variational Inference (UIVI) and related methods address this difficulty through reparameterized ELBO gradients, but existing approaches estimate the required score indirectly through reverse-conditional sampling or approximation, which can be computationally costly and unstable. We propose Denoising Implicit Variational Inference (DIVI), which learns the marginal score directly via denoising score matching. The learned score is used as a plug-in estimator in the pathwise ELBO gradient, replacing inner reverse-conditional sampling with a score-network evaluation. We analyze DIVI as inexact stochastic gradient ascent and show, under stated assumptions, that the averaged stationarity measure is controlled by the usual optimization term and the average score-matching error. Empirically, we evaluate DIVI on both synthetic distributions and a variety of real-data Bayesian inference tasks. The results show that DIVI improves over the current UIVI baselines while reducing the cost of ELBO-gradient estimation. Code is available at `https://github.com/VincentWangzt/DIVI`.

1 Introduction

Bayesian inference provides a principled framework for uncertainty quantification, but exact posterior inference is often intractable in modern probabilistic models. Variational Inference (VI) addresses this difficulty by casting posterior approximation as an optimization problem: given a variational family $\mathcal{Q}$, VI seeks a tractable approximation $q \in \mathcal{Q}$ by maximizing the Evidence Lower Bound (ELBO), equivalently minimizing a KL divergence up to an additive constant [Jordan et al., 1999, Zhang et al., 2018] . The choice of $\mathcal{Q}$ determines the tradeoff between computational tractability and approximation quality. Classical mean-field VI is computationally convenient due to its factorized structure, but is often too restrictive to capture posterior dependencies and complex geometry [Peterson and Anderson, 1987, Saul et al., 1996, Turner et al., 2011, Blei et al., 2017].

Subsequent work has relaxed these restrictions through black-box gradient estimators and richer variational families, including auxiliary-variable models, structured approximations, and normalizing flows, substantially improving the expressiveness of VI [Ranganath et al., 2014, Mnih and Gregor, 2014, Paisley et al., 2012, Giordano et al., 2015, Kingma et al., 2016, Dinh et al., 2016]. However, many such methods still require evaluating and differentiating the variational density, which restricts the admissible variational family.

Implicit VI relaxes this restriction by defining variational distributions through sampling procedures, such as neural generators, without requiring tractable densities [Tran et al., 2017, Huszár, 2017, Shi et al., 2018]. The central obstacle then shifts to ELBO optimization, since marginal log-density

*Equal contribution.

†Correspondence to Cheng Zhang <chengzhang@math.pku.edu.cn>.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and marginal score are no longer directly available. Semi-implicit variational inference (SIVI) [Yin and Zhou, 2018, Moens et al., 2021] offers a compromise by combining an explicit conditional distribution with an implicit mixing distribution. The resulting variational family is easy to sample from and highly expressive, but exact ELBO gradients require the marginal score $\nabla_z \log q_\phi(z)$, which is intractable because it involves integrating over the mixing variable.

Existing ELBO-based SIVI methods address this marginal-score bottleneck indirectly. Classical SIVI optimizes biased surrogate ELBOs, whereas unbiased implicit variational inference (UIVI) [Titsias and Ruiz, 2019] derives an unbiased gradient estimator of the exact ELBO assuming access to samples from the reverse conditional distribution over the mixing variable. In practice, this reverse conditional is approximated by inner-loop Markov chain Monte Carlo (MCMC) sampling, which can be expensive and unstable when the reverse conditional is high-dimensional or multimodal. Importance-weighted variants reduce the reliance on MCMC by using importance sampling from proposal distributions [Sobolev and Vetrov, 2019], and recent AISIVI-style methods improve the proposal through conditional normalizing flows (CNFs) [Pielok et al., 2025]. However, these approaches still depend on an inner approximation to the reverse conditional, resulting in difficulties to control the variance of the gradient estimator in practice. The sampling-based approximation techniques also add computational overhead to the optimization procedure.

In this paper, we propose Denoising Implicit Variational Inference (DIVI), which estimates the marginal score of a semi-implicit variational distribution directly through denoising score matching (DSM) [Vincent, 2011]. The key observation is that samples from the semi-implicit hierarchy provide pairs (z, ϵ) for which the conditional score $\nabla_z \log q_\phi(z|\epsilon)$ is tractable. Instead of sampling from an approximate reverse conditional, DIVI trains a score network to regress the marginal score from these conditional-score targets. The learned score is then used as a plug-in estimator in the pathwise ELBO gradient, replacing inner proposal-based approximation with a single score-network evaluation.

This formulation yields both an efficient algorithm and a tractable optimization analysis. Our main contributions are summarized as follows:

- **Direct marginal-score learning for semi-implicit VI.** We propose DIVI, which learns the marginal score of a semi-implicit variational distribution via denoising score matching and uses the learned score as a plug-in estimator in the pathwise ELBO gradient, replacing inner reverse-conditional sampling with a single score-network evaluation.
- **Convergence analysis with inexact score estimates.** We analyze DIVI through the lens of inexact stochastic gradient ascent and show, under standard smoothness and boundedness assumptions, that the averaged stationarity measure is controlled by the usual optimization error and the average score-matching error.
- **Empirical validation on synthetic distribution and real-data Bayesian inference tasks.** Across these settings, DIVI recovers challenging posterior structure, matches or improves upon current baselines, and achieves favorable wall-clock efficiency in high-dimensional diffusion and Bayesian neural-network experiments.

2 Background

2.1 Semi-implicit Variational Family

SIVI [Yin and Zhou, 2018] posits a flexible family that mixes an explicit tractable conditional layer $q_\phi(z|\epsilon)$ with a simple mixing distribution $q(\epsilon)$:

$$\epsilon \sim q(\epsilon), \qquad z \sim q_\phi(z|\epsilon), \qquad q_\phi(z) = \int q_\phi(z|\epsilon)\, q(\epsilon)\, \mathrm{d}\epsilon.$$

This construction makes $q_\phi(z)$ semi-implicit: sampling is straightforward via drawing from $q(\epsilon)$ followed by $q_\phi(z|\epsilon)$, while evaluating the marginal density $q_\phi(z)$ is generally intractable due to the integral. This intractability of $\log q_\phi(z)$ makes optimizing the exact ELBO challenging, and SIVI employs tractable surrogate objectives that become tight asymptotically. We write $s_\phi(z) := \nabla_z \log q_\phi(z)$ for the marginal score and $s_\phi(z|\epsilon) := \nabla_z \log q_\phi(z|\epsilon)$ for the conditional score.

2.2 Unbiased Implicit Variational Inference

Instead of optimizing surrogate ELBOs for semi-implicit variational families, UIVI targets the exact ELBO and derives an unbiased gradient estimator [Titsias and Ruiz, 2019]. We follow the setup in UIVI, adopting the mild assumption of reparameterizability: there exists a base noise distribution $q(u)$ and a deterministic map h_ϕ such that

$$u \sim q(u), \qquad z = h_\phi(u;\epsilon) \quad \Longleftrightarrow \quad z \sim q_\phi(z|\epsilon). \tag{1}$$

Denote the dataset as $\mathcal{D}$. With the reparameterization $z = h_\phi(u;\epsilon)$, the exact ELBO can be rewritten as

$$\begin{aligned}\mathcal{L}(\phi) &= \mathbb{E}_{q_\phi(z)}\big[\log p(\mathcal{D},z) - \log q_\phi(z)\big] \\ &= \mathbb{E}_{q(\epsilon)\,q(u)}\Big[\log p(\mathcal{D},z) - \log q_\phi(z)\Big]\Big|_{z=h_\phi(u;\epsilon)}.\end{aligned} \tag{2}$$

Pushing ∇_ϕ inside the expectation yields the following decomposition, where the term $\mathbb{E}_{z\sim q_\phi}\nabla_\phi \log q_\phi(z)$ vanishes under the usual regularity conditions.

$$\nabla_\phi \mathcal{L}(\phi) = \mathbb{E}_{q(\epsilon)q(u)}\Big(\nabla_z \log p(\mathcal{D},z) - \nabla_z \log q_\phi(z)\Big)\Big|_{z=h_\phi(u,\epsilon)} \cdot \nabla_\phi h_\phi(u,\epsilon), \tag{3}$$

The score of the joint distribution $\log p(\mathcal{D},z)$ is accessible. The main difficulty lies in the term $s_\phi = \nabla_z \log q_\phi(z)$, which is intractable due to the mixture integral. UIVI resolves this by using the identity

$$\nabla_z \log q_\phi(z) = \mathbb{E}_{\epsilon\sim q_\phi(\epsilon|z)}\big[\nabla_z \log q_\phi(z|\epsilon)\big] \tag{4}$$

expressing the marginal score as an expectation under the reverse conditional $q_\phi(\epsilon|z) \propto q(\epsilon)q_\phi(z|\epsilon)$. This setting is also applicable to sampling from an unnormalized distribution: one only needs to replace $\nabla_z \log p(\mathcal{D},z)$ with the gradient of the log unnormalized density.

In practice, UIVI estimates the expectation in (4) using an inner-loop MCMC sampler targeting $q_\phi(\epsilon|z)$, which can be computationally expensive in high-dimensional settings. Titsias and Ruiz [2019] argue that only a few MCMC steps may suffice because chains are initialized at a plausible state. However, the samples may fail to decorrelate when $q_\phi(\epsilon|z)$ has complex geometry. Importance-sampling variants replace MCMC samples with samples from a proposal distribution [Sobolev and Vetrov, 2019], while AISIVI [Pielok et al., 2025] takes a step further by incorporating CNFs to approximate the reverse conditional. These methods still rely on an inner approximation to $q_\phi(\epsilon|z)$, so the resulting score estimator can still suffer from numerical instability and high variance, especially when the reverse distribution has not yet been approximated well.

3 Methodology

In this section, we present Denoising Implicit Variational Inference (DIVI). The standard pathwise ELBO gradient in (3) is available once the marginal score $s_\phi(z) = \nabla_z \log q_\phi(z)$ is known. The proposed component of DIVI is to learn a proxy score network $s_\psi(z)$ for this marginal score, avoiding the reverse-conditional sampling used by UIVI. Training alternates between two steps: a ψ-step that fits $s_\psi(z)$ using tractable conditional scores, and a ϕ-step that updates the semi-implicit variational distribution using a plug-in ELBO gradient.

ψ-step (learning the marginal score) For a fixed variational parameter ϕ, the conditional score $s_\phi(z|\epsilon) = \nabla_z \log q_\phi(z|\epsilon)$ is tractable, whereas the marginal score $s_\phi(z)$ is not. By the score identity in (4), averaging the conditional score over the reverse conditional recovers the marginal score. This motivates fitting $s_\psi(z)$ by regressing against the conditional score on samples generated from the semi-implicit hierarchy. We therefore minimize the denoising score matching objective

$$\min_\psi \mathbb{E}_{\epsilon\sim q(\epsilon),u\sim q(u)}\left[\|s_\psi(z) - \nabla_z \log q_\phi(z|\epsilon)\|^2\right]$$

This objective is tractable because it only requires samples from $q(\epsilon)q(u)$ and evaluations of the conditional score. Let the population loss for this step be denoted as $\mathcal{J}(\phi,\psi)$.

Algorithm 1 DIVI Algorithm

Require: Dataset $\mathcal{D}$, Generator $z = h_\phi(u, \epsilon)$, Score Network $s_\psi(z)$, Priors $q(\epsilon), q(u)$.
Require: Learning rates η_ϕ, η_ψ, Inner steps L_ψ, Batch sizes N, B.
Initialize ϕ_0, ψ_{-1}.
for $n = 0, 1, 2, \ldots$ until convergence **do**
  Set $\phi \leftarrow \phi_n$ and $\psi \leftarrow \psi_{n-1}$.
  // ψ-step: fitting the marginal score.
  for $k = 1$ **to** L_ψ **do**
    Sample $\epsilon^{(m)} \sim q(\epsilon), u^{(m)} \sim q(u),\ z^{(m)} \leftarrow h_\phi(u^{(m)}, \epsilon^{(m)})$ for $m = 1 \ldots B$.
    $\mathcal{J}_\psi \leftarrow \frac{1}{B} \sum_{m=1}^{B} \left\| s_\psi(z^{(m)}) - \nabla_z \log q_\phi(z^{(m)} | \epsilon^{(m)}) \right\|^2$.
    $\psi \leftarrow \psi - \eta_\psi \nabla_\psi \mathcal{J}_\psi$.
  end for
  Set $\psi_n \leftarrow \psi$.
  // ϕ-step: optimizing the ELBO.
  Sample $\epsilon^{(i)} \sim q(\epsilon), u^{(i)} \sim q(u),\ z^{(i)} \leftarrow h_\phi(u^{(i)}, \epsilon^{(i)})$ for $i = 1 \ldots N$.
  $\hat{g}_n \leftarrow \frac{1}{N} \sum_{i=1}^{N} \left(\nabla_z \log p(\mathcal{D}, z^{(i)}) - s_{\psi_n}(z^{(i)}) \right) \nabla_\phi h_\phi(u^{(i)}, \epsilon^{(i)})$.
  $\phi_{n+1} \leftarrow \phi_n + \eta_\phi \hat{g}_n$.
end for

Proposition 3.1. *Denote* $\mathcal{E}^2(\phi, \psi) = \mathbb{E}_{q_\phi(z)} \| s_\psi(z) - \nabla_z \log q_\phi(z) \|^2$. *Then we have*

$$\mathcal{J}(\phi, \psi) = \mathcal{E}^2(\phi, \psi) + \mathbb{E}_z \{ \mathcal{I}(q_\phi(\epsilon | z)) \}$$

where $\mathcal{I}(q_\phi(\epsilon|z)) := \mathbb{E}_{q_\phi(\epsilon|z)} \| \nabla_z \log q_\phi(\epsilon|z) \|^2$.

For fixed ϕ, the second term in Proposition 3.1 is independent of ψ. Thus, minimizing the tractable denoising score matching objective $\mathcal{J}(\phi, \psi)$ with respect to ψ is equivalent to minimizing the marginal score error $\mathcal{E}^2(\phi, \psi)$. In particular, the optimal score model satisfies $s_{\psi^*}(z) = \nabla_z \log q_\phi(z)$ q_ϕ-a.e., yielding $\mathcal{E}(\phi, \psi^*) = 0$. The proof is in Appendix A.1.

ϕ-step (ELBO optimization with a plug-in score) Given the learned score network $s_\psi(z)$, we approximate the exact ELBO gradient by replacing the intractable marginal score. Using the reparameterization $z = h_\phi(u, \epsilon)$ with $\epsilon \sim q(\epsilon)$ and $u \sim q(u)$, this gives the averaged surrogate gradient:

$$\hat{g}(\phi; \psi) := \frac{1}{N} \sum_{i=1}^{N} \left(\nabla_z \log p(\mathcal{D}, z_i) - s_\psi(z_i) \right) \nabla_\phi h_\phi(u_i, \epsilon_i)$$

We then update ϕ by stochastic gradient ascent on the ELBO:

$$\phi_{n+1} = \phi_n + \eta_\phi \hat{g}(\phi_n, \psi_n).$$

At gradient-evaluation time, this approximation replaces reverse-conditional sampling with one evaluation of the learned score network, which, by design, reduces the computational cost of each ELBO-gradient estimate compared to previous sampling-based UIVI variants.

We implement DIVI by alternating between the score-fitting step and the ELBO-gradient step. The complete procedure is presented in Algorithm 1.

To compare the different strategies concerning gradient approximation, we list the key features of related methods in Table 1. The original SIVI provides a surrogate lower bound [Yin and Zhou, 2018], not directly optimizing the ELBO. However, the gradient of the surrogate lower bound also gives an approximation of the ELBO gradient [Pielok et al., 2025], with ϵ sampled from the prior distribution.

4 Theoretical Analysis

In this section, we analyze the optimization error introduced by replacing the exact marginal score with the learned score network. The convergence analysis of black-box variational inference (BBVI) has been studied in general settings [Ranganath et al., 2016, Kim et al., 2024, Domke et al., 2023],

Table 1: Comparison of gradient estimation strategies.

Method	Score Estimator $\hat{s}_\phi(z)$	Estimate Approach	Strategy
SIVI	$\nabla_z \log\left(\frac{1}{K}\sum_{i=1}^K q_\phi(z\|\epsilon_i)\right)$	$\epsilon_i \overset{\text{i.i.d.}}{\sim} q(\epsilon)$	Surrogate Lower Bound
UIVI	$\frac{1}{K}\sum_{i=1}^K \nabla_z \log q_\phi(z\|\epsilon_i)$	$\epsilon_i \sim q_\phi(\epsilon\|z)$ (approximately)	Hamiltonian Monte Carlo
AISIVI	$\nabla_z \log\left(\frac{1}{K}\sum_{i=1}^K \frac{q(\epsilon_i)q_\phi(z\|\epsilon_i)}{q_\psi(\epsilon_i\|z)}\right)$	$\epsilon_i \overset{\text{i.i.d.}}{\sim} q_\psi(\epsilon\|z)$	Importance Sampling
DIVI	$s_\psi(z)$	$\min_\psi \mathbb{E}\left[\|s_\psi(z) - \nabla_z \log q_\phi(z\|\epsilon)\|^2\right]$	**Denoising Score Matching**

* We compute $\nabla_\phi \mathcal{L}_{K-1}$ of SIVI under the reparameterization framework, which provides an approximation for $\nabla_\phi \mathcal{L}$.

but convergence guarantees for specific semi-implicit approaches remain limited [Moens et al., 2021, Cheng et al., 2024]. For ELBO-based semi-implicit methods such as UIVI and AISIVI, the main difficulty is to control the error of the score estimator used inside the pathwise gradient. DIVI makes this error explicit through the score-matching residual $\mathcal{E}(\phi, \psi)$ and the corresponding optimization process can be viewed as an inexact stochastic gradient procedure, allowing us to use tools for biased stochastic gradient ascent [Ajalloeian and Stich, 2020, Demidovich et al., 2023].

We use the Gaussian reparameterized form $z = h_\phi(u, \epsilon) = \mu_\phi(\epsilon) + u \odot \sigma_\phi(\epsilon)$, where $u \sim \mathcal{N}(0, I_d)$ and $\epsilon \in \mathbb{R}^{d_\epsilon}$ is typically Gaussian or uniform. At iteration n, let $\mathcal{F}_n$ denote the filtration generated by all randomness before the score estimator is updated. After obtaining ψ_n, define the enlarged filtration $\mathcal{G}_n = \mathcal{F}_n \vee \sigma(\psi_n)$. We use the following assumptions.

Assumption 1. *The objective $\mathcal{L}(\phi)$ is twice continuously differentiable and L-smooth with respect to ϕ.*

Assumption 2. *There exists a constant $M > 0$ such that $\mathbb{E}_{z\sim q_\phi}\|\nabla_z \log p(\mathcal{D}, z)\|^4 \leq M$, $\mathbb{E}_{z\sim q_\phi}\|s_\psi(z)\|^4 \leq M$ holds uniformly for ϕ, ψ.*

Assumption 3. *There exists a constant $C > 0$ such that $\max\left\{\mathbb{E}_\epsilon\|\nabla_\phi \sigma_\phi(\epsilon)\|_2^4, \mathbb{E}_\epsilon\|\nabla_\phi \mu_\phi(\epsilon)\|_2^4\right\} \leq C$ uniformly holds for ϕ, where $\|\cdot\|_2$ denotes the operator norm.*

The uniform boundedness conditions are understood on the algorithmic domain, where the iteration sequence (ϕ_n, ψ_n) is assumed to reside almost surely in a deterministic compact domain. In practice, this can be guaranteed by projected updates, which are standard in stochastic optimization analysis [Ghadimi et al., 2016]. The smoothness condition is standard in non-convex stochastic optimization. Assumption 2 controls the fourth moment of the score-error factor in the surrogate gradient through a uniform moment bound in expectation. We provide an additional derivation of this assumption using the linear growth conditions in [Cheng et al., 2024] in Appendix A.3. Assumption 3 is specific to our Gaussian reparameterization and controls the component-wise Jacobians of μ_ϕ and σ_ϕ. Numerical validations along the optimization path are provided in Appendix C.

Lemma 4.1. *The fourth moment of the operator norm of the reparameterization Jacobian is bounded, i.e.,*

$$\mathbb{E}_{u,\epsilon}\|\nabla_\phi h_\phi(u, \epsilon)\|_2^4 \leq C_h, \quad C_h := 8\left(4\log^2(2d) + 8\log(2d) + 9\right)C = \mathcal{O}(C(\log d)^2).$$

We next analyze the stochastic optimization dynamics. Since ϕ_n and ψ_n are $\mathcal{G}_n$-measurable, conditioned on $\mathcal{G}_n$, the samples $z_n^{(1)}, \ldots, z_n^{(N)} \overset{i.i.d.}{\sim} q_{\phi_n}$, where $z_n^{(i)} = h_{\phi_n}(u_n^{(i)}, \epsilon_n^{(i)})$, are independently generated from the current variational distribution. Let $\hat{g}_n$ be the mini-batch surrogate gradient defined in Algorithm 1. We use the conditional decomposition

$$\hat{g}_n = \nabla_\phi \mathcal{L}(\phi_n) + b_n + \xi_n,$$

where $b_n := \mathbb{E}[\hat{g}_n \mid \mathcal{G}_n] - \nabla_\phi \mathcal{L}(\phi_n)$ is the conditional bias, and $\xi_n := \hat{g}_n - \mathbb{E}[\hat{g}_n \mid \mathcal{G}_n]$ is the martingale difference noise with $\mathbb{E}[\xi_n \mid \mathcal{G}_n] = 0$.

Denote $\mathcal{E}_n^2 := \mathcal{E}^2(\phi_n, \psi_n) = \mathbb{E}_{z\sim q_{\phi_n}}\left[\|s_{\phi_n}(z_n) - s_{\psi_n}(z_n)\|^2 \mid \mathcal{G}_n\right]$, which is $\mathcal{G}_n$-measurable. We can control the bias and variance as follows:

Proposition 4.2. *For constants $J = \sqrt{C_h}$ and $\Sigma^2 = 4\sqrt{MC_h}$ we have $\|b_n\|^2 \le J\mathcal{E}_n^2$, $\mathbb{E}[\|\xi_n\|^2 \mid \mathcal{G}_n] \le \frac{\Sigma^2}{N}$.*

Using these bounds, we obtain the following modified ascent lemma for DIVI optimization. Compared with the standard stochastic ascent inequality, the bound contains two additional error terms: the Monte Carlo variance term, scaled by Σ^2/N, and the score-estimation bias term, scaled by $\mathcal{E}_n^2$.

Lemma 4.3. *If the fixed stepsize $\eta \le 1/(4L)$, we have*

$$\mathbb{E}[\mathcal{L}(\phi_{n+1}) \mid \mathcal{G}_n] \ge \mathcal{L}(\phi_n) + \frac{\eta}{4}\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 - \frac{\eta^2 L\Sigma^2}{2N} - \frac{3\eta J}{4}\mathcal{E}_n^2.$$

By summing the one-step ascent inequality over $n = 0, \dots, T-1$, we obtain an averaged stationarity bound:

Theorem 4.4. *Let $\Delta_0 := \mathcal{L}^* - \mathcal{L}(\phi_0)$. Under the assumptions above, if $\eta \le 1/(4L)$, then*

$$\frac{1}{T}\sum_{n=0}^{T-1} \mathbb{E}\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 \le \left(\frac{4\Delta_0}{\eta T} + \frac{2\eta L\Sigma^2}{N}\right) + 3J\left(\frac{1}{T}\sum_{n=0}^{T-1}\mathbb{E}\mathcal{E}_n^2\right).$$

In particular, if $T \ge 32L\Delta_0 N/\Sigma^2$, then the balancing stepsize $\eta \asymp \sqrt{\Delta_0 N/(L\Sigma^2 T)}$ is admissible, and the stochastic optimization part is $\mathcal{O}\left(\Sigma\sqrt{\frac{L\Delta_0}{NT}}\right)$.

See detailed proofs in Appendix A. Theorem 4.4 separates the optimization error into two parts. The first part is the usual stochastic optimization error caused by finite mini-batch sampling; when the balancing stepsize is admissible, it decays as $\mathcal{O}(T^{-1/2})$ for fixed N. The second part is the averaged score-estimation error, which captures the bias introduced by replacing the intractable score s_{ϕ_n} with the learned estimator s_{ψ_n}. By Proposition 3.1, the denoising score-matching objective has the marginal score as its population minimizer, hence this residual can be reduced by sufficiently optimizing the inner score-learning problem, up to the statistical and approximation errors of the score model. We provide empirical evaluations of this quantity during the training process in Section 6.4. Thus, DIVI approaches stationarity to the extent that the learned score tracks the marginal score along the optimization path.

5 Related Work

ELBO-based SIVI. Semi-implicit variational inference (SIVI) defines expressive variational distributions by mixing an explicit conditional distribution with an implicit mixing distribution, enabling efficient sampling while leaving the marginal density and marginal score intractable [Yin and Zhou, 2018, Moens et al., 2021]. The original SIVI objective optimizes tractable surrogate ELBO bounds, whereas UIVI targets the exact ELBO by estimating the marginal score through reverse-conditional sampling over the mixing variable [Titsias and Ruiz, 2019]. Importance-weighted and AISIVI-style methods reduce the reliance on MCMC by using proposal distributions, including expressive conditional normalizing flows [Sobolev and Vetrov, 2019, Pielok et al., 2025]. DIVI is closest to these ELBO-gradient methods, but replaces inner reverse-conditional score estimation with a learned marginal-score proxy trained by denoising score matching.

Alternative SIVI objectives. Other SIVI methods avoid marginal-density evaluation by changing the optimization criterion. Fisher-divergence-based SIVI uses score matching but leads to a minimax problem [Yu and Zhang, 2023], while KSIVI uses kernel Stein discrepancy to obtain a standard minimization objective [Cheng et al., 2024, Liu et al., 2016]. Particle semi-implicit VI uses particles to represent the mixing distribution in a special semi-implicit family construction and approximates Euclidean–Wasserstein gradient flows [Lim and Johansen, 2025]. These approaches are complementary to DIVI: they modify the objective, representation, or gradient-flow geometry, whereas DIVI keeps the ELBO-gradient framework and addresses its marginal-score bottleneck through denoising score matching [Hyvärinen, 2005, Vincent, 2011]. Convergence guarantees for ELBO-based black-box VI and related semi-implicit methods have been studied under stochastic-gradient or surrogate-objective settings [Ranganath et al., 2016, Kim et al., 2024, Moens et al., 2021, Cheng et al., 2024]. Our convergence analysis treats the resulting update as inexact ELBO-gradient optimization, with residual stationarity error controlled by the denoising score-matching error.

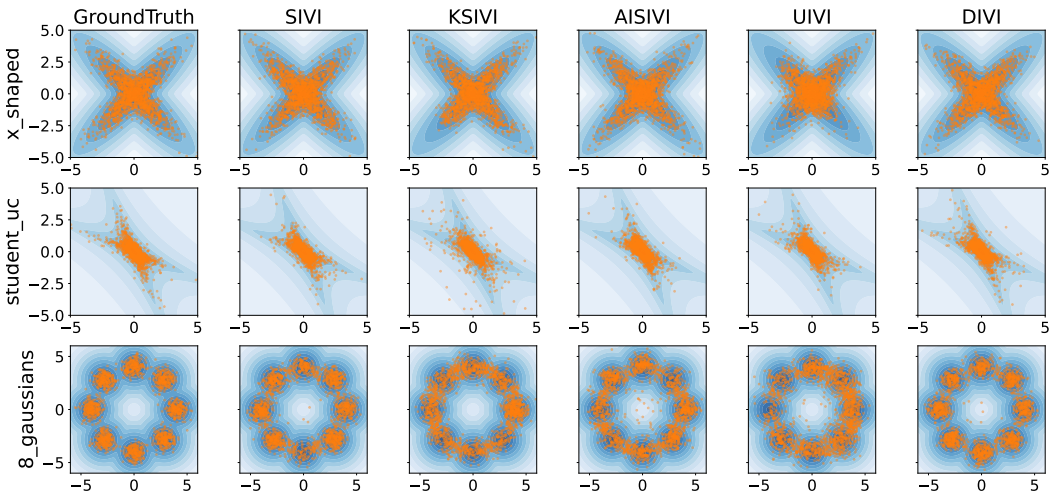

Figure 1: Qualitative comparison on two-dimensional targets. Each panel is generated from 2,000 samples from the ground-truth baseline or a learned variational distribution; blue contours show the target density.

6 Experiments

We evaluate DIVI across posterior inference problems that increase in dimension and structure. The experiments investigate whether DIVI can capture visible posterior geometry on two-dimensional synthetic targets, approximate a 100-dimensional conditioned diffusion posterior accurately, and provide predictive posterior distributions for Bayesian neural network (BNN) regression. In line with prior SIVI evaluations, we incorporate annealing techniques where appropriate. Unless stated otherwise, tables report the mean and standard error over 10 independent runs.

Our main comparisons are against UIVI and AISIVI [Titsias and Ruiz, 2019, Pielok et al., 2025]. We also include SIVI [Yin and Zhou, 2018] and KSIVI [Cheng et al., 2024] where their objectives apply. For the instantiation of the DIVI score model ψ, we use a multi-layer perceptron (MLP), which is warm-started before joint training. For each experiment, all methods share the same neural-network architecture for the semi-implicit variational family. The variational family, evaluation metrics, and per-experiment configurations are collected in Appendix B. We also provide additional arguments to support our theoretical Assumption 2 and Assumption 3 in Appendix C.

6.1 Toy two-dimensional targets

We first use low-dimensional unnormalized distribution targets to check whether the learned semi-implicit distribution captures geometric features that are easy to inspect visually.

Following previous works [Yu and Zhang, 2023, Li et al., 2023, Yu et al., 2023], we focus on the three plotted targets in Figure 1: `x_shaped`, `student_uc`, and `8_gaussians`. These targets probe for intersecting ridges, heavy-tailed dependence, and multi-modal coverage. We train SIVI, AISIVI, UIVI, and DIVI for 10,000 iterations and KSIVI for 50,000 iterations, with learning rate 10^{-3} and batch size 128. For DIVI, we train the proxy score MLP for 5 steps using 2048 samples and learning rate 10^{-3} after each update to the main VI model. Full per-method configurations and quantitative diagnostics are given in Appendix B.3.

Figure 1 shows that DIVI preserves the main geometry across the plotted targets, with sample clouds that closely follow the target contours in these visual comparisons. The quantitative diagnostics in Appendix B.3 show that DIVI obtains the lowest mean on four of the six KL-style and Wasserstein metrics and the second-lowest on the remaining two. We also provide a comparison of the learned score against the true target distribution score in Appendix B.7, which demonstrates that DIVI achieves the lowest L^2 error on example X-shaped target.

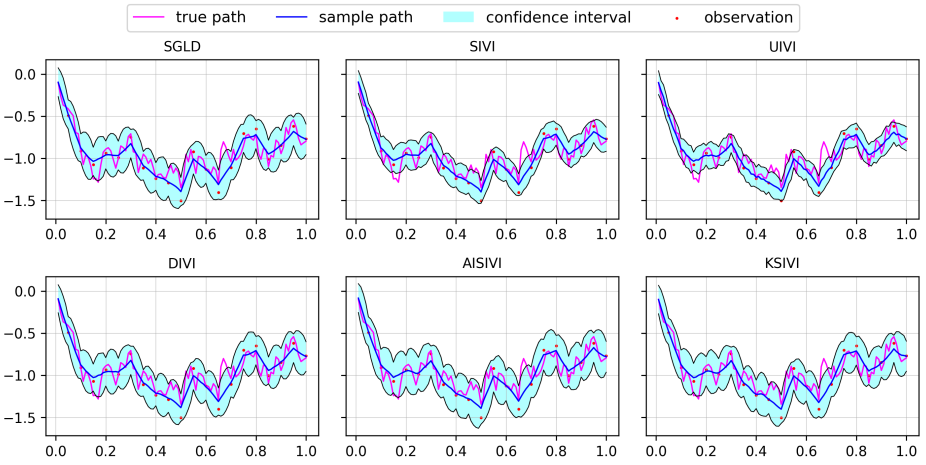

Figure 2: Posterior path approximations on the conditioned diffusion benchmark. The magenta curve is the true latent path, red points are observations, blue curves are posterior mean paths, and shaded regions are marginal 95% confidence intervals.

6.2 Conditioned diffusion process

We next evaluate a 100-dimensional conditioned diffusion posterior used in recent SIVI evaluations [Cheng et al., 2024, Pielok et al., 2025]. The latent path follows the double-well Langevin SDE

$$dx_t = 10x_t(1 - x_t^2)\,dt + dw_t, \qquad 0 \le t \le 1, \qquad x_0 = 0, \tag{5}$$

where w_t is one-dimensional Brownian motion. We discretize the path with Euler-Maruyama step size $\Delta t = 0.01$, giving $x = (x_{\Delta t}, x_{2\Delta t}, \ldots, x_{100\Delta t}) \in \mathbb{R}^{100}$. The corresponding path prior is

$$p_{\text{prior}}(x) = \prod_{i=1}^{100} (2\pi\Delta t)^{-1/2} \exp\left\{ -\frac{\left[x_{i\Delta t} - x_{(i-1)\Delta t} - 10x_{(i-1)\Delta t}(1 - x_{(i-1)\Delta t}^2)\Delta t\right]^2}{2\Delta t} \right\}. \tag{6}$$

Observations are made every five time steps,

$$y = (y_{5\Delta t}, y_{10\Delta t}, \ldots, y_{100\Delta t}), \qquad y_{5k\Delta t} \mid x_{5k\Delta t} \sim \mathcal{N}(x_{5k\Delta t}, \sigma^2), \tag{7}$$

for $k = 1, \ldots, 20$ and $\sigma = 0.1$. The inference target is therefore $p(x \mid y) \propto p_{\text{prior}}(x)p(y \mid x)$, with likelihood

$$p(y \mid x) = \prod_{k=1}^{20} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_{5k\Delta t} - x_{5k\Delta t})^2}{2\sigma^2} \right\}. \tag{8}$$

Posterior quality can be assessed through the mean path and the marginal uncertainty band. Following Cheng et al. [2024], we approximate the posterior by running stochastic gradient Langevin dynamics (SGLD) with 1000 particles, 100,000 iterations and a step size of 10^{-4}, and use the resulting samples as a high-quality reference for evaluation. In the evaluation, we include SIVI and KSIVI as additional baselines. We train SIVI, UIVI, AISIVI, and DIVI for 10,000 main VI iterations with learning rate 10^{-3} and batch size 128; KSIVI is trained for 100,000 iterations with the same learning rate and batch size and a slower learning rate decay schedule. For DIVI, we train the proxy score MLP for 5 steps using 2048 samples and learning rate 10^{-3} after each update to the main VI model. Further method configurations are in Appendix B.4.

Figure 2 shows that DIVI tracks the latent trajectory while maintaining a posterior band comparable to the SGLD reference. Following Pielok et al. [2025], Table 2 reports the expected log marginal likelihood of the 1,000 high-quality reference samples obtained from SGLD. For each method, we use 100,000 samples from the final trained variational model to approximate the marginal likelihood, and use the marginal likelihood estimated from 100,000 SGLD samples as the reference value. Larger values indicate better agreement with the reference posterior. We also report the wall-clock end-to-end training time for each method, including any warmup phases. As shown in Table 2, DIVI obtains the

Table 2: Conditioned diffusion results for expected log marginal likelihood and wall-clock time.

Method	ELM	wall-clock time (s)	iterations
SGLD	75.316 ± 0.00	–	–
SIVI	58.309 ± 0.11	368	10K
UIVI	21.866 ± 0.23	756	10K
AISIVI	74.687 ± 0.03	736	10K
KSIVI	74.808 ± 0.04	948	100K
DIVI	**75.162** ± 0.02	**335**	10K

Table 3: BNN test performance and computational cost. RMSE and NLL are reported as mean ± standard error. Lower is better. **Bold** indicates the best result; underline indicates the second best.

Dataset	Metric	SIVI	UIVI	AISIVI	KSIVI	DIVI
Boston	RMSE	2.826 ± 0.06	3.266 ± 0.11	3.294 ± 0.10	**2.661** ± 0.02	2.677 ± 0.06
	NLL	2.518 ± 0.01	2.606 ± 0.02	2.615 ± 0.02	2.490 ± 0.00	**2.477** ± 0.01
Concrete	RMSE	6.932 ± 0.09	7.668 ± 0.16	8.243 ± 0.28	6.351 ± 0.02	**6.342** ± 0.08
	NLL	3.373 ± 0.02	3.490 ± 0.03	3.539 ± 0.04	**3.262** ± 0.00	3.283 ± 0.02
Power	RMSE	3.936 ± 0.01	3.977 ± 0.01	4.007 ± 0.01	4.265 ± 0.14	**3.878** ± 0.01
	NLL	2.802 ± 0.00	2.813 ± 0.00	2.823 ± 0.00	2.931 ± 0.06	**2.781** ± 0.00
Protein	RMSE	4.892 ± 0.03	5.021 ± 0.00	4.934 ± 0.02	4.967 ± 0.00	**4.493** ± 0.00
	NLL	3.019 ± 0.00	3.029 ± 0.00	3.028 ± 0.00	3.021 ± 0.00	**2.927** ± 0.00
Wine Red	RMSE	0.623 ± 0.01	0.627 ± 0.00	0.625 ± 0.01	**0.604** ± 0.00	0.613 ± 0.00
	NLL	0.941 ± 0.01	0.947 ± 0.01	0.943 ± 0.01	**0.915** ± 0.00	0.930 ± 0.01
Yacht	RMSE	1.911 ± 0.12	2.720 ± 0.10	2.539 ± 0.10	**1.274** ± 0.07	1.358 ± 0.04
	NLL	1.862 ± 0.06	2.281 ± 0.08	2.181 ± 0.07	1.595 ± 0.04	**1.557** ± 0.06
Avg. wall-clock time		487	842	704	1100	**264**

highest expected log marginal likelihood among the VI methods and is closest to the SGLD reference, while also using the lowest wall-clock training time among the reported variational methods. Further metric and configuration details can be found in Appendix B.4.

6.3 Bayesian neural-network regression

We finally evaluate posterior inference for Bayesian neural networks (BNN) on six UCI regression targets: Boston Housing, Concrete, Power Plant, Protein, Wine Red, and Yacht. Following the BNN settings used for KSIVI and SIVI-SM [Cheng et al., 2024, Yu and Zhang, 2023], each target uses a single-hidden-layer ReLU network with 50 hidden units. These targets provide a diverse high-dimensional suite, with posterior dimensionality ranging from 301 to 751. Because these posteriors are data-dependent, we evaluate posterior predictive performance through test root mean square error (RMSE) and negative test log-likelihood (NLL), where lower values indicate better performance. We select the learning rate for each method on each target independently from $\{10^{-5}, 10^{-4}, 10^{-3}\}$ via cross-validation. We train all methods for 10,000 VI iterations except KSIVI, which, following the setting in Cheng et al. [2024], is trained for 100,000 iterations. We train the proxy score MLP for DIVI for 2 steps using 2048 samples and learning rate 10^{-3} after each optimization step to the variational distribution. Dataset and evaluation details are in Appendix B.5.

Table 3 gives the full predictive comparison across SIVI, UIVI, AISIVI, KSIVI, and DIVI. DIVI is best or second-best on most reported RMSE and NLL entries, with particularly strong performance on Power and Protein. The results also show cases where KSIVI remains stronger, such as Wine Red and Yacht RMSE. Overall, DIVI remains competitive with the strongest baselines while requiring the lowest average wall-clock training time.

6.4 Ablation studies

First we ablate on the effect of the score estimators of each method. We evaluate SIVI, UIVI, AISIVI, and DIVI as score estimators on the same fixed variational distributions obtained after 2,000 to 10,000 DIVI updates on the X-shaped target. Table 4 reports the squared L^2 discrepancy of the estimated score with respect to a reference score produced by UIVI-styled long HMC chains, together with the score-evaluation time of each method. DIVI achieves the lowest mean score-estimation error at every checkpoint and the lowest score-evaluation latency under the tested configurations. We may also observe that the score estimates of DIVI improves continually throughout the training process. More details of the evaluation are given in Appendix B.6. We also provide a comparison of the learned score against the true target distribution score along the trajectory of each method in Appendix B.7, which demonstrates that DIVI achieves the lowest L^2 error on all reported checkpoints on the example X-shaped target.

Table 4: Score-estimation error on fixed DIVI variational checkpoints for the X-shaped target (mean $\pm$ standard error over five seeds). Bold indicates the lowest mean.

Checkpoint	SIVI	UIVI	AISIVI	DIVI
2,000 steps	154.381 $\pm$ 68.264	1188.717 $\pm$ 509.566	45.529 $\pm$ 34.394	**0.899 $\pm$ 0.406**
4,000 steps	113.301 $\pm$ 27.096	889.808 $\pm$ 239.577	13.478 $\pm$ 8.219	**0.802 $\pm$ 0.388**
6,000 steps	78.763 $\pm$ 17.876	880.869 $\pm$ 314.731	7.249 $\pm$ 2.089	**0.407 $\pm$ 0.211**
8,000 steps	81.909 $\pm$ 27.307	972.443 $\pm$ 358.567	20.203 $\pm$ 6.899	**0.303 $\pm$ 0.137**
10,000 steps	107.791 $\pm$ 21.953	899.236 $\pm$ 273.263	6.881 $\pm$ 3.126	**0.331 $\pm$ 0.245**
Time (ms)	5.316 $\pm$ 0.082	204.523 $\pm$ 10.375	8.096 $\pm$ 0.099	**0.134 $\pm$ 0.001**

Additionally, we ablate on the expressiveness of the semi-implicit variational distributional family, providing comparison with a RealNVP variational family in Appendix B.9. We also perform ablations of DIVI's choice of score-update count in each variational optimization step and the reverse batch size in Appendix B.8. We believe this would provide practical guidance for future users of DIVI, as the score-update count and reverse batch size are two important hyperparameters in the accuracy-computaion tradeoff of DIVI.

7 Conclusion

We introduced Denoising Implicit Variational Inference (DIVI), an ELBO-based method for semi-implicit variational families that replaces reverse-conditional sampling with direct marginal-score learning. By fitting a score network via denoising score matching against tractable conditional scores, DIVI yields a simple plug-in pathwise ELBO-gradient estimator and avoids the inner sampling step used by previous UIVI-style methods. We analyzed the resulting procedure as inexact stochastic gradient ascent, showing that the averaged stationarity measure is controlled by the standard optimization term and the average score-matching error. Empirically, DIVI performs competitively across synthetic targets, conditioned diffusion posteriors, and Bayesian neural-network regression benchmarks, while achieving favorable wall-clock time in high-dimensional settings. Overall, these results suggest that denoising score estimation provides a practical route to scalable ELBO optimization with semi-implicit variational distributions. As a limitation, our analysis relies on smoothness of the target objective and boundedness of the optimization trajectory, and finite-capacity score networks may leave non-vanishing residual score-estimation error. We also acknowledge the performance in complex multimodal settings as a to-be-improved aspect, which might be mitigated by complex hierarchical variational family, though we leave these for future work.

Acknowledgments and Disclosure of Funding

This work was supported by National Natural Science Foundation of China (grant no. 12671341, grant no. 12292980 and grant no. 12292983). The research of Cheng Zhang was support in part by National Engineering Laboratory for Big Data Analysis and Applications, the Key Laboratory of Mathematics and Its Applications (LMAM) and the Key Laboratory of Mathematical Economics and

Quantitative Finance (LMEQF) of Peking University. The authors thank Zhifei Li and Haoyu Tao for the helpful discussions and feedback in the preliminary stages of this work.

References

Ahmad Ajalloeian and Sebastian U. Stich. Analysis of SGD with biased gradient estimators. *arXiv preprint arXiv:2008.00051*, 2020.

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.

Ziheng Cheng, Longlin Yu, Tianyu Xie, Shiyue Zhang, and Cheng Zhang. Kernel semi-implicit variational inference. In *Forty-first International Conference on Machine Learning*, 2024.

Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtarik. A guide through the zoo of biased sgd. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 23158–23171. Curran Associates, Inc., 2023.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Justin Domke, Robert Gower, and Guillaume Garrigos. Provable convergence guarantees for black-box variational inference. *Advances in neural information processing systems*, 36:66289–66327, 2023.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

Ryan J Giordano, Tamara Broderick, and Michael I Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. *Advances in neural information processing systems*, 28, 2015.

Ferenc Huszár. Variational inference using implicit distributions. *ArXiv*, abs/1702.08235, 2017.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL `http://jmlr.org/papers/v6/hyvarinen05a.html`.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kyurae Kim, Jisu Oh, Kaiwen Wu, Yi-An Ma, and Jacob R. Gardner. On the convergence of black-box variational inference, 2024. URL `https://arxiv.org/abs/2305.15349`.

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.

Lingxiao Li, Qiang Liu, Anna Korba, Mikhail Yurochkin, and Justin Solomon. Sampling with mollified interaction energy descent, 2023. URL `https://arxiv.org/abs/2210.13400`.

Jen Ning Lim and Adam M. Johansen. Particle semi-implicit variational inference, 2025. URL `https://arxiv.org/abs/2407.00649`.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, pages 1791–1799. PMLR, 2014.

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

John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

Carsten Peterson and James R Anderson. A mean field theory learning algorithm for neural networks. *Complex systems*, 1(5), 1987.

Tobias Pielok, Bernd Bischl, and David Rügamer. Revisiting unbiased implicit variational inference. In *Forty-second International Conference on Machine Learning*, 2025.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Lawrence K Saul, Tommi Jaakkola, and Michael I Jordan. Mean field theory for sigmoid belief networks. *Journal of artificial intelligence research*, 4:61–76, 1996.

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

Artem Sobolev and Dmitry P. Vetrov. Importance weighted hierarchical variational inference. *ArXiv*, abs/1905.03290, 2019.

Michalis K Titsias and Francisco Ruiz. Unbiased implicit variational inference. In *The 22nd international conference on artificial intelligence and statistics*, pages 167–176. PMLR, 2019.

Dustin Tran, Rajesh Ranganath, and David M. Blei. Hierarchical implicit models and likelihood-free variational inference. In *Neural Information Processing Systems*, 2017.

Richard Turner, M. Sahani, D. Barber, Ali Cemgil, and S. Chiappa. Two problems with variational expectation maximisation for time-series models. *Bayesian Time Series Models*, pages 109–130, 01 2011.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

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.

Longlin Yu, Tianyu Xie, Yu Zhu, Tong Yang, Xiangyu Zhang, and Cheng Zhang. Hierarchical semi-implicit variational inference with application to diffusion model acceleration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

A Proofs of Propositions

A.1 Proof of Proposition 3.1

Proof. Expanding the squared norm with $\nabla_z \log q_\phi(z|\epsilon) - \nabla_z \log q_\phi(z) = \nabla_z \log q_\phi(\epsilon|z)$:

$$\begin{aligned}
\mathbb{E}_{z,\epsilon}\|s_\psi(z) - \nabla_z \log q_\phi(z|\epsilon)\|^2 &= \mathbb{E}_{z,\epsilon}\|(s_\psi(z) - \nabla_z \log q_\phi(z)) - \nabla_z \log q_\phi(\epsilon|z)\|^2 \\
&= \underbrace{\mathbb{E}_z\|s_\psi(z) - \nabla_z \log q_\phi(z)\|^2}_{\mathcal{E}^2(\phi,\psi)} + \underbrace{\mathbb{E}_{z,\epsilon}\|\nabla_z \log q_\phi(\epsilon|z)\|^2}_{\mathbb{E}_z\mathcal{I}(q_\phi(\epsilon|z))} \\
&\quad - 2\mathbb{E}_z\big\langle s_\psi(z) - \nabla_z \log q_\phi(z), \underbrace{\mathbb{E}_{q_\phi(\epsilon|z)}\nabla_z \log q_\phi(\epsilon|z)}_{\text{zero}} \big\rangle \\
&= \mathcal{E}^2(\phi,\psi) + \mathbb{E}_z\mathcal{I}(q_\phi(\epsilon|z))
\end{aligned}$$

where the last equality holds under the standard regularity conditions that allow differentiation under the integral sign. If the score model class is sufficiently rich, the population minimizer satisfies $s_{\psi^*}(z) = \nabla_z \log q_\phi(z)$ q_ϕ-a.e., and hence $\mathcal{E}(\phi,\psi^*) = 0$. □

A.2 Proof of Lemma 4.1

Proof. Under the reparameterization $z = \mu_\phi(\epsilon) + u \odot \sigma_\phi(\epsilon)$, $u \sim \mathcal{N}(0, I_d), \epsilon \in \mathbb{R}^{d_\epsilon}, \phi \in \mathbb{R}^{d_\phi}$, the gradient $\nabla_\phi \mu_\phi(\epsilon) \in \mathbb{R}^{d\times d_\phi}, \nabla_\phi \sigma_\phi(\epsilon) \in \mathbb{R}^{d\times d_\phi}$. Throughout the proof, all vectors are treated as row vectors.

First we give an upper bound for $\mathbb{E}\|u\|_\infty^4 = \mathbb{E}\max_{1\le i\le d}|u_i|^4$. For a standard Gaussian random variable, we use the tail bound $\mathbb{P}(|u_i| > y) \le 2e^{-y^2/2}, y > 0$. Therefore, for any $x > 0$, $\mathbb{P}(u_i^4 > x) = \mathbb{P}(|u_i| > x^{1/4}) \le 2e^{-\sqrt{x}/2}$. Hence,

$$\mathbb{E}\|u\|_\infty^4 = \int_0^\infty \mathbb{P}\left(\max_{1\le i\le d} u_i^4 > x\right)\mathrm{d}x \le \int_0^\infty \min\{1, \sum_{i=1}^d \mathbb{P}(u_i^4 > x)\}\mathrm{d}x \le \int_0^\infty \min\{1, 2d\cdot e^{-\sqrt{x}/2}\}\mathrm{d}x.$$

Solving for x_0 such that $2d \cdot e^{-\sqrt{x}/2} = 1$ yields $x_0 = 4\log^2(2d)$. Splitting the integral at x_0, we obtain:

$$\begin{aligned}
\int_0^\infty \min\{1, 2d\cdot e^{-\sqrt{x}/2}\}\mathrm{d}x &= 4\log^2(2d) + 2d\int_{4\log^2(2d)}^\infty e^{-\sqrt{x}/2}\mathrm{d}x. \\
&= 4\log^2(2d) + 16d\int_{\log(2d)}^\infty te^{-t}\mathrm{d}t \\
&= 4\log^2(2d) + 8\log(2d) + 8.
\end{aligned}$$

So we arrive at the upper bound for the fourth moment $\mathbb{E}\|u\|_\infty^4 \le 4\log^2(2d) + 8\log(2d) + 8$. Based on this result, we can control the fourth order moment $\mathbb{E}\|\nabla_\phi h_\phi(u,\epsilon)\|_2^4$:

$$\begin{aligned}
\mathbb{E}\|\nabla_\phi h_\phi(u,\epsilon)\|_2^4 &= \mathbb{E}\|\nabla_\phi \mu_\phi(\epsilon) + \mathrm{diag}(u)\nabla_\phi \sigma_\phi(\epsilon)\|_2^4 \\
&\le 8\,\mathbb{E}\left(\|\nabla_\phi \mu_\phi(\epsilon)\|_2^4 + \|u\|_\infty^4 \cdot \|\nabla_\phi \sigma_\phi(\epsilon)\|_2^4\right) \\
&= 8\mathbb{E}_\epsilon\|\nabla_\phi \mu_\phi(\epsilon)\|_2^4 + 8\mathbb{E}_u\|u\|_\infty^4 \cdot \mathbb{E}_\epsilon\|\nabla_\phi \sigma_\phi(\epsilon)\|_2^4 \\
&\le 8\mathbb{E}_\epsilon\|\nabla_\phi \mu_\phi(\epsilon)\|_2^4 + 8(4\log^2(2d) + 8\log(2d) + 8)\mathbb{E}_\epsilon\|\nabla_\phi \sigma_\phi(\epsilon)\|_2^4 \\
&\le 8(4\log^2(2d) + 8\log(2d) + 9)C = \mathcal{O}(C(\log d)^2).
\end{aligned}$$

Here, $\|\cdot\|_2$ denotes the $2-$ operator norm for matrices, $\mathrm{diag}(u)$ is the diagonal matrix with $(\mathrm{diag}(u))_{ii} = u_i$. □

A.3 Bound Based on Linear Growth Condition

We show that Assumption 2 can be derived from a linear growth condition together with a bounded fourth moment condition on the reparameterization functions.

Similar to the conditions of linear growth and fourth-moment boundedness in [Cheng et al., 2024], we assume that there exists a constant $\lambda > 0$ such that

$$\max\left\{\|\nabla_z \log p(\mathcal{D}, z)\|, \|s_\psi(z)\|\right\} \leq \lambda(\|z\| + 1)$$

holds uniformly for z and ψ and there exists a constant $\zeta > 0$ such that

$$\max\left\{\mathbb{E}_\epsilon\|\mu_\phi(\epsilon)\|^4, \mathbb{E}_\epsilon\|\sigma_\phi(\epsilon)\|^4\right\} \leq \zeta$$

holds uniformly for ϕ. Then Assumption 2 holds with

$$M = 8\lambda^4\left[8\zeta\left(4\log^2(2d) + 8\log(2d) + 9\right) + 1\right] = \mathcal{O}(\zeta\lambda^4\log^2 d).$$

Proof. It suffices to prove the bound for $s_\psi(z)$; the bound for $\nabla_z \log p(\mathcal{D}, z)$ follows identically. By the linear growth condition and $(a+b)^4 \leq 8(a^4 + b^4)$,

$$\begin{aligned}
\mathbb{E}_{q_\phi}\|s_\psi(z)\|^4 &\leq \lambda^4\mathbb{E}_{q_\phi}(\|z\| + 1)^4 \leq 8\lambda^4\left(\mathbb{E}_{q_\phi}\|z\|^4 + 1\right) \\
&= 8\lambda^4\left(\mathbb{E}_{u,\epsilon}\|\mu_\phi(\epsilon) + u \odot \sigma_\phi(\epsilon)\|^4 + 1\right) \\
&\leq 8\lambda^4\left(8\mathbb{E}_\epsilon\|\mu_\phi(\epsilon)\|^4 + 8\mathbb{E}_{u,\epsilon}\|u \odot \sigma_\phi(\epsilon)\|^4 + 1\right) \\
&\leq 8\lambda^4\left(8\zeta + 8\mathbb{E}_u\|u\|_\infty^4\mathbb{E}_\epsilon\|\sigma_\phi(\epsilon)\|^4 + 1\right) \\
&\leq 8\lambda^4\left[8\zeta + 8\zeta\left(4\log^2(2d) + 8\log(2d) + 8\right) + 1\right] \\
&= 8\lambda^4\left[8\zeta\left(4\log^2(2d) + 8\log(2d) + 9\right) + 1\right].
\end{aligned}$$

Therefore, $\mathbb{E}_{q_\phi}\|s_\psi(z)\|^4 \leq M$, where $M := 8\lambda^4\left[8\zeta\left(4\log^2(2d) + 8\log(2d) + 9\right) + 1\right]$. The same argument gives $\mathbb{E}_{q_\phi}\|\nabla_z \log p(\mathcal{D}, z)\|^4 \leq M$. We also provide numerical validations on the reasonable bound of ζ and λ, which is discussed concisely in Appendix C □

A.4 Proof of Proposition 4.2

Proof. Based on the preparations in Appendix A.2, we can respectively bound the bias and variance of the surrogate gradient.

First, we fix (ϕ, ψ) and take all expectations with respect to the fresh sampling variables (u, ϵ), with $z = h_\phi(u, \epsilon)$. Equivalently, expectations of functions depending only on z are taken with respect to $z \sim q_\phi$. We concentrate on two estimators only using a single sample z:

$$g(z;\phi) = \left(\nabla_z \log p(\mathcal{D}, z) - s_\phi(z)\right)\nabla_\phi h_\phi(u, \epsilon) \tag{9}$$

$$\hat{g}(z;\phi,\psi) = \left(\nabla_z \log p(\mathcal{D}, z) - s_\psi(z)\right)\nabla_\phi h_\phi(u, \epsilon) \tag{10}$$

Since $\mathbb{E}_{u,\epsilon}g(z;\phi) = \nabla_\phi\mathcal{L}(\phi)$, (9) gives an unbiased estimator of $\nabla_\phi\mathcal{L}(\phi)$. However, the semi-implicit score function $s_\phi(z) = \nabla_z \log q_\phi(z)$ is intractable. Therefore, (10) gives a biased estimator of $\nabla_\phi\mathcal{L}(\phi)$ with bias-variance decomposition:

$$\hat{g}(z;\phi,\psi) = \nabla_\phi\mathcal{L}(\phi) + b(\phi,\psi) + \xi(z;\phi,\psi) \tag{11}$$

where $b(\phi,\psi) = \mathbb{E}\hat{g}(z;\phi,\psi) - \nabla_\phi\mathcal{L}(\phi)$, $\xi(z;\phi,\psi) = \hat{g}(z;\phi,\psi) - \nabla_\phi\mathcal{L}(\phi) - b(\phi,\psi)$. Then $\mathbb{E}\xi(z;\phi,\psi) = 0$.

By Lemma 4.1 and Cauchy-Schwarz inequality, we can control the bias norm $\|b(\phi,\psi)\|$ by the score matching error $\mathcal{E}(\phi,\psi)$:

$$\begin{aligned}
\|b(\phi,\psi)\|^2 &= \|\mathbb{E}_{u,\epsilon}\left[\hat{g}(z;\phi,\psi) - g(z;\phi)\right]\|^2 \\
&\leq \left(\mathbb{E}_{u,\epsilon}\|\hat{g}(z;\phi,\psi) - g(z;\phi)\|\right)^2 \\
&= \left(\mathbb{E}_{u,\epsilon}\|\left(s_\phi(z) - s_\psi(z)\right)\nabla_\phi h_\phi(u,\epsilon)\|\right)^2, \\
&\leq \left(\mathbb{E}_{u,\epsilon}\left[\|s_\phi(z) - s_\psi(z)\| \cdot \|\nabla_\phi h_\phi(u,\epsilon)\|_2\right]\right)^2 \\
&\leq \mathbb{E}_{z\sim q_\phi}\|s_\phi(z) - s_\psi(z)\|^2 \cdot \mathbb{E}_{u,\epsilon}\|\nabla_\phi h_\phi(u,\epsilon)\|_2^2 \\
&\leq \mathbb{E}_{z\sim q_\phi}\|s_\phi(z) - s_\psi(z)\|^2 \cdot \sqrt{\mathbb{E}_{u,\epsilon}\|\nabla_\phi h_\phi(u,\epsilon)\|_2^4} \\
&\leq \sqrt{C_h} \cdot \mathcal{E}^2(\phi,\psi) = J \cdot \mathcal{E}^2(\phi,\psi).
\end{aligned}$$

Next, $\mathbb{E}\|\xi\|^2$ could be controlled similarly.

$$\begin{aligned}
\mathbb{E}_{u,\epsilon}\|\xi\|^2 &= \mathbb{E}_{u,\epsilon}\left\|\hat{g}(z;\phi,\psi) - \mathbb{E}_{u,\epsilon}\hat{g}(z;\phi,\psi)\right\|^2 \\
&\le \mathbb{E}_{u,\epsilon}\|\hat{g}(z;\phi,\psi)\|^2 \\
&= \mathbb{E}_{u,\epsilon}\left\|\left(\nabla_z \log p(\mathcal{D},z) - s_\psi(z)\right)\nabla_\phi h_\phi(u,\epsilon)\right\|^2 \\
&\le \mathbb{E}_{u,\epsilon}\left[\|\nabla_z \log p(\mathcal{D},z) - s_\psi(z)\|^2 \cdot \|\nabla_\phi h_\phi(u,\epsilon)\|_2^2\right] \\
&\le \sqrt{\mathbb{E}_{z\sim q_\phi}\|\nabla_z \log p(\mathcal{D},z) - s_\psi(z)\|^4} \cdot \sqrt{\mathbb{E}_{u,\epsilon}\|\nabla_\phi h_\phi(u,\epsilon)\|_2^4} \\
&\le \sqrt{8(\mathbb{E}_{z\sim q_\phi}\|\nabla_z \log p(\mathcal{D},z)\|^4 + \mathbb{E}_{z\sim q_\phi}\|s_\psi(z)\|^4)} \cdot \sqrt{\mathbb{E}_{u,\epsilon}\|\nabla_\phi h_\phi(u,\epsilon)\|_2^4} \\
&\le 4\sqrt{C_h \cdot M} = \Sigma^2.
\end{aligned}$$

We now relate these conditional quantities to the preceding single-sample bounds. The bounds derived above were fixed-parameter bounds: for any fixed (ϕ,ψ), with fresh samples $z = h_\phi(u,\epsilon)$, we have $\|b(\phi,\psi)\|^2 \le J\mathcal{E}^2(\phi,\psi), \quad \mathbb{E}\|\xi(z;\phi,\psi)\|^2 \le \Sigma^2$. Conditioning on $\mathcal{G}_n$, the iterates ϕ_n and ψ_n are fixed, and the only remaining randomness comes from the fresh mini-batch. Therefore, the preceding fixed-parameter bias bound can be applied with $(\phi,\psi) = (\phi_n,\psi_n)$, yielding $\|b_n\|^2 \le J\mathcal{E}_n^2$.

It remains to justify the mini-batch variance bound. Define the single-sample surrogate gradient at iteration n by $g_n^{(i)} = \left(\nabla_z \log p(\mathcal{D}, z_n^{(i)}) - s_{\psi_n}(z_n^{(i)})\right)\nabla_\phi h_{\phi_n}(u_n^{(i)}, \epsilon_n^{(i)})$, Then $\hat{g}_n = \frac{1}{N}\sum_{i=1}^N g_n^{(i)}$, and $\xi_n = \frac{1}{N}\sum_{i=1}^N \left(g_n^{(i)} - \mathbb{E}[g_n^{(i)} \mid \mathcal{G}_n]\right)$. Conditional on $\mathcal{G}_n$, the variables $g_n^{(1)},\dots,g_n^{(N)}$ are i.i.d. Hence the cross terms vanish, and

$$\mathbb{E}[\|\xi_n\|^2 \mid \mathcal{G}_n] = \frac{1}{N}\mathbb{E}\left[\left\|g_n^{(1)} - \mathbb{E}[g_n^{(1)} \mid \mathcal{G}_n]\right\|^2 \mid \mathcal{G}_n\right] \le \frac{1}{N}\mathbb{E}\left[\|g_n^{(1)}\|^2 \mid \mathcal{G}_n\right] \le \frac{\Sigma^2}{N}.$$

This completes the proof. □

A.5 Proof of Lemma 4.3

Proof. Recall that the update rule is $\phi_{n+1} = \phi_n + \eta\hat{g}_n$. By the L-smoothness of $\mathcal{L}(\phi)$, we have

$$\begin{aligned}
\mathcal{L}(\phi_{n+1}) &\ge \mathcal{L}(\phi_n) + \langle \phi_{n+1} - \phi_n, \nabla_\phi \mathcal{L}(\phi_n)\rangle - \frac{L}{2}\|\phi_{n+1} - \phi_n\|^2 \\
&= \mathcal{L}(\phi_n) + \eta\left\langle \nabla_\phi \mathcal{L}(\phi_n), \hat{g}_n\right\rangle - \frac{\eta^2 L}{2}\|\hat{g}_n\|^2.
\end{aligned}$$

By Proposition 4.2, we have the decomposition $\hat{g}_n = \nabla_\phi \mathcal{L}(\phi_n) + b_n + \xi_n, \quad \mathbb{E}[\xi_n \mid \mathcal{G}_n] = 0$, with $\|b_n\|^2 \le J\mathcal{E}_n^2, \quad \mathbb{E}[\|\xi_n\|^2 \mid \mathcal{G}_n] \le \frac{\Sigma^2}{N}$. Taking conditional expectation with respect to $\mathcal{G}_n$, we obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\phi_{n+1}) \mid \mathcal{G}_n] &\ge \mathcal{L}(\phi_n) + \eta\left\langle \nabla_\phi \mathcal{L}(\phi_n), \nabla_\phi \mathcal{L}(\phi_n) + b_n\right\rangle - \frac{\eta^2 L}{2}\mathbb{E}\left[\left\|\nabla_\phi \mathcal{L}(\phi_n) + b_n + \xi_n\right\|^2 \mid \mathcal{G}_n\right] \\
&= \mathcal{L}(\phi_n) + \eta\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 + \eta\left\langle \nabla_\phi \mathcal{L}(\phi_n), b_n\right\rangle - \frac{\eta^2 L}{2}\left\|\nabla_\phi \mathcal{L}(\phi_n) + b_n\right\|^2 \\
&\quad - \frac{\eta^2 L}{2}\mathbb{E}[\|\xi_n\|^2 \mid \mathcal{G}_n] \\
&\ge \mathcal{L}(\phi_n) + \eta\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 - \frac{\eta}{2}\left(\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 + \|b_n\|^2\right) \\
&\quad - \eta^2 L\left(\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 + \|b_n\|^2\right) - \frac{\eta^2 L}{2}\mathbb{E}[\|\xi_n\|^2 \mid \mathcal{G}_n] \\
&= \mathcal{L}(\phi_n) + \frac{\eta}{2}(1 - 2\eta L)\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 - \frac{\eta}{2}(1 + 2\eta L)\|b_n\|^2 - \frac{\eta^2 L}{2}\mathbb{E}[\|\xi_n\|^2 \mid \mathcal{G}_n] \\
&\ge \mathcal{L}(\phi_n) + \frac{\eta}{2}(1 - 2\eta L)\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 - \frac{\eta J}{2}(1 + 2\eta L)\mathcal{E}_n^2 - \frac{\eta^2 L\Sigma^2}{2N}.
\end{aligned}$$

In the last inequality, we used $\langle a,b\rangle \ge -\frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2$, $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$.

Under the stepsize condition $\eta \le 1/(4L)$, we have $\frac{\eta}{2}(1-2\eta L) \ge \frac{\eta}{4}$, $\frac{\eta J}{2}(1+2\eta L) \le \frac{3\eta J}{4}$. Therefore,

$$\mathbb{E}[\mathcal{L}(\phi_{n+1}) \mid \mathcal{G}_n] \ge \mathcal{L}(\phi_n) + \frac{\eta}{4}\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 - \frac{\eta^2 L \Sigma^2}{2N} - \frac{3\eta J}{4}\mathcal{E}_n^2.$$

This completes the proof. □

A.6 Proof of Theorem 4.4

Proof. Taking total expectation on both sides of Lemma 4.3, we obtain

$$\mathbb{E}\mathcal{L}(\phi_{n+1}) \ge \mathbb{E}\mathcal{L}(\phi_n) + \frac{\eta}{4}\mathbb{E}\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 - \frac{\eta^2 L \Sigma^2}{2N} - \frac{3\eta J}{4}\mathbb{E}\mathcal{E}_n^2.$$

Summing from $n = 0$ to $T-1$, we get

$$\mathbb{E}\mathcal{L}(\phi_T) \ge \mathcal{L}(\phi_0) + \frac{\eta}{4}\sum_{n=0}^{T-1}\mathbb{E}\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 - \frac{\eta^2 L \Sigma^2 T}{2N} - \frac{3\eta J}{4}\sum_{n=0}^{T-1}\mathbb{E}\mathcal{E}_n^2.$$

Rearranging and using $\mathbb{E}\mathcal{L}(\phi_T) \le \mathcal{L}^*$,

$$\frac{\eta}{4}\sum_{n=0}^{T-1}\mathbb{E}\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 \le \Delta_0 + \frac{\eta^2 L \Sigma^2 T}{2N} + \frac{3\eta J}{4}\sum_{n=0}^{T-1}\mathbb{E}\mathcal{E}_n^2.$$

Dividing both sides by $\eta T/4$, we obtain

$$\frac{1}{T}\sum_{n=0}^{T-1}\mathbb{E}\|\nabla_\phi \mathcal{L}(\phi_n)\|^2 \le \frac{4(\Delta_0)}{\eta T} + \frac{2\eta L \Sigma^2}{N} + 3J\left(\frac{1}{T}\sum_{n=0}^{T-1}\mathbb{E}\mathcal{E}_n^2\right) \tag{12}$$

$$= \mathcal{O}\left(\frac{1}{\eta T} + \frac{\eta}{N} + \frac{1}{T}\sum_{n=0}^{T-1}\mathbb{E}\mathcal{E}_n^2\right). \tag{13}$$

The right-hand side consists of three terms. The first term is the finite-iteration optimization error, the second term comes from the Monte Carlo variance of the mini-batch estimator, and the third term reflects the accumulated bias caused by the imperfect score estimator s_{ψ_n}. Balancing the first two terms suggests the choice $\eta \asymp \min\left\{\frac{1}{4L}, \sqrt{\frac{N(\Delta_0)}{L\Sigma^2 T}}\right\}$. With the balancing stepsize, when it is admissible, the stochastic optimization part satisfies $\frac{4\Delta_0}{\eta T} + \frac{2\eta L\Sigma^2}{N} = \mathcal{O}\left(\Sigma\sqrt{\frac{L\Delta_0}{NT}}\right)$. Therefore, if N is fixed and L, Δ_0, Σ are treated as constants, the stochastic optimization error is of order $\mathcal{O}(T^{-1/2})$. □

B Experimental Details

B.1 Shared setup

Variational family. All SIVI variants use a conditional Gaussian semi-implicit variational family. We write the mixing variable as ϵ and the latent variable as z, with

$$q_\phi(z) = \int q_\phi(z \mid \epsilon) q(\epsilon)\, d\epsilon, \qquad q_\phi(z \mid \epsilon) = \mathcal{N}(z; \mu_\phi(\epsilon), \Sigma_\phi(\epsilon)). \tag{14}$$

For two-dimensional targets, the mixing prior is standard Gaussian, and μ_ϕ and the diagonal variance are produced by a two-hidden-layer MLP with hidden width 128. For the conditioned diffusion task, $\epsilon \in [0,1]^{100}$ uses a uniform prior, μ_ϕ is a two-hidden-layer ReLU MLP with hidden width 128, and the conditional covariance is a learned global diagonal variance initialized near 0.01725. For BNN regression, $\epsilon \in \mathbb{R}^3$ uses a standard Gaussian prior, μ_ϕ and the diagonal log variance are produced by a two-hidden-layer ReLU MLP with hidden width 10, and the log variance is initialized at -8 with a lower clamp at -20. These BNN choices follow the compact semi-implicit architecture used in the KSIVI and SIVI-SM comparisons.

Common training protocol. Unless otherwise noted, quantitative results use ten random seeds and report the mean and standard error. The supplementary score-error comparisons, reverse-update ablations, and RealNVP comparison use five seeds and report standard errors. For n independent runs, the standard error is $s/\sqrt{n}$, where s is the sample standard deviation across runs. Unless otherwise noted, SIVI, UIVI, AISIVI, and DIVI use 10,000 main VI iterations with batch size 128, Adam optimization, and a step learning-rate scheduler. Unless otherwise noted, all methods are trained with annealing enabled for half of the total training iterations. All experiments are conducted on a Linux server with Nvidia RTX 4090 GPUs and Intel Xeon Gold 6430 @ 2.10GHz.

SIVI. SIVI approximates the semi-implicit marginal $\log q_\phi(z)$ using Monte Carlo averages over auxiliary ϵ samples. We draw 4,096 auxiliary samples on diffusion targets and 2,048 on BNN targets due to GPU memory constraints.

UIVI. UIVI samples from the reverse conditional $q_\phi(\epsilon \mid z)$ via Hamiltonian Monte Carlo. After 5 burn-in transitions, a single HMC trajectory of 5 leapfrog steps provides $K = 5$ auxiliary samples, one per intermediate position, following Titsias and Ruiz [2019]. The leapfrog step size is 0.2.

B.2 Metrics

KL-style ELBO. For toy and diffusion targets, we report a KL-style objective,

$$D_{\mathrm{KL}} = -\widehat{\mathcal{L}}(q_\phi) = -\frac{1}{N}\sum_{i=1}^{N}\left[\log p(z_i) - \widehat{\log q_\phi(z_i)}\right], \qquad z_i \sim q_\phi. \tag{15}$$

The estimate uses $N = 5{,}000$ VI samples. For each z_i, $\widehat{\log q_\phi(z_i)}$ is approximated with 20 batches of 2,048 auxiliary ϵ samples, plus the generating ϵ sample. We utilize the batching logic in Pielok et al. [2025] to increase the effective number of Monte Carlo auxiliary samples to 40960.

Sliced Wasserstein-2. Sliced Wasserstein-2 is computed by averaging one-dimensional Wasserstein distances over random unit projections,

$$\mathrm{SW}_2^2(q, r) = \mathbb{E}_{\theta \sim \mathrm{Unif}(\mathbb{S}^{d-1})}\left[W_2^2\big(\theta^\top q, \theta^\top r\big)\right], \tag{16}$$

estimated using 10,000 samples and 5,000 random projections. For heavy-tailed distributions, we report a truncated variant following Li et al. [2023]: both variational and reference samples are retained only if every coordinate has absolute value less than a threshold, the *edge length*, and sliced Wasserstein-2 is then computed on the accepted samples. This provides a stable central-region transport diagnostic when tail draws dominate finite-sample Wasserstein estimates. We use 6 as the edge length for `8_gaussians` and edge length 8 for `student_uc`.

Expected log marginal (ELM). For the conditioned diffusion benchmark, we report the expected log marginal likelihood,

$$\mathrm{ELM} = \mathbb{E}_{z \sim r}\left[\sum_{j=1}^{d} \log \widehat{q}_{\phi,j}(z_j)\right], \tag{17}$$

where r is the empirical distribution from the SGLD reference samples and $\widehat{q}_{\phi,j}$ is a one-dimensional Gaussian KDE fit to coordinate j of model samples. We follow Pielok et al. [2025] in using this metric. Direct Monte Carlo approximation of the full semi-implicit marginal likelihood is difficult because it integrates over a 100-dimensional mixing distribution, so the coordinate-wise KDE estimate provides a stable diagnostic, though it does not capture the full joint marginal likelihood. The reported evaluation uses 1,000 reference samples and 100,000 model samples; the SGLD baseline uses 100,000 SGLD samples as the model sample set.

BNN predictive metrics. For BNN targets, evaluation draws 2,000 posterior samples (network weights) from the learned variational distribution. RMSE is computed from the posterior predictive mean on the test set. Test log-likelihood is estimated by Monte Carlo marginalization over posterior samples, and the reported NLL is its negative.

B.3 Two-dimensional targets

The main text displays and summarizes `x_shaped`, `student_uc`, and `8_gaussians`. Visual panels use 2,000 samples per method and overlay the samples on target-density contours.

Target density definitions. We now state the precise density for each of the three two-dimensional synthetic targets used in Section 6.1. All targets have $z \in \mathbb{R}^2$.

X-shaped. The `x_shaped` target is an equal-weight mixture of two zero-mean Gaussians whose covariance ellipses cross at the origin, producing an X-shaped density:

$$p(z) = \tfrac{1}{2}\,\mathcal{N}(z;\,0,\,\Sigma_1) + \tfrac{1}{2}\,\mathcal{N}(z;\,0,\,\Sigma_2), \tag{18}$$

where

$$\Sigma_1 = \begin{pmatrix} 2 & 1.8 \\ 1.8 & 2 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 2 & -1.8 \\ -1.8 & 2 \end{pmatrix}.$$

The high off-diagonal correlation ($\rho = \pm 0.9$) in each component creates two narrow ridges that intersect at the origin. Approximating both ridges simultaneously requires the variational family to represent sharp, non-convex geometry.

Student-t (uncorrelated marginals). The `student_uc` target is a bivariate Student-t distribution with $\nu = 2$ degrees of freedom and a fixed scale matrix $A \in \mathbb{R}^{2\times 2}$ with $\det(A) = 1$. Its density is defined through the transformation $\tilde{z} = A^{-1}z$ with independent univariate Student-t marginals:

$$p(z) \;=\; |\det A|^{-1} \prod_{i=1}^{2} \frac{\Gamma\!\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\!\left(\frac{\nu}{2}\right)} \left(1 + \frac{\tilde{z}_i^{\,2}}{\nu}\right)^{-(\nu+1)/2}, \qquad \nu = 2. \tag{19}$$

The matrix A is generated deterministically (seed 50) as $A = (BB^\top)^{-1}/\det\!\big((BB^\top)^{-1}\big)$, where $B_{ij} \sim \mathrm{Uniform}(-1, 1)$. The resulting density has polynomial tails ($\propto \|z\|^{-3}$ marginally), which stress-tests whether the variational approximation can allocate sufficient probability mass far from the mode.

Eight Gaussians. The `8_gaussians` target is an equally weighted mixture of eight isotropic Gaussians arranged uniformly on a circle of radius r:

$$p(z) = \frac{1}{8}\sum_{k=1}^{8} \mathcal{N}\big(z;\,\mu_k,\,\sigma^2 I_2\big), \qquad \mu_k = r\big(\cos(k\pi/4),\,\sin(k\pi/4)\big), \tag{20}$$

with $r = 4$ and $\sigma = 0.5$. The well-separated modes (inter-mode distance ≈ 3.1, versus component standard deviation 0.5) make this a challenging multi-modal target: the variational family must discover and cover all eight modes without spurious bridging density between them.

Table 5 collects the key parameters of each target.

Table 5: Summary of two-dimensional synthetic target densities.

Target	Density form	Parameters	Key challenge
`x_shaped`	$\frac{1}{2}\mathcal{N}(0,\Sigma_1) + \frac{1}{2}\mathcal{N}(0,\Sigma_2)$	$\Sigma_{1,2} = \left(\begin{smallmatrix} 2 & \pm 1.8 \\ \pm 1.8 & 2 \end{smallmatrix}\right)$	Intersecting ridges
`student_uc`	$\lvert A\rvert^{-1}\prod_{i=1}^{2} t_\nu(\tilde{z}_i),\ \ \tilde{z} = A^{-1}z$	$\nu = 2,\ \ \det(A) = 1$	Heavy polynomial tails
`8_gaussians`	$\frac{1}{8}\sum_{k=1}^{8}\mathcal{N}(\mu_k, \sigma^2 I)$	$r = 4,\ \ \sigma = 0.5$	Multi-modal coverage

Training configuration details. SIVI, UIVI, AISIVI, and DIVI use 10,000 iterations, VI learning rate 10^{-3}, batch size 128, and a step scheduler with decay factor 0.7 every 1,000 iterations (0.9 on `8_gaussians`). Linear annealing over 5,000 steps is applied on all targets except `student_uc`, where annealing is disabled for these methods.

Table 6: Toy target final evaluation metrics (mean ± standard error). Wall-clock time is averaged across the three targets; bold indicates the best mean.

Target	Metric	SIVI	KSIVI	AISIVI	UIVI	DIVI
x_shaped	D_{KL}	0.012 ± 0.01	0.054 ± 0.01	0.070 ± 0.02	0.038 ± 0.02	**0.009 ± 0.00**
	W2	0.075 ± 0.03	0.222 ± 0.03	0.191 ± 0.03	0.152 ± 0.04	**0.074 ± 0.01**
student_uc	D_{KL}	**2.676 ± 0.00**	4.512 ± 1.79	2.744 ± 0.06	2.682 ± 0.00	2.677 ± 0.00
	W2 $\lvert x\rvert < 8$	0.124 ± 0.01	0.749 ± 0.63	0.166 ± 0.05	0.145 ± 0.00	**0.086 ± 0.01**
8_gaussians	D_{KL}	**0.236 ± 0.03**	0.830 ± 0.08	1127.982 ± 793.92	0.801 ± 0.01	0.355 ± 0.08
	W2 $\lvert x\rvert < 6$	0.348 ± 0.08	0.569 ± 0.21	0.789 ± 0.24	0.445 ± 0.06	**0.311 ± 0.07**
Wall-clock time (s)		266	377	387	803	**244**

SIVI. SIVI uses 4,096 auxiliary samples from the mixing prior to approximate the semi-implicit marginal density.

AISIVI. AISIVI uses a conditional RealNVP reverse model with 8 coupling layers and hidden width 12. Before joint training, the flow is pre-trained on initial VI samples for 500 warmup epochs with learning rate 10^{-2} and batch size 8,192. During main training, the flow uses 1,024 reverse samples, reverse learning rate 10^{-3}, and one reverse update per VI step.

DIVI. The proxy score MLP has hidden width 64 and three hidden layers. The proxy is warm-started for 500 epochs with learning rate 10^{-3} and batch size 8,192. During main training the proxy uses reverse batch size 2,048, reverse learning rate 10^{-3}, and 5 updates per VI step.

KSIVI. KSIVI is trained for 50,000 iterations with batch size 128, learning rate 10^{-3}, 25,000-step linear annealing, and a step-decay schedule with factor 0.9 every 1,000 iterations, following Cheng et al. [2024]. We use a Gaussian kernel on `x_shaped` and `8_gaussians`, and an affine-invariant Riesz kernel on `student_uc`. On `x_shaped`, the conditional variances are initialized to the constant 1.313; the mean network uses learning rate 10^{-3} and the variance head uses 10^{-5}, with the same decay schedule. On `student_uc`, the median kernel bandwidth is detached during backpropagation, and we apply log probability regularization with coefficient 0.05 during annealing following Cheng et al. [2024].

Quantitative results. Table 6 presents the quantitative results across all five methods. DIVI obtains the lowest mean on four of the six quality metrics and the second-lowest on the remaining two, while SIVI is strongest on the Student UC and eight-Gaussian KL-style metrics. DIVI also has the lowest mean end-to-end wall-clock time. The Wasserstein-2 evaluation follows the protocol in Appendix B.2. We are observing divergence on `8_gaussians` for AISIVI, as a result of drifting support. The same divergence behavior has been observed in the official code of AISIVI, indicating a potential weakness in training stability.

B.4 Conditioned diffusion process

The diffusion posterior is the 100-dimensional discretized path posterior described in Section 6.2. Reference samples are produced by SGLD with 1,000 particles, 100,000 iterations, and step size 10^{-4}. Trace plots show posterior mean paths and marginal 95% posterior intervals against the true latent path and noisy observations.

SIVI, UIVI, AISIVI, and DIVI are trained for 10,000 VI iterations with learning rate 10^{-3}, batch size 128, step scheduler (decay 0.7 every 1,000 steps), and 5,000-step linear annealing. KSIVI is trained for 100,000 iterations.

AISIVI. AISIVI uses a conditional RealNVP with 16 coupling layers and hidden width 256. Before joint training, the reverse flow is pre-trained on samples from the initial variational distribution for 2,000 warmup epochs with learning rate 2×10^{-4} and batch size 128. During main training, the flow uses 256 reverse samples, reverse learning rate 2×10^{-4}, and one reverse update per VI step.

DIVI. The proxy score MLP has hidden width 128 and three hidden layers. The proxy is warm-started for 2,000 epochs with learning rate 10^{-3} and batch size 8,192. During main training the proxy uses reverse batch size 2,048, reverse learning rate 10^{-3}, and 5 updates per VI step.

KSIVI. KSIVI follows the default schedule from Cheng et al. [2024]: 100,000 iterations, 50,000-step linear annealing, Gaussian kernel, and a step-decay schedule with factor 0.9 every 10,000 iterations.

Evaluation. We evaluate using the expected log marginal likelihood defined in Appendix B.2, with 1,000 reference samples from the SGLD chain and 100,000 model samples from each trained variational distribution.

B.5 Bayesian neural-network regression

Each BNN target uses a one-hidden-layer ReLU network with 50 hidden units. The input dimensions are 13 for Boston, 8 for Concrete, 4 for Power, 9 for Protein, 11 for Wine Red, and 6 for Yacht. The datasets can be found on the UCI Machine Learning website [3], and all follow CC BY 4.0 License to the best of our knowledge. Including weights, biases, and two precision variables, the corresponding posterior dimensions are 751, 501, 301, 551, 651, and 401. The datasets contain 506, 1,030, 9,568, 45,730, 1,599, and 308 observations respectively. Data preparation uses a fixed random seed, a 90/10 train-test split, input standardization based on the training split, and target standardization for training labels. Test labels are evaluated on their original scale.

Common BNN protocol. All methods use batch size 128, gradient clipping at norm 1, and a 100-step MSE pretraining stage that initializes the variational sampler toward predictive weights before posterior training begins. For evaluation, we apply exponential moving averages of the VI parameters with decay 0.999, following the stabilized BNN evaluation protocol used by KSIVI and SIVI-SM [Cheng et al., 2024, Yu and Zhang, 2023]. The VI learning rate is selected separately for each method and target from $\{10^{-3}, 10^{-4}, 10^{-5}\}$ by cross-validation, since high-learning-rate divergence depends on both the method and the target posterior.

AISIVI. AISIVI uses a conditional RealNVP with 16 coupling layers and hidden width 256. The reverse flow is pre-trained on initial VI samples for 500 warmup epochs with learning rate 5×10^{-5} and batch size 100. During main training, the flow uses 256 reverse samples, reverse learning rate 5×10^{-5}, one reverse update per VI step, and 10,000-step linear annealing.

DIVI. The proxy score MLP has hidden width 512 and four hidden layers. The proxy is warm-started for 5,000 epochs with learning rate 2×10^{-3} and batch size 2,048. During main training the proxy uses 2 updates per VI step. Linear annealing over 10,000 steps is applied.

KSIVI. KSIVI is trained for 100,000 iterations without annealing, using a Gaussian kernel and log-likelihood regularization term with weight 1, with a step-decay schedule of factor 0.9 every 3,000 iterations, following Cheng et al. [2024].

Evaluation. BNN evaluation draws 2,000 posterior samples and reports test RMSE and NLL as described in Appendix B.2.

B.6 Direct score-estimation evaluation

We evaluate the marginal-score error on fixed variational models obtained after 2,000, 4,000, 6,000, 8,000, and 10,000 DIVI training steps on the X-shaped target. For each model, we generate $N = 1{,}024$ evaluation pairs by sampling $\epsilon_i^{(0)} \sim \mathcal{N}(0, I)$ and then $z_i \sim q_\phi(z \mid \epsilon_i^{(0)})$. Following the reverse-conditional score identity used in UIVI [Titsias and Ruiz, 2019],

$$\nabla_z \log q_\phi(z) = \mathbb{E}_{q_\phi(\epsilon|z)} \left[\nabla_z \log q_\phi(z \mid \epsilon)\right], \qquad q_\phi(\epsilon \mid z) \propto q(\epsilon) q_\phi(z \mid \epsilon), \tag{21}$$

we construct a reference score $s_{\mathrm{HMC}}(z_i)$ by sampling the reverse conditional. For each z_i, we run 20 HMC chains initialized at the generating $\epsilon_i^{(0)}$, with 1,000 burn-in steps and 5,000 retained samples

[3] `https://archive.ics.uci.edu/`

per chain. We pool the retained samples and average their conditional Gaussian scores. The reported HMC diagnostic is $\hat{R} = 1.0065$. The empirical squared score-estimation error is

$$\widehat{\mathcal{E}}^2 = \frac{1}{N}\sum_{i=1}^{N} \|\widehat{s}_{\text{method}}(z_i) - s_{\text{HMC}}(z_i)\|_2^2 . \tag{22}$$

Every estimator is applied to the same fixed DIVI variational model at each checkpoint. SIVI uses 4,096 mixing-prior samples; UIVI uses its native posterior-HMC estimator with five burn-in transitions and five auxiliary samples; AISIVI uses 1,024 importance samples from a normalizing-flow proposal optimized for 10,000 steps separately for each fixed variational distribution; and DIVI uses its checkpointed denoising score network. Table 4 reports score error means and standard errors across five seeds. Score-evaluation time is measured for a batch of 128 over 100 synchronized calls after fitting the score network or proposal.

B.7 Discrepancy from the target score

We also evaluate the squared L^2 discrepancy between each method's estimated variational score and the analytic score of the X-shaped target throughout training. At each checkpoint, we use the variational model obtained by that method and evaluate

$$D_{\text{target}} = \mathbb{E}_{z\sim q_\phi} \|\widehat{s}_{\text{method}}(z) - \nabla_z \log p(z)\|_2^2 . \tag{23}$$

This diagnostic measures the joint effect of variational approximation and score estimation along each method's own training trajectory. Table 7 reports the mean and standard error across five seeds. DIVI obtains the lowest mean target-score discrepancy at every reported checkpoint.

Table 7: Squared L^2 discrepancy from the analytic X-shaped target score along each method's training trajectory (mean $\pm$ standard error over five seeds). Bold indicates the lowest mean.

Method	2K steps	4K steps	6K steps	8K steps	10K steps
SIVI	4.479 ± 1.186	0.546 ± 0.306	0.298 ± 0.248	0.270 ± 0.221	0.190 ± 0.152
UIVI	5.062 ± 0.226	3.265 ± 0.676	3.023 ± 0.276	3.052 ± 0.321	3.040 ± 0.356
AISIVI	1704.846 ± 1662.003	282.647 ± 161.966	581.071 ± 429.919	100.476 ± 27.904	88.381 ± 38.811
DIVI	$\mathbf{3.694 \pm 0.200}$	$\mathbf{0.411 \pm 0.049}$	$\mathbf{0.148 \pm 0.028}$	$\mathbf{0.142 \pm 0.038}$	$\mathbf{0.079 \pm 0.017}$

B.8 Effect and selection of score-update steps and batch size

The number of inner score updates per outer VI update, L_ψ (reverse steps), and the score batch size, B_ψ (reverse batch size), control the accuracy and computational cost of score tracking. DIVI warm-starts ψ from the preceding iteration, so each inner loop tracks the score of the current variational distribution. The main experiments use two inner updates for BNN regression and five for the toy and conditioned diffusion targets. The appropriate choice depends on dimension, network capacity, batch size, and how rapidly q_ϕ changes. A practical selection rule starts with a larger update count, such as $L_\psi = 10$, and halves it while monitoring a distributional metric such as the ELBO or Wasserstein-2 distance. The selected value balances distributional accuracy and computation. For batch-size selection, $B_\psi = 2{,}048$ provides a starting point; larger batches can be evaluated using the same distributional metric within the available memory budget.

Tables 8 and 9 give ablations on the X-shaped target over five seeds, with the remaining training settings fixed. The first varies L_ψ with $B_\psi = 2{,}048$; the second varies B_ψ with $L_\psi = 5$. Increasing the update count improves both metrics and increases runtime. Larger score batches improve both metrics with similar measured runtimes over the tested range. The main toy experiments use $L_\psi = 5$ and $B_\psi = 2{,}048$.

B.9 Comparison with a RealNVP variational family

Normalizing-flow VI uses invertible transformations with tractable densities and scores. Semi-implicit VI represents complex multimodal structure through an implicit mixture. DIVI learns the marginal score of this family. As a concrete comparison, we train a four-layer RealNVP [Dinh et al., 2016]

Table 8: Effect of reverse steps on the X-shaped target with reverse batch size 2,048 (mean $\pm$ standard error over five seeds). Bold indicates the lowest mean quality metric.

Reverse steps	$D_{\mathrm{KL}} \downarrow$	W2 $\downarrow$	Runtime (s)
1	0.2083 ± 0.1503	0.2517 ± 0.0526	171.6 ± 1.1
2	0.02253 ± 0.00725	0.2402 ± 0.0375	216.5 ± 3.4
5 (default)	0.004619 ± 0.001524	0.1127 ± 0.0279	352.8 ± 4.8
10	$\mathbf{0.002708 \pm 0.000824}$	$\mathbf{0.07333 \pm 0.00553}$	571.7 ± 7.2

Table 9: Effect of reverse batch size on the X-shaped target with five reverse steps (mean $\pm$ standard error over five seeds). Bold indicates the lowest mean quality metric.

Reverse batch size	$D_{\mathrm{KL}} \downarrow$	W2 $\downarrow$	Runtime (s)
512	0.01058 ± 0.00520	0.1592 ± 0.0387	339.6 ± 3.6
1,024	0.008070 ± 0.001921	0.1480 ± 0.0249	339.4 ± 5.5
2,048 (default)	0.004619 ± 0.001524	0.1127 ± 0.0279	352.8 ± 4.8
4,096	0.004079 ± 0.001026	0.07773 ± 0.01072	351.1 ± 4.1
8,192	$\mathbf{0.001978 \pm 0.000539}$	$\mathbf{0.06823 \pm 0.00606}$	344.7 ± 5.2

for 10,000 optimization steps with batch size 128 on the eight-Gaussian target and aggregate the results across five seeds. Table 10 reports ELBO, Wasserstein-2 distance, and parameter count. DIVI achieves a higher mean ELBO and a lower mean Wasserstein-2 distance in this benchmark with fewer parameters.

Table 10: DIVI and a four-layer RealNVP variational family on the eight-Gaussian target (mean $\pm$ standard error over five seeds). Bold indicates the best mean or lowest parameter count.

Method	ELBO $\uparrow$	W2 $\downarrow$	Parameter count
DIVI	$\mathbf{-0.355 \pm 0.082}$	$\mathbf{0.327 \pm 0.060}$	**26,054**
RealNVP	-0.374 ± 0.006	0.402 ± 0.053	137,236

C Numerical Validations

In this section, we provide empirical diagnostics supporting Assumption 2 and Assumption 3 through numerical experiments along the observed training trajectories.

For Assumption 2, we provide empirical diagnostics by tracking the fourth moment of scores, both for the reverse model and for the target score. We plot the mean value of the fourth moment of norms of the reverse model score $\mathbb{E}_{z\sim q_\phi}\|s_\psi(z)\|^4$ as the solid line, and the minimum and maximum values as the shaded regions in Figure 3, with each of the values calculated across 10 independent seeds. The same plot for fourth moment of the target score $\mathbb{E}_{z\sim q_\phi}\|\nabla_z \log p(\mathcal{D}, z)\|^4$ is shown in Figure 4. The figures provide empirical evidence that there exists a uniform bound on the 4th moment of the norm of the scores over the training trajectories, consistent with Assumption 2.

For Assumption 3, we report the calculated $\max\left\{\mathbb{E}_\epsilon\|\nabla_\phi \sigma_\phi(\epsilon)\|_2^4, \mathbb{E}_\epsilon\|\nabla_\phi \mu_\phi(\epsilon)\|_2^4\right\}$ for each checkpoint respectively throughout the training process, and aggregate across 10 independent seeds. We track the metrics for the toy `x_shaped` and Conditioned Diffusion target distribution. The results are shown in Figure 5, where the solid line indicates the mean while the shaded region demonstrates the minimum and maximum values. These diagnostics suggest that we can obtain a uniform upper bound for the expected ℓ_2 norm's fourth moment of the VI networks' gradient, justifying Assumption 3.

We also provide empirical evidence to the alternative version of Assumption 2, as present in Appendix A.3. In Appendix A.3 we proved that Assumption 2 can be derived from another linear growth condition $\max\{\|\nabla_z \log p(\mathcal{D}, z)\|, \|s_\psi(z)\|\} \le \lambda(\|z\| + 1)$ together with a bounded fourth-moment condition $\max\{\mathbb{E}_\epsilon\|\mu_\phi(\epsilon)\|^4, \mathbb{E}_\epsilon\|\sigma_\phi(\epsilon)\|^4\} \le \zeta$, both of which we assume to hold uniformly over ϕ. As empirical evidence, we demonstrate for three checkpoints taken at 2000, 5000 and 10000 iterations, that $\log\max\{\|\nabla_z \log p(\mathcal{D}, z)\|, \|s_\psi(z)\|\} - \log(\|z\| + 1)$ is bounded in Figure 6. The figure demonstrates different values for samples of z drawn with uniform distribution in $\|z\|$.

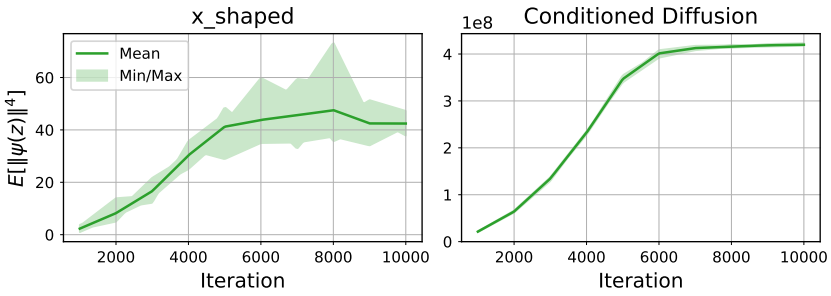

Figure 3: Fourth moment of the norm of the approximated score $\mathbb{E}_{z\sim q_\phi}\|s_\psi(z)\|^4$ throughout DIVI training for `x_shaped` and the conditioned diffusion benchmark, aggregated over 10 seeds.

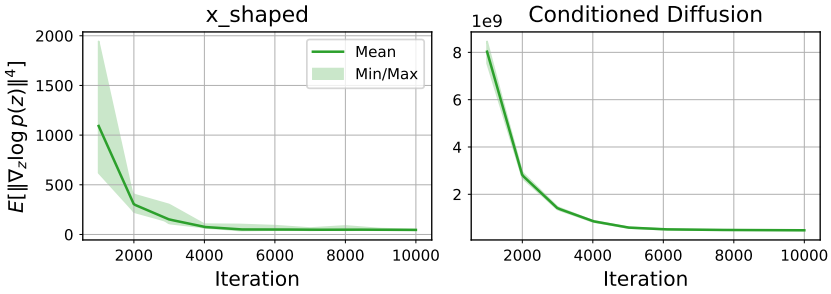

Figure 4: Fourth moment of the norm of the target score $\mathbb{E}_{z\sim q_\phi}\|\nabla_z \log p(\mathcal{D}, z)\|^4$ throughout DIVI training for `x_shaped` and the conditioned diffusion benchmark, aggregated over 10 seeds.

We also plot the 4th moment of the VI model checkpoints $\max\{\mathbb{E}_\epsilon\|\mu_\phi(\epsilon)\|^4, \mathbb{E}_\epsilon\|\sigma_\phi(\epsilon)\|^4\}$ using 10 different seeds and plot them along the 10 checkpoints we obtained during the DIVI training regime in Figure 7. The solid line indicates the 10-seed-average of the fourth moments, while the shaded regions indicate the minimum and maximum of recorded values. These plots together demonstrate empirical evidence to support the alternate version of Assumption 2 in Appendix A.3.

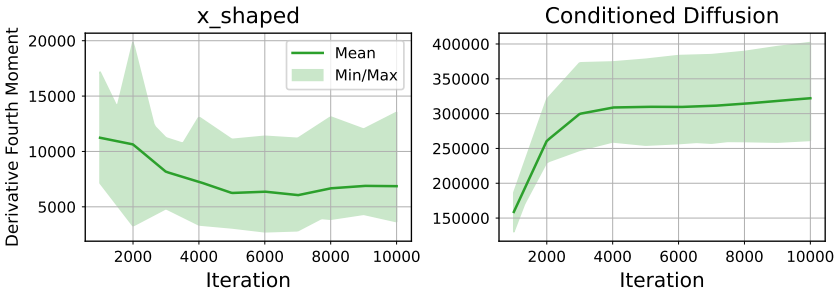

Figure 5: The expected ℓ_2 norm's fourth moment of the VI model gradient throughout the DIVI training process for `x_shaped` and Conditioned Diffusion experiments, aggregated across 10 different seeds. This supports Assumption 3 that the VI model gradients are bounded.

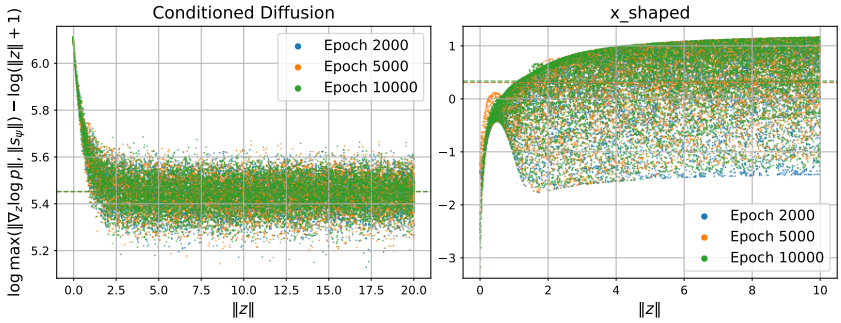

Figure 6: $\log \max\{\|\nabla_z \log p(\mathcal{D}, z)\|, \|s_\psi(z)\|\} - \log(\|z\| + 1)$ over different z's, ϕ's and ψ's throughout DIVI training for `x_shaped` and the conditioned diffusion benchmark.

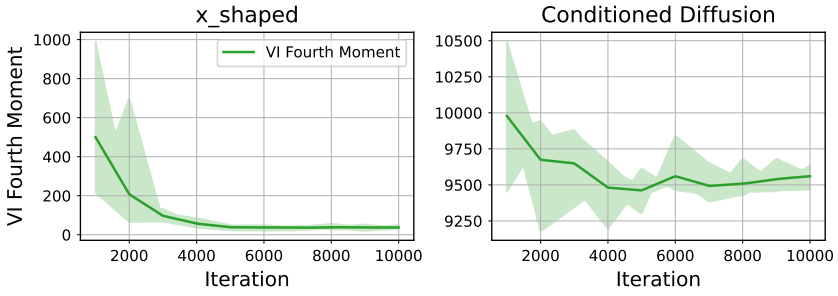

Figure 7: Fourth moment of the VI model reparameterization norm $\max\{\mathbb{E}_\epsilon\|\mu_\phi(\epsilon)\|^4, \mathbb{E}_\epsilon\|\sigma_\phi(\epsilon)\|^4\}$ throughout DIVI training for `x_shaped` and the conditioned diffusion benchmark, aggregated over 10 seeds.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state our contributions clearly: a novel semi-implicit variational inference method utilizing denoising score matching, a theoretical convergence analysis (in Section 4) justifying the method's rigor and experiment results (in Section 6) that demonstrate the efficiency and effectiveness of our method.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have acknowledged the limitations in Section 7. Theoretically, the convergence of our method relies on the smoothness of the target optimization objective and the boundedness of the optimization trajectory, while the capacity of the score networks may leave non-vanishing residues for score-estimation error. We also identify enhancing the performance under complex multimodal settings as a potential future direction. We have made the assumptions we used in the theoretical analysis explicit in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See Section 4 and Appendix A, where we provide complete assumptions and proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementaion is publicly available at `https://github.com/VincentWangzt/DIVI`

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have included all our code and a complete experiment results reproducing guideline in the supplementary materials.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Justification: See Section 6 and Appendix B for the experimental settings and evaluation procedures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For each experiment setup, we have repeated the experiments for 10 different random seeds and reported the standard error for all of the key evaluation metrics.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have reported the average run time for each method in our experiments, and we have also reported the type of GPU and CPU workers used to perform such experiments. The entire reproduction of experiments is estimated to complete within the scale of a hundred GPU hours. The preliminary experiments used to sweep the appropriate hyperparameter is in the scale of tens of GPU hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. The work is a Bayesian inference method that aims to provide better performance and efficiency. No human subjects or sensitive personal data are involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is a Bayesian inference method that lies within the traditional machine learning framework, with no significant societal impacts within the view of the authors.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper proposes a novel method of semi-implicit variational inference, with no checkpoints or curated data that are of significant importance.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: We have included the BNN dataset details, including the publicly available links and acknowledged the licenses in Appendix B.5. Other datasets we used are all synthetic and have also cited the proper original proposers, to the best of our knowledge.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper introduces a novel semi-implicit method, with no new assets released or used in the experiments.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper proposes a novel semi-implicit variational inference method, no human subjects have been involved in any sense.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We have not incorporated LLMs as a core contributor in the development of our methods, apart from paper writing and coding for experiments and result post-processing. We have carefully reviewed the NeurIPS handbook on LLM policy and strictly adhered to the instructions.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.