ONSAGER–MACHLUP POSTERIOR TRANSPORT FOR DEEP GAUSSIAN PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Approximate inference over inducing variables is the central computational bottleneck of Deep Gaussian Processes (DGPs). Existing methods either fit an explicit density $q_\phi(\mathbf{U})$ by an ELBO (DSVI, IPVI, DDVI, DBVI) or sample by MCMC (SGHMC). We instead frame DGP inference as *posterior transport*: learn a deterministic sampler that maps a tractable reference measure to posterior-relevant inducing variables, regularised by a path prior derived from the Doob-bridged reference diffusion. Our realisation, **OM-Path** (formally FBVI-bridge-Path), applies Song's probability-flow ODE to DBVI's Doob-bridged forward SDE; the reference drift is closed-form from the bridge marginal coefficients (no score matching) and the path regulariser is the **Onsager–Machlup action**. At the finite ϵ used in training, the objective is a data-fit term plus the classical Onsager–Machlup small-tube path prior; Theorem 1 identifies its small-noise limit with the conditional MAP path of the corresponding tempered Doob-bridge path posterior. Two strict path-space ELBO variants on the same backbone are derived as ablations. Under a matched-seed paired Wilcoxon test against DBVI on seven UCI regression benchmarks, OM-Path wins significantly on the two largest datasets (*power* $p = 0.007$, NLL $\mathbf{0.009}$ against the DSVI baseline of 0.017; *protein* $p = 0.002$, RMSE $\mathbf{0.718}$ vs. 0.761), ties on *yacht* / *qsar*, and concedes the small-N noisy cells to DBVI. It also outperforms the strict-ELBO variants on $13/14$ UCI cells, and our gradient-variance measurements show a substantial Hutchinson penalty on *power*, though the size of that effect is dataset-dependent.

1 INTRODUCTION

Deep Gaussian Processes (DGPs; Damianou & Lawrence (2013)) stack Gaussian-process layers to obtain a Bayesian deep model with calibrated uncertainty. Inference is intractable because the posterior over the per-layer inducing variables $\mathbf{U} = \{U^{(\ell)}\}_{\ell=1}^{L}$ is non-Gaussian and highly correlated across layers. Doubly-stochastic mean-field VI (DSVI; Salimbeni & Deisenroth (2017)) sidesteps this by enforcing a Gaussian factorisation with closed-form KL—a strong restriction whenever the true posterior is multimodal or heavily correlated across the M inducing points. Three published lines remove that restriction: GAN-style implicit VI (IPVI; Yu et al., 2019), SG-HMC sampling (Havasi et al., 2018), and the score-based ELBO methods DDVI (Xu et al., 2024; reverse VP-SDE + denoising score matching) and DBVI (Xu et al., 2026; Doob-bridged reverse SDE + conditional DSM). A separate body of work builds *general* unnormalised-posterior samplers from controlled SDEs—PIS (Zhang & Chen, 2022), DDS (Vargas et al., 2023), DGFS (Zhang et al., 2024), SGDS (Kim et al., 2025), NFS^2 (Chen et al., 2025)—but targets generic densities and has not been instantiated for hierarchical-GP inducing variables.

This paper: posterior transport. We propose *posterior transport* for DGP inducing variables: the variational object is not a density $q_\phi(\mathbf{U})$ fit by an ELBO, but a *deterministic sampler* $T_\phi : p_{\mathrm{ref}} \to$ posterior-relevant samples realised by an ODE on $\mathcal{C}([0,1];\mathbb{R}^M)$. The reference measure is the noise-side marginal of a Doob-bridged forward diffusion, and the sampler is regularised by a *path prior* derived from the reference probability flow ODE. Two consequences follow. First, Song et al. (2021)'s probability flow ODE applied to DBVI's Doob-bridged forward SDE gives a *closed-form* reference drift $v_{\mathrm{ref}}^{\mathrm{Bri}}$ in terms of the bridge mean and variance coefficients (Lemmas 1–2), so no score matching is needed. Second, the path regulariser cannot be a Girsanov KL: for two SDEs sharing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

diffusion g Girsanov gives a finite path-measure KL, but as $g \to 0$ the deterministic and stochastic path measures become mutually singular and the KL blows up. The correct deterministic-limit object is the *Onsager–Machlup action*, the Freidlin–Wentzell small-noise rate functional (Freidlin & Wentzell, 1998), which with a data NLL gives our training loss

$$\mathcal{L}_{\mathrm{OM}}(\phi) \;=\; -\mathbb{E}_{q_\phi}\big[\log p(\mathbf{y} \mid \mathbf{F}^{(L)})\big] \;+\; \tfrac{1}{2}\int \mathbb{E}\big[\|v_\phi(U_\tau, 1-\tau, \mathrm{ctx}) - \bar{v}_{\mathrm{ref}}(U_\tau, \tau)\|^2\big]\,\mathrm{d}\tau, \quad (1)$$

where $\bar{v}_{\mathrm{ref}}$ is the reverse-time reference drift defined in Eq. 7. This loss is *trace-free*: no denoising score matching (DDVI/DBVI), no Hutchinson trace (FFJORD), no Stein control variate for the partition-function gradient (NFS^2).

Theoretical status. Two statements should be kept apart. *(i) Finite ϵ (what we optimise).* At any $\epsilon > 0$, S_{OM} is the classical Onsager–Machlup functional of the path under the Gaussian reference path measure—i.e. the small-tube log-probability ratio of Ikeda & Watanabe (1989), a path-prior term, *not* a density with respect to any dominating measure on an infinite-dimensional path space. $\mathcal{L}_{\mathrm{OM}}$ is then a data fit plus that path prior. *(ii) $\epsilon \to 0$ (the theorem).* Take the reference diffusion $\mathrm{d}U_\tau = \bar{v}_{\mathrm{ref}}(U_\tau, \tau)\,\mathrm{d}\tau + \epsilon\,\mathrm{d}W_\tau$, and tilt it by the endpoint likelihood raised to $1/\epsilon^2$. The Freidlin–Wentzell LDP yields the good rate functional $I_y(u) = -\log p(\mathbf{y}|u_1) + S_{\mathrm{OM}}(u; \bar{v}_{\mathrm{ref}})$ on $[0,1]$, whose minimiser is exactly equation 1: a *small-noise conditional MAP identification* (Theorem 1). The framing is rigorous but is path-space MAP rather than an ELBO (Remark 1). One practical gap remains and we state it plainly: the path term is estimated under the reference bridge marginal rather than under the learned trajectory law (Eq. 12). We also derive two strict path-space ELBO variants on the *same* bridge backbone—**FBVI-bridge-CNF** (FFJORD change-of-variables) and **FBVI-bridge-CNFOM** (FFJORD log-det + OM regulariser)—which are honest lower bounds on $\log p(\mathbf{y}|\mathbf{x})$ yet *lose* to the trace-free MAP objective on 13 of the 14 UCI cells. Our gradient-variance measurements support a substantial Hutchinson penalty on *power*, while the size of the effect is dataset-dependent (Appendix R); the qualitative pattern is the familiar one (score matching, contrastive divergence, consistency models) of low-variance surrogates beating exact-likelihood objectives.

Contributions.

1. **Posterior transport for DGP inducing variables**, realised by a deterministic ODE whose reference drift is closed-form from the Doob-bridge marginal via Song's probability flow ODE (Section 4.1).
2. **A small-noise conditional-MAP identification, with an explicit finite-α reading** (Theorem 1, Remark 2). It identifies $\mathcal{L}_{\mathrm{OM}}$ with the $\epsilon \to 0$ conditional MAP path of a tempered Doob-bridge path posterior, certifying S_{OM} as that measure's Freidlin–Wentzell rate functional and fixing the relative weight of the two terms at 1. At the $\alpha = 1$ we actually train with, S_{OM} is the classical Onsager–Machlup small-tube functional of the path—a path prior, not a density on path space—and the sampler still produces a non-degenerate endpoint distribution, which is what the BO acquisition (Section 5.5) and the heteroscedastic toy (Appendix T) consume.
3. **Strict-ELBO comparisons** (CNF, CNFOM) on the same backbone, quantifying the practical cost of the Hutchinson trace (Section 4.3, Appendix J).
4. **Empirical results.** Under matched-seed paired Wilcoxon testing, FBVI-bridge-Path wins significantly over DBVI on the two largest UCI datasets (*power* $p = 0.007$, *protein* $p = 0.002$), ties on *yacht* / *qsar*, concedes the small-N noisy cells, is mean-best on all six large-regression cells (up to 515k rows) and in the tied-best group on SUSY / HIGGS.

2 Related Work

Inducing-variable VI for DGPs. Our baselines are DSVI (Salimbeni & Deisenroth, 2017), SG-HMC (Havasi et al., 2018), IPVI (Yu et al., 2019) and the score-based ELBO methods DDVI (Xu et al., 2024) (reverse-time VP-SDE trained by denoising score matching) and DBVI (Xu et al., 2026) (a Doob bridge anchoring the SDE at a data-conditioned initial mean, with a conditional score trained by conditional DSM). Both maximise a Girsanov-form ELBO: they are density-VI methods whose variational density is implicit in an SDE terminal, not posterior transport in the sense used here.

General unnormalised posterior samplers. PIS (Zhang & Chen, 2022), DDS (Vargas et al., 2023), DGFS (Zhang et al., 2024), SGDS (Kim et al., 2025) and NFS^2 (Chen et al., 2025) build controlled-SDE samplers for generic unnormalised targets. They are not formulated for hierarchical-GP inducing variables, but their path-space view (Girsanov KL against a fixed reference SDE) is the closest precedent for our deterministic-ODE transport, which replaces that KL by the Freidlin–Wentzell rate functional. NFS^2 in particular keeps density tracking through a continuity-equation residual and needs velocity-driven SMC plus a Stein control variate for $\partial_t \log Z_t$; the MAP framing avoids both.

Theoretical anchors and OM action in generative modelling. Theorem 1 draws on path-integral control (Kappen, 2005), Schrödinger bridges (Léonard, 2014) and the Freidlin–Wentzell framework (Freidlin & Wentzell, 1998). Kretschmann (2023) shows that OM-functional minimisers are *strong posterior modes* under suitable regularity—the identification we use—and analyses the finite-ϵ Ikeda–Watanabe formulation matching Remark 2. Concurrently, Raja et al. (2025) minimises the OM action over trajectories of a *pre-trained* diffusion or flow model for transition-path sampling, with the learned score as reference drift and fixed endpoints; we instead use it as a *training* objective with a closed-form Doob-bridge PF-ODE drift and an endpoint Bayesian likelihood. Flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023), its few-step variants (Frans et al., 2025) and its use in samplers (Yoon & Lee, 2024) are the deterministic counterpart to score-based diffusion; ours is, to our knowledge, the first probability-flow-ODE posterior transport for hierarchical GPs.

3 Background

DGP with inducing variables. A DGP with L layers defines $\mathbf{f}^{(\ell)} \sim \mathcal{GP}(0, k^{(\ell)})$ with $\mathbf{F}^{(\ell)} = \mathbf{f}^{(\ell)}(\mathbf{F}^{(\ell-1)})$ and $\mathbf{F}^{(0)} = \mathbf{x}$. For scalability one introduces M inducing locations $\mathbf{Z}^{(\ell)} \in \mathbb{R}^{M \times d_{\ell-1}}$ and values $\mathbf{U}^{(\ell)} \in \mathbb{R}^{M \times d_\ell}$ with prior $p(\mathbf{U}^{(\ell)}) = \mathcal{N}\Big(0, K^{(\ell)}_{\mathbf{ZZ}}\Big)$ and sparse-GP conditional $p(\mathbf{F}^{(\ell)} \mid \mathbf{F}^{(\ell-1)}, \mathbf{U}^{(\ell)})$, and maximises

$$\mathcal{L}(\theta, \phi) \;=\; \mathbb{E}_{q_\phi(\mathbf{U})}\big[\log p(\mathbf{y} \mid \mathbf{F}^{(L)})\big] \;-\; D_{\mathrm{KL}}\big(q_\phi(\mathbf{U}) \,\|\, p(\mathbf{U})\big). \tag{2}$$

DSVI sets $q(\mathbf{U}^{(\ell)}) = \mathcal{N}\big(m_\ell, L_\ell L_\ell^\top\big)$, making the KL analytic at the cost of the Gaussian restriction.

DDVI and DBVI. Xu et al. (2024) instead define $q(\mathbf{U})$ as the terminal of a reverse-time variance-preserving SDE, $\mathrm{d}U_t = [-\frac{1}{2}\beta(t)U_t - \beta(t)s_\phi(U_t, t)]\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\bar{W}_t$ from $U_T \sim \mathcal{N}(0, \mathbf{I})$, training s_ϕ by denoising score matching. Xu et al. (2026) condition this SDE on a learned initial distribution via Doob's h-transform: an amortiser $\mu_\theta(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^M$ gives a data-anchored $p_0(U_0 \mid \mathbf{x}) = \mathcal{N}\big(\mu_\theta(\mathbf{x}), \sigma_0^2\mathbf{I}\big)$, the forward SDE gains the drift correction $g(t)^2 h(U_t, t, U_0)$, and the score becomes conditional. Under affine drift the bridge marginal $p_t^{\mathrm{Bri}}(U_t \mid \mathbf{x}) = \mathcal{N}(m_t, \kappa_t\mathbf{I})$ is closed form (Xu et al., 2026, Prop. 2). Both maximise a Girsanov-form ELBO, so both are density-VI methods whose variational density is implicit in an SDE terminal. Appendix C gives a self-contained review.

Flow matching. Flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023) learns a velocity field by regression onto a target velocity u_t on a probability path p_t and samples by integrating $\mathrm{d}x_t = v_\phi(x_t, t)\mathrm{d}t$. Its use for variational inference in hierarchical Bayesian models has been limited, in part because there is no obvious analogue of denoising score matching when the target is an intractable posterior—a gap the present objective fills.

4 Method

Section 4.1 reduces the bridge SDE to its probability flow ODE and gives the closed-form reference drift; Section 4.2 adds the Doob bridge anchor and derives the Onsager–Machlup posterior-transport objective (Theorem 1); Section 4.3 contrasts it with two strict path-space ELBO variants on the same backbone.

4.1 FROM BRIDGE SDE TO PROBABILITY FLOW ODE

Setup. Following DBVI (Xu et al., 2026), consider the Doob h-transformed forward bridge SDE in bridge-time $s \in [0, 1]$:

$$\mathrm{d}U_s = b(U_s, s)\,\mathrm{d}s + g\,\mathrm{d}W_s, \quad b(U_s, s) = -\lambda(s)\,U_s + g^2 h(U_s, s, U_0), \tag{3}$$

where h is the Doob h-transform of Xu et al. (2026, Prop. 1) and $U_0 \sim p_0^\theta(\cdot|\mathbf{x}) = \mathcal{N}\big(\mu_\theta(\mathbf{Z}), \sigma_0^2 I\big)$ is amortised by a per-layer net μ_θ. Under affine drift the marginal at every s stays Gaussian (Prop. 1).

Lemma 1 (Score-based probability-flow-ODE equivalence; Song et al., 2021)**.** *Let p_s denote the marginal at time s of the SDE $\mathrm{d}U_s = b(U_s, s)\,\mathrm{d}s + g\,\mathrm{d}W_s$. The deterministic ODE*

$$\frac{\mathrm{d}U_s}{\mathrm{d}s} \;=\; b(U_s, s) - \tfrac{1}{2} g^2\,\nabla_U \log p_s(U_s) \tag{4}$$

shares the same marginals $\{p_s\}_{s\in[0,T]}$ as the original SDE (proof: Appendix D.1).

Lemma 2 (Closed-form probability-flow drift for Gaussian marginals)**.** *If $p_s(U) = \mathcal{N}(U;\, m_s,\, \kappa_s I)$ is Gaussian with smooth m_s, κ_s, then the probability-flow-ODE drift simplifies to the linear field*

$$v_{\mathrm{ref}}(U, s) \;=\; \dot{m}_s + \frac{\dot{\kappa}_s}{2\,\kappa_s}\big(U - m_s\big). \tag{5}$$

Proof: Appendix D.2.

Combining Lemma 1 and Lemma 2 with Prop. 1's bridge marginal $p_s^{\mathrm{Bri}} = \mathcal{N}(\phi(s)\mu_\theta(\mathbf{Z}), \kappa(s)I)$ gives the central closed-form expression we use throughout:

$$\boxed{v_{\mathrm{ref}}^{\mathrm{Bri}}(U, s) \;=\; \dot{\phi}(s)\,\mu_\theta(\mathbf{Z}) \;+\; \frac{\dot{\kappa}(s)}{2\,\kappa(s)}\,\big(U - \phi(s)\,\mu_\theta(\mathbf{Z})\big).} \tag{6}$$

This is the reference probability-flow drift of the Doob bridge, in *forward* bridge time s: integrated forward it pushes the data-anchored start to the noise-side marginal while preserving the bridge marginals, and it depends on $\mu_\theta(\mathbf{Z})$ but not on $\mathbf{y}$ — bridge geometry, not posterior. Our sampler runs in reverse time $\tau = 1 - s$, so, once and for all, we set

$$\bar{v}_{\mathrm{ref}}(U, \tau) \;:=\; -\,v_{\mathrm{ref}}^{\mathrm{Bri}}(U,\, 1 - \tau) \qquad (\text{chain rule } \mathrm{d}/\mathrm{d}\tau = -\mathrm{d}/\mathrm{d}s). \tag{7}$$

Every path-space expression below uses $\bar{v}_{\mathrm{ref}}$, so no further sign or time-argument bookkeeping is needed.

A bridge-free variant, **vanilla FBVI**, which pushes the GP prior $\mathcal{N}(0, K_{\mathbf{ZZ}})$ forward through a learned ODE with no data anchoring, is described in Appendix A.1; it is the deterministic analogue of DDVI and we carry it through the experiments as an ablation.

4.2 FBVI-BRIDGE-PATH: BRIDGE ANCHORING + ONSAGER–MACHLUP POSTERIOR TRANSPORT

Doob bridge anchoring. We replace the unconditional prior $\mathcal{N}(0, K_{\mathbf{ZZ}})$ as the base of the sampler ODE with the noise-side marginal of the forward Doob bridge:

$$U_0 \sim \mathcal{N}(\phi(1)\mu_\theta(\mathbf{Z}), \kappa(1)I)\,, \qquad \frac{\mathrm{d}U_\tau}{\mathrm{d}\tau} = v_\phi^{(\ell)}\big(U_\tau,\, 1 - \tau,\, \mu_\theta(\mathbf{Z})\big), \qquad \widehat{\mathbf{U}}^{(\ell)} := U_1, \tag{8}$$

in reverse-time $\tau = 1 - s \in [0, 1]$. The base is now *data-anchored* through the amortiser $\mu_\theta(\mathbf{Z})$ with mean attenuation $\phi(1)$ and isotropic variance $\kappa(1)$ from Prop. 1; we obtain $\kappa(1) \approx 0.50$ in our default setting (Prop. 2), so the bridge halves the initial variance relative to vanilla FBVI.

Posterior transport, not density VI. A learned ODE implicitly defines an endpoint density $q_\phi(\widehat{\mathbf{U}})$, but the object we parameterise is the *transport sampler* $T_\phi\colon U_0 \mapsto U_1$, not its density. Density VI regularises the endpoint by $\mathrm{KL}(q_\phi \| p)$ and so needs either a tractable density (DSVI) or a log-det estimator (FFJORD/Hutchinson); we instead place a prior on the *path* traced by the sampler, derived from the reference Doob-bridge probability flow. The data term selects posterior-relevant endpoints; the path prior penalises deviation from the reference bridge.

Tempered Doob-bridge path posterior. Fix the reference diffusion $\mathrm{d}U_\tau = \bar{v}_{\mathrm{ref}}(U_\tau, \tau)\,\mathrm{d}\tau + \epsilon\,\mathrm{d}W_\tau$ with $U_0 \sim \mathcal{N}(\phi(1)\mu_\theta(\mathbf{Z}), \kappa(1)I)$ and path measure $\mathbb{P}^\epsilon_{\mathrm{ref}}$ on $\mathcal{C}([0,1];\mathbb{R}^M)$; tilting by the endpoint likelihood at inverse temperature $\beta = 1/\epsilon^2$ gives a *tempered path posterior*

$$\frac{\mathrm{d}\mathbb{P}^\epsilon_y}{\mathrm{d}\mathbb{P}^\epsilon_{\mathrm{ref}}}(U_{0:1}) \;=\; \frac{1}{Z^\epsilon}\, p\big(\mathbf{y} \mid \mathrm{dgp}(U_1)\big)^{1/\epsilon^2}, \qquad Z^\epsilon = \mathbb{E}_{\mathbb{P}^\epsilon_{\mathrm{ref}}}\big[p(\mathbf{y}|U_1)^{1/\epsilon^2}\big]. \tag{9}$$

This is Bayes' rule on path space, with $\beta = 1/\epsilon^2$ the standard rescaling that keeps likelihood and reference-prior contributions of the same order under small-noise asymptotics. By (A1) the reference drift is Lipschitz on the whole of $[0,1]$ — the h-transform singularity at the anchor cancels in the coefficient ODEs — so no endpoint cutoff is needed and the LDP below applies to $\mathbb{P}^\epsilon_y$ itself, on the same interval $[0,1]$ that the sampler integrates.

The small-noise argument needs three regularity conditions: a reference drift that is Lipschitz in U uniformly on $[0,1]$, a Lipschitz DGP forward with a Varadhan tail bound, and the standard uniform-topology Freidlin–Wentzell LDP. They are stated in full as **(A1)–(A3)** in Appendix A.3, together with a per-dataset numerical check of the tail condition.

Theorem 1 (Small-noise conditional MAP path)**.** *Under (A1)–(A3), let $\mathbb{P}^\epsilon_y$ be the tempered Doob-bridge path posterior of equation 9 and $\bar{v}_{\mathrm{ref}}$ the reverse-time reference drift of equation 7. Then, conditional on the initial state $U_0 = u_0$, the $\epsilon \to 0$ MAP path of $\mathbb{P}^\epsilon_y$ solves*

$$u^\star(u_0) \;=\; \arg\min_{\substack{u \in H^1([0,1];\mathbb{R}^M) \\ u_0 \text{ given}}} \Big\{ -\log p(\mathbf{y} \mid u_1) \;+\; \underbrace{\tfrac{1}{2}\int_0^1 \big\|\dot{u}_\tau - \bar{v}_{\mathrm{ref}}(u_\tau, \tau)\big\|^2 \,\mathrm{d}\tau}_{S_{\mathrm{OM}}(u;\, \bar{v}_{\mathrm{ref}})} \Big\}. \tag{10}$$

Both terms carry unit weight: the $1/\epsilon^2$ tempering multiplies the whole bracket, so it cancels in the argmin.

Rather than solve equation 10 for each u_0, we share one velocity field via $\dot{u}_\tau = v_\phi(u_\tau, 1-\tau, \mathrm{ctx})$ and minimise the u_0-expectation of the objective in braces. This amortisation is a modelling restriction, not part of the theorem: $\phi^\star$ is optimal only within the ODE-parameterised submanifold of H^1 paths, and the resulting gap—which may shrink with increasing model capacity under suitable approximation assumptions—is slack separate from the LDP (Appendix A.3). Substituting the ODE trajectory gives the practical loss

$$\boxed{\mathcal{L}_{\mathrm{OM}}(\theta,\phi) \;=\; -\,\mathbb{E}_{q_\phi}\big[\log p(\mathbf{y} \mid \mathbf{F}^{(L)})\big] \;+\; \tfrac{\alpha}{2}\int_0^1 \mathbb{E}_{U_\tau \sim q^\phi_\tau}\Big[\big\|v_\phi(U_\tau, 1-\tau, \mathrm{ctx}) \;-\; \bar{v}_{\mathrm{ref}}(U_\tau, \tau)\big\|^2\Big]\,\mathrm{d}\tau,} \tag{11}$$

where q^ϕ_τ is the law of the learned trajectory at time τ and the velocity net takes the forward bridge time $1-\tau$ as its time argument.

Two points about equation 11 must be stated precisely. **(i) α is not a temperature.** Theorem 1 puts unit weight on both terms — the $1/\epsilon^2$ tempering multiplies the whole bracket in equation 10 and cancels in the argmin — so the MAP objective is exactly $\alpha = 1$, which is what every experiment uses. We retain a scalar α only as a regularisation-strength knob; no limit in α recovers the theorem, and the sweep in Appendix Q is a robustness check around the prescribed value, not a temperature schedule. **(ii) How the path term is estimated.** We draw the regulariser state in closed form from the bridge marginal, giving (with $s = 1-\tau$)

$$\widehat{\mathcal{R}}(\phi) = \tfrac{\alpha}{2}\int_0^1 \mathbb{E}_{U_s \sim p^{\mathrm{Bri}}_s}\Big[\big\|v_\phi(U_s, s, \mathrm{ctx}) + v^{\mathrm{Bri}}_{\mathrm{ref}}(U_s, s)\big\|^2\Big]\,\mathrm{d}s, \quad p^{\mathrm{Bri}}_s = \mathcal{N}(\phi(s)\mu_\theta(\mathbf{Z}), \kappa(s)I)\,. \tag{12}$$

which is exact at the configuration that zeroes the regulariser (Prop. 3) and needs no trace, log-det or partition-function estimate.

Remark 1 (Theoretical status: rigorous, but not an ELBO). Theorem 1 rigorously grounds $\mathcal{L}_{\mathrm{OM}}$ as a **path-space MAP estimator** via the Freidlin–Wentzell LDP. It is not an evidence lower bound on $\log p(\mathbf{y} \mid \mathbf{x})$: the OM action is a path *prior* regulariser (a small-tube log-probability ratio for a deterministic trajectory under the small-noise reference diffusion) and the data term is a single endpoint likelihood. Strict-ELBO variants on the same backbone, which *do* bound $\log p(\mathbf{y} \mid \mathbf{x})$, are obtained by retaining the FFJORD log-det term (Section 4.3).

Remark 2 (Finite ϵ). Theorem 1 is an $\epsilon \to 0$ statement, but at every ϵ the same S_{OM} is the Ikeda–Watanabe small-tube functional (Ikeda & Watanabe, 1989), so $\mathcal{L}_{\mathrm{OM}}$ is a data fit plus a path prior throughout, with relative weight 1 (Appendix A.4).

A self-contained proof of Theorem 1 via the Freidlin–Wentzell rate functional is in Appendix D.3. Three features distinguish equation 11 from existing methods: it is *trace-free* (no Hutchinson estimator of $\nabla{\cdot}v_\phi$, unlike FFJORD), *auxiliary-loss-free* (no denoising score matching against the bridge marginal, unlike DBVI, because $v_{\mathrm{ref}}^{\mathrm{Bri}}$ is closed form), and *partition-function-free* (no $\partial_t \log Z_t$ estimate, unlike NFS^2's velocity-driven SMC with a Stein control variate).

Monte Carlo estimation. The data term is a single Euler trajectory through equation 8. The regulariser equation 12 is a single MC pair (s_t, U_{s_t}) per minibatch with $s_t \sim \mathcal{U}[0,1]$ and $U_{s_t} \sim \mathcal{N}(\phi(s_t)\mu_\theta(\mathbf{Z}), \kappa(s_t)I)$ drawn in closed form from the forward bridge marginal — no time clipping is applied, and none is needed, since $\bar{v}_{\mathrm{ref}}$ is Lipschitz on all of $[0,1]$ by (A1). Both terms reparameterise, so (θ, ϕ) are updated by joint SGD on the sum of the data term and $\widehat{\mathcal{R}}$.

Why no flow-matching auxiliary loss? Generative flow matching (Lipman et al., 2023) needs an explicit regression target because the data distribution is only available through samples; for posterior transport the unnormalised posterior is available pointwise through the data NLL so the path term can be written against a closed-form reference drift instead of a regression target estimated from samples. Adding an annealed-Langevin flow-matching auxiliary does not improve results (Appendix E).

4.3 Strict path-space ELBO variants on the same bridge backbone

If a strict lower bound on $\log p(\mathbf{y} \mid \mathbf{x})$ is preferred to the MAP framing, the bridge-anchored sampler in equation 8 admits two natural strict-ELBO objectives that retain the FFJORD instantaneous change-of-variables term (Chen et al., 2018):

$$\textbf{(CNF)} \quad \mathcal{L}_{\mathrm{CNF}}(\theta,\phi) = -\,\mathbb{E}_{q_\phi}\big[\log p(\mathbf{y} \mid \mathbf{F}^{(L)})\big] + \mathrm{KL}\big(q_\phi(\widehat{\mathbf{U}}) \,\|\, p(\mathbf{U})\big), \tag{13}$$

$$\text{with} \quad \log q_\phi(U_1) = \log p_0(U_0) - \int_0^1 \nabla{\cdot}v_\phi(U_\tau, 1-\tau)\,\mathrm{d}\tau,$$

$$\textbf{(CNFOM)} \quad \mathcal{L}_{\mathrm{CNFOM}}(\theta,\phi) = \mathcal{L}_{\mathrm{CNF}}(\theta,\phi) + \alpha\, S_{\mathrm{OM}}(v_\phi;\, \bar{v}_{\mathrm{ref}}). \tag{14}$$

$-\mathcal{L}_{\mathrm{CNF}}$ is a strict lower bound on $\log p(\mathbf{y} \mid \mathbf{x})$ (tight up to FFJORD discretisation), and $-\mathcal{L}_{\mathrm{CNFOM}}$ is looser by exactly $\alpha S_{\mathrm{OM}} \geq 0$; we keep CNFOM not to tighten the bound but because it injects the same path-space regularisation as $\mathcal{L}_{\mathrm{OM}}$, isolating the Hutchinson-vs-OM trade on a common backbone. The divergence is a single-sample Hutchinson trace per Euler step, and bridge initialisation, velocity net, step count and MC budget are shared with $\mathcal{L}_{\mathrm{OM}}$, so only the path-space objective differs.

Despite their strict-ELBO status, both variants lose to the trace-free MAP objective on 13 of the 14 UCI regression cells (Appendix J, Table A9; the exception is *qsar* RMSE), with the largest margins on *power* and *protein*. Our gradient-variance measurements (Appendix R) show a substantial Hutchinson penalty on *power*, though the effect is dataset-dependent, so we present this as the leading explanation rather than an established cause. This is why the trade has not been made before—at any $g > 0$ Girsanov KL is finite and a valid bound is free, so ELBO methods (DSVI, IPVI, DDVI, DBVI; PIS, DDS, DGFS, SGDS, NFS^2) never have to give it up. Only the deterministic limit $g \to 0$ forces the choice between absorbing an explicit log-det term and accepting a rigorous non-ELBO MAP objective.

4.4 Theoretical properties

Four results, stated and proved in Appendix A.5, characterise the construction. (i) *Bridge marginals* (Prop. 1, restating Xu et al. (2026, Prop. 2)): under affine drift the bridge marginal is Gaussian with coefficients ϕ, κ solving a closed two-dimensional ODE system, giving $\phi(1) \approx 0.37$, $\kappa(1) \approx 0.50$ at our default $\lambda = g = \sigma_0 = 1$; the h-transform factor $c_s \sim 1/s$ diverges at the anchor, but cancels in $\dot{\phi}$ and $\dot{\kappa}$, leaving the reference drift regular on all of $[0,1]$ (A1). (ii) *Initial-state variance reduction*

(Prop. 2): the bridge start has per-coordinate variance $\kappa(1) \approx 0.50$ against $\mathrm{tr}(K_{\mathbf{ZZ}})/M = 1$ for the unbridged prior, i.e. half the initialisation variance. (iii) ***DBVI correspondence*** (Prop. 3, Remark 4): the setting $v_\phi(U, 1-\tau, \mathrm{ctx}) \equiv \bar{v}_{\mathrm{ref}}(U, \tau)$ zeroes the path term *alone* (it is not the optimum of $\mathcal{L}_{\mathrm{OM}}$, whose data term pulls away from it), and there our sampler matches the marginals of DBVI's bridge SDE — the correspondence is Song's PF-ODE identity, *not* the naive $g \to 0$ limit — and the learned velocity, carrying no factorisation constraint, is less structurally constrained than DBVI's score-parameterised drift and needs no conditional-DSM auxiliary. (iv) ***Few-step Euler error*** (Prop. 4): $\mathbb{E}\|U_1^{(N)} - U_1^{(\infty)}\|^2 \le C_v e^{2L_v} N^{-2}(L_v^2 \kappa(1) + L_s^2)$, so bridge anchoring halves the variance part of the few-step prefactor $\kappa(1)$ — the formal counterpart of Appendix B.4.

Computationally, each gradient step shares its leading $\mathcal{O}(LM^3)$ Cholesky and $\mathcal{O}(LBMd_{\ell-1})$ kernel cost with DSVI and adds $\mathcal{O}(NLMd_\ell d_h)$ for the velocity field; at $M = 128$, $d_h = 128$, $N = 10$, $L = 2$ this is a $\sim 17\%$ overhead over DSVI (Section 5.4, Appendix M).

5 Experiments

Datasets and protocol. Seven small/medium UCI regression datasets (*yacht* 308, *boston* 506, *energy* 768, *qsar* 908, *concrete* 1030, *power* 9568, *protein* 45730), two large regression benchmarks (*airline*, 200k rows; *year*, reported both as a 200k subsample and in full at 5.15×10^5 rows), and two 200k-row binary classification benchmarks with a Bernoulli head (*SUSY*, *HIGGS*). All methods share a 2-layer DGP backbone with $M = 128$ inducing variables and an ARD-RBF kernel, implemented from scratch in one PyTorch file with no GPyTorch dependency, so the sparse-GP layer, kernel, optimiser, batch size, MC budget and evaluation protocol are identical. We use Adam at lr 10^{-2}, batch 256 (small/medium) or 1024 (large), $T = 100$ epochs for small/medium data and proportionally fewer for large, 2 MC samples per gradient step, 32 per evaluation, and $80/20$ splits. Small/medium cells report mean$\pm$std over 10 seeds (SGHMC 5); large regression and classification use 3–5. **Bold** marks the column-best entry and anything within one standard deviation of it.

Methods. We compare **DSVI** (Salimbeni & Deisenroth, 2017), **SGHMC** (Havasi et al., 2018), **IPVI** (Yu et al., 2019), **DDVI** (Xu et al., 2024), **DBVI** (Xu et al., 2026) and our **FBVI-bridge-Path** (alias **OM-Path**), the trace-free MAP estimator of Eq. 11. On the *same* bridge backbone we also report three ablations that differ only in the training loss: the strict-ELBO **FBVI-bridge-CNF** and **FBVI-bridge-CNFOM** (Eqs. 13–14, Appendix J), and a biased **implicit-q surrogate** that drops the path-prior term (Appendix K); **FBVI** denotes the bridge-free variant of Appendix A.1. The implicit-q rows appear in the large-data and classification tables because they are the configuration earlier drafts used, and they show what the principled objective recovers; they are *not* the method of this paper. A NFS^2-style PINN-residual variant (Chen et al., 2025) is in Appendix L.

5.1 Main results on small/medium regression ($L = 2$)

Table 1 reports test RMSE and NLL. The relevant comparison is **OM-Path versus DBVI**: both take their reference process from the same Doob-bridged forward SDE and differ only in whether the objective is a Girsanov ELBO on the reverse SDE (DBVI, with a conditional-DSM auxiliary) or a Freidlin–Wentzell rate functional on the deterministic reverse ODE (ours, closed-form reference drift, no auxiliary). Under matched compute and 10 matched seeds, the paired Wilcoxon signed-rank test (Appendix S) gives **significant OM-Path wins** on the two largest datasets—*power* ($p = 0.007$ on both metrics; OM-Path NLL $\mathbf{0.009}$, matching DSVI's 0.017, against DBVI at 0.125) and *protein* ($p = 0.002^{**}$ on both; RMSE $0.761 \to \mathbf{0.718}$, NLL $1.146 \to \mathbf{1.088}$)—**ties** ($p \in [0.19, 0.35]$) on *yacht* and *qsar*, and **DBVI ahead** ($p \gtrsim 0.97$) on *boston*, *energy* and *concrete*. This is consistent with a MAP-vs-stochastic-Bayes trade-off (deterministic MAP being sample-efficient when the posterior concentrates, losing to stochastic Bayes at low N and high noise), though the FW result predicts no such finite-sample transition. The within-1σ bolding is a deliberately loose flag for parity candidates; the Wilcoxon test is the definitive verdict. On *qsar* the bolded group spans nearly every method with DSVI narrowly leading—when the posterior is close to Gaussian, the flexibility of flow- or score-based inference does not pay for its overhead—and SGHMC is competitive only on the smallest datasets, not reaching *power* within our budget. The DBVI runs diverge on *power* seed 8 and *protein* seeds $\{2, 4, 6\}$; these are excluded from its means but counted as losses by the rank test (Appendix S). Appendix B.1 visualises the table and the per-epoch *protein* trajectories.

Table 1: UCI regression at $L = 2$: test RMSE (top) and test NLL (bottom). DBVI and OM-Path are re-run from scratch on the *same* 10 matched seeds, and these are exactly the runs the paired Wilcoxon test of Appendix S uses—the two tables report one set of numbers. Baselines follow the cited protocols (DSVI / IPVI / DDVI 10 seeds, SGHMC 5; the DSVI *power* cell is re-run on 3 matched seeds inside our compute envelope), and DDVI on *power* / *protein* is re-run with lr 3×10^{-3} and gradient clip 1.0 so the comparison is non-pathological. Means and standard deviations are over the seeds that produced a finite result (a run is excluded only on a hard numerical divergence: non-finite loss, or RMSE $> 5\times$ the mean predictor). No OM-Path run was excluded anywhere; DBVI diverged on *power* seed 8 and *protein* seeds $\{2, 4, 6\}$, so those cells average 9 and 7 seeds respectively (the divergent pairs are still counted, as DBVI losses, by the rank test in Appendix S). **Bold** = best mean per column and every method whose mean lies within one standard deviation of the best entry.

Method	yacht	boston	energy	qsar	concrete	power	protein
DSVI	.496±.067	.440±.049	.298±.025	**.633±.060**	.458±.024	**.245±.005**	.841±.009
SGHMC	.476±.075	.443±.053	.332±.044	**.679±.076**	.491±.043	–	.871±.029
IPVI	.758±.174	.580±.060	.396±.045	.795±.125	.782±.076	.439±.096	.958±.034
DDVI	.909±.292	.454±.091	.300±.078	.936±.326	.641±.068	.467±.250	.762±.011
DBVI	**.339±.047**	**.390±.047**	**.175±.013**	**.643±.049**	**.403±.038**	.273±.041	.761±.014
Path (ours)	**.337±.088**	.453±.145	.201±.034	**.641±.069**	**.430±.030**	**.243±.009**	**.718±.031**

Method	yacht	boston	energy	qsar	concrete	power	protein
DSVI	.904±.044	.799±.050	.629±.016	**1.028±.054**	.832±.023	**.017±.017**	1.253±.009
SGHMC	1.023±.071	**.688±.054**	.865±.077	1.176±.078	.967±.065	–	1.434±.168
IPVI	1.219±.362	.982±.091	.738±.062	1.332±.183	1.227±.098	.725±.132	1.670±.377
DDVI	1.210±.159	**.712±.058**	**.572±.025**	1.271±.078	1.025±.066	.656±.635	1.146±.014
DBVI	**.721±.044**	**.687±.049**	**.495±.009**	**1.007±.060**	**.676±.034**	.125±.157	1.146±.018
Path (ours)	**.699±.066**	.783±.147	**.468±.150**	**1.016±.066**	.717±.032	**.009±.034**	**1.088±.042**

5.2 Large-data regression: 200k–515k rows

With the principled OM objective the deterministic-ODE method is **mean-best on all six cells** of Table 2: on *year_full* it drops RMSE from DBVI's 0.837 to $\mathbf{0.830}$ and NLL from 1.242 to $\mathbf{1.232}$; on the 200k *year* subsample, RMSE $0.864 \to \mathbf{0.827}$ and NLL $1.291 \to \mathbf{1.230}$. We report *mean-best*, not a significance verdict: these cells were not run as matched pairs, and at 3–5 seeds a one-sided exact Wilcoxon bottoms out at $p = 2^{-n}$ (0.031 at $n = 5$, 0.125 at $n = 3$). On *year_full* the 0.007 RMSE gap sits inside 1σ of the seed noise in any case. The NLL margins on *year* are the largest and most robust signal here. The implicit-q surrogate was losing on these datasets (0.889, 0.890) and the principled objective closes that gap, while unbridged FBVI loses stability on *year_full* (one seed diverged, RMSE 1.38 ± 0.49)—the bridge anchoring is what lets the deterministic flow scale to high-dimensional inputs.

Table 2: Large regression benchmarks (airline 200k, year 200k subsample, year full 515k), 3–5 seeds. **Bold** = tied-best per column; no cell is paired-tested (see text).

	RMSE ↓			NLL ↓		
Method	airline	year (200k)	year_full	airline	year (200k)	year_full
DSVI	0.983±0.001	0.886±0.017	0.885±0.006	1.412±0.001	1.313±0.016	1.299±0.005
SGHMC	–	–	0.940±0.042	–	–	1.387±0.061
IPVI	1.127±0.388	0.889±0.005	–	1.473±0.146	1.316±0.007	–
DDVI	1.000±0.010	–	–	1.422±0.008	–	–
DBVI	0.985±0.001	0.864±0.019	0.837±0.003	1.414±0.001	1.291±0.020	1.242±0.003
FBVI	**0.956±0.003**	0.890±0.010	1.383±0.489	1.386±0.003	1.319±0.009	2.027±0.717
FBVI-bridge (impl-q)	**0.956±0.002**	0.890±0.008	0.889±0.004	1.384±0.002	1.319±0.007	1.304±0.004
OM-Path (ours)	0.949±0.004	**0.827±0.003**	**0.830±0.022**	**1.369±0.004**	**1.230±0.004**	**1.232±0.027**

5.3 BINARY CLASSIFICATION: SUSY AND HIGGS

To test a non-Gaussian likelihood we add a Bernoulli head and run all inference families on two 200k-row high-energy-physics benchmarks; the full error / NLL / AUC / F1 table is Table A1 (Appendix B.2; see also Appendices I and G).

OM-Path sits in the tied-best group on both datasets, alongside DBVI on every metric (SUSY Err 0.198 vs. 0.199, AUC 0.873 vs. 0.872; HIGGS Err 0.271 vs. 0.274, AUC 0.807 vs. 0.804). The HIGGS margins are within one standard deviation, so we read them as ties; the salient point is that the implicit-q surrogate was *losing* on HIGGS (4.7 AUC points behind DBVI) and the principled MAP objective closes that gap entirely, leaving no score-vs-velocity penalty on non-Gaussian likelihoods. Two baselines misbehave: DDVI needs a symmetric ± 15 logit clip before the BCE (otherwise the unconditional VP-SDE initialisation saturates the likelihood and NLL exceeds 10^{10}) and even then trails every bridge-anchored method, so data anchoring is the missing ingredient rather than a numerical tweak; IPVI collapses to a degenerate all-positive predictor (Appendix I).

5.4 DEPTH, COST, FEW-STEP INFERENCE AND STABILITY

Four secondary checks, in full in Appendices B.3, M, B.4 and H. **Depth ($L \in \{2, 3, 4, 5\}$).** DSVI is depth-stable but never improves with depth. Unbridged FBVI loses $\sim 50\%$ RMSE on *yacht* as depth grows (one seed explodes), whereas OM-Path stays within $\pm 10\%$ on *energy*, *power*, *qsar* and *concrete* and *improves* on *protein* ($0.745 \to 0.669$ from $L = 2$ to 4); its losses at $L = 5$ are confined to the smallest noisy datasets.

Cost. Training slowdowns relative to DSVI on *protein* are FBVI $1.01\times$, DDVI $1.15\times$, OM-Path $1.17\times$, DBVI $1.28\times$; inference costs 0.2–0.7 ms per posterior sample.

Few-step inference and stability. Because the sampler is a deterministic ODE, the 10 trained Euler steps can be cut to 1–4 at evaluation without retraining, and OM-Path is the most step-count-stable method: within $\sim 30\%$ of its 10-step RMSE at every count with no blow-up, where the implicit-q surrogate fails at 1 step on *power* (2.541 vs. 0.250) and unbridged FBVI at 2 steps on *concrete* (0.955 vs. 0.384). Separately, IPVI is seed-unstable under our backbone (*yacht* NLL 1.22 ± 0.36 against 0.526 ± 0.020), a property of the adversarial family.

5.5 BAYESIAN OPTIMIZATION WITH DGP SURROGATES

Bayesian optimization is a natural downstream test: the surrogate's posterior uncertainty drives the acquisition, so a miscalibrated posterior translates directly into wasted queries. We compare **DSVI**, **OM-Path** and a **Random** lower bound on four synthetic functions under one-sample Thompson sampling (one posterior sample, argmin over a fresh pool of 1000 candidates, evaluate, append, refit), with 50 random initial points, 100 iterations, a 2-layer DGP surrogate at $M = 64$ with 80-epoch refits, 5 seeds. OM-Path is mean-best on **Hartmann-6** (0.389 ± 0.144 vs. DSVI 0.482 ± 0.218, Random 1.106 ± 0.285) and **Levy-20** (85.7 ± 9.5 vs. 93.0 ± 16.7, 97.9 ± 18.7), with the lowest seed variance in both. It ties DSVI on **Ackley-50** and leads on **Rosenbrock-100** without statistical separation; neither high-dimensional cell is rescued by any surrogate. The seed-replacement rule behind Ackley-50's $n = 3$, the full table and the regret curves are in Appendix B.5.

6 CONCLUSION AND LIMITATIONS

FBVI-bridge-Path makes DGP inducing-variable inference deterministic transport: Song's probability flow ODE on DBVI's Doob-bridged SDE gives a closed-form reference drift, and the Onsager–Machlup action a trace-free objective that Theorem 1 ties to a small-noise conditional MAP path. It wins the two largest UCI datasets, gives no bound on $\log p(\mathbf{y} \mid \mathbf{x})$, and concedes the small-N cells.

Reproducibility statement

Every number in this paper is produced by a single file `fbvi_native.py` ($\sim$ 2,200 LOC, no GPyTorch dependency) and an aggregation script `aggregate_table.py`, both released anonymously as supplementary material. The model, objective and training procedure are stated in full in Sections 4 and 5, with the complete experimental protocol (splits, seeds, optimiser settings and hyperparameters) in Appendix C and the surrounding appendices. All assumptions of the theoretical statements are given with their propositions and theorems, and the corresponding proofs appear in the appendix.

References

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Mireille Capitaine. Onsager–machlup functional for some smooth norms on Wiener space. *Probability Theory and Related Fields*, 102(2):189–201, 1995.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Wuhao Chen, Zijing Ou, and Yingzhen Li. Neural flow samplers with shortcut models. In *Frontiers in Probabilistic Inference: Learning meets Sampling*, 2025.

Andreas Damianou and Neil D Lawrence. Deep gaussian processes. In *Artificial intelligence and statistics*, pp. 207–215. PMLR, 2013.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer, 2 edition, 2010.

Detlef Dürr and Alexander Bach. The Onsager–Machlup function as lagrangian for the most probable path of a diffusion process. *Communications in Mathematical Physics*, 60(2):153–170, 1978.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, volume 2025, pp. 34668–34684, 2025.

Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer, New York, 2 edition, 1998.

Marton Havasi, José Miguel Hernández-Lobato, and Juan José Murillo-Fuentes. Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo. *Advances in neural information processing systems*, 31, 2018.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic Differential Equations and Diffusion Processes*, volume 24 of *North-Holland Mathematical Library*. North-Holland/Kodansha, 2 edition, 1989.

Hilbert J Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011–P11011, 2005.

Minkyu Kim, Kiyoung Seong, Dongyeop Woo, Sungsoo Ahn, and Minsu Kim. On scalable and efficient training of diffusion samplers. In *Advances in Neural Information Processing Systems*, volume 38, pp. 26042–26068, 2025.

Remo Kretschmann. Are minimizers of the Onsager–Machlup functional strong posterior modes? *SIAM/ASA Journal on Uncertainty Quantification*, 11(4):1105–1138, 2023.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

Sanjeev Raja, Martin Šípka, Michael Psenka, Tobias Kreiman, Michal Pavelka, and Aditi S Krishnapriyan. Action-minimization meets generative modeling: Efficient transition path sampling with the Onsager–Machlup functional. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50972–51008. PMLR, 2025.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. *Advances in neural information processing systems*, 30, 2017.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023.

Francisco Vargas, Will Sussman Grathwohl, and Arnaud Doucet. Denoising diffusion samplers. In *The Eleventh International Conference on Learning Representations*, 2023.

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

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

Jongmin Yoon and Juho Lee. Sequential flow straightening for generative modeling. *arXiv preprint arXiv:2402.06461*, 2024.

Haibin Yu, Yizhou Chen, Bryan Kian Hsiang Low, Patrick Jaillet, and Zhongxiang Dai. Implicit posterior variational inference for deep gaussian processes. *Advances in neural information processing systems*, 32, 2019.

Dinghuai Zhang, Ricky TQ Chen, Chenghao Liu, Aaron Courville, and Yoshua Bengio. Diffusion generative flow samplers: Improving learning signals through partial trajectory optimization. In *International Conference on Learning Representations*, volume 2024, pp. 47211–47232, 2024.

Qinsheng Zhang and Yongxin Chen. Path integral sampler: A stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022.

A Deferred method material

A.1 Vanilla FBVI: velocity-field VI without the bridge

Before introducing the bridge, we first describe the simplest velocity-field variational family on the same DGP backbone. **Vanilla FBVI** defines the variational posterior over each layer's inducing variables as the push-forward of the GP prior $p(\mathbf{U}^{(\ell)}) = \mathcal{N}\Big(0, K_{\mathbf{ZZ}}^{(\ell)}\Big)$ through a learned ODE:

$$U_0^{(\ell)} \sim \mathcal{N}\Big(0, K_{\mathbf{ZZ}}^{(\ell)}\Big), \quad \frac{\mathrm{d}U_t^{(\ell)}}{\mathrm{d}t} = v_\phi^{(\ell)}(U_t^{(\ell)}, t), \quad \widehat{\mathbf{U}}^{(\ell)} := U_1^{(\ell)}. \tag{15}$$

The ELBO is the standard sparse-GP ELBO of Eq. 2; the expectation under $q_\phi(\mathbf{U})$ is estimated by a single reparameterised Euler trajectory through equation 15. The velocity net is zero-initialised on its last layer so that $\widehat{\mathbf{U}} \approx U_0$ and q_ϕ starts close to the prior, reproducing the standard DSVI initialisation.

Vanilla FBVI is the deterministic-ODE analogue of DDVI (Xu et al., 2024) (see Appendix C for a full review), with two differences: (i) no Brownian noise during integration, and (ii) no denoising score-matching auxiliary loss. The variational density $q_\phi(\mathbf{U})$ is implicit; in our experiments we estimate the ELBO either explicitly via FFJORD's instantaneous change of variables (Chen et al., 2018) (Appendix J) or by dropping $\log q_\phi$ as in Xu et al. (2024)'s implicit-q surrogate (Appendix K).

Vanilla FBVI suffers two limitations that motivate the bridge. First, as depth L or input dimension d grows, the gap between the unconditional prior $\mathcal{N}(0, K_{\mathbf{ZZ}})$ and the true posterior widens; the velocity field must traverse this whole gap in N Euler steps and the optimisation becomes harder. Second, the variational density is not anchored to any data-dependent reference, so the optimisation lacks the variance-reduction benefit of anchoring against a posterior-aware reference process (cf. DBVI's Doob bridge anchoring on the score side). We address both in the next subsection.

A.2 Bridge marginal coefficients

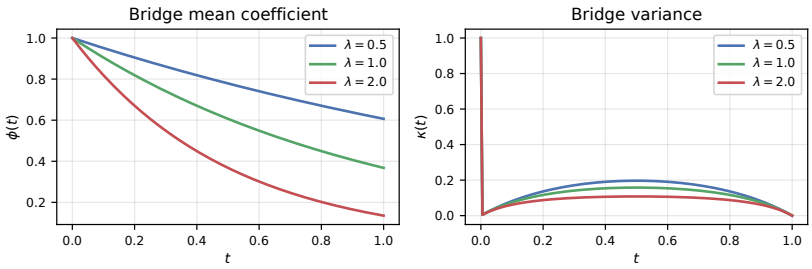

Figure A1: Bridge marginal coefficients $\phi(s)$ (mean attenuation) and $\kappa(s)$ (variance) from the closed-form ODE system in Prop. 1, for three diffusion strengths λ. With our default $\lambda = g = \sigma_0 = 1$ we obtain $\phi(1) \approx 0.37$ and $\kappa(1) \approx 0.50$, both of which appear in the reference drift equation 6 and in the bridge initial distribution sampled in the sampler's initial distribution.

A.3 Technical conditions for Theorem 1

These are the regularity conditions under which the small-noise large-deviation argument behind Theorem 1 applies.

Definition 1 (Technical conditions)**.** *We assume throughout:*

- ***(A1) Regularity of the reference drift on all of*** $[0, 1]$**.** *The Doob h-transform factor c_s of Prop. 1 does diverge as $c_s \sim 1/s$ at the anchor, but the reference drift does* not*: the divergences cancel identically in the coefficient ODEs. Using $\phi(0) = a_0 = 1$ and $\kappa(0) = \sigma_0^2$ in equations 16–17,*

$$\dot{\phi}(s) \to -\lambda, \qquad \dot{\kappa}(s) \to g^2 - 2\lambda\sigma_0^2 \qquad (s \downarrow 0),$$

because the two c_s terms in each equation differ by the factor $(a_s - \phi(s))$ resp. $(a_s\sigma_0^2 - \kappa(s))$, both of which vanish at $s = 0$. Indeed $\phi(s) = e^{-\lambda s}$ solves equation 16 exactly, killing the h-transform term along the whole trajectory. Consequently $\bar{v}_{\mathrm{ref}}(U, \tau)$ of Eq. 7 is affine in U with coefficients that are continuous and bounded on the closed interval, hence globally Lipschitz in U with linear growth, uniformly in $\tau \in [0, 1]$, and the Freidlin–Wentzell regularity hypothesis (Freidlin & Wentzell, 1998, Ch. 3, Cond. (3.1.1)) holds on the whole of $[0, 1]$ with no endpoint cutoff. At our defaults $\lambda = g = \sigma_0 = 1$ the Lipschitz constant is $\sup_s |\dot{\kappa}(s)/2\kappa(s)| = 0.5$, attained at $s = 0$ (Appendix A.2).

- ***(A2) Lipschitz DGP forward and Varadhan tail.*** *The DGP composition $\mathrm{dgp}(\cdot)$ mapping U_1 to $\mathbf{F}^{(L)}$ is L_f-Lipschitz. With Gaussian observation noise this gives $-\log p(\mathbf{y} \mid$*

$U_1) = \frac{1}{2\sigma^2}\|\mathbf{y} - \mathrm{dgp}(U_1)\|^2 + \mathrm{const}$, *which grows at most quadratically in* $\|U_1\|$. *Varadhan's lemma requires the tail condition* $\lim_{R\to\infty} \limsup_{\epsilon\to 0} \epsilon^2 \log \mathbb{E}_{\mathbb{P}^\epsilon_{\mathrm{ref}}}\big[\mathbf{1}\{\|U_1\| > R\}\, p(\mathbf{y}|U_1)^{1/\epsilon^2}\big] = -\infty$. *Under (A1), the reference-marginal large-deviation rate at* U_1 *grows as* $\|U_1\|^2/(2\kappa_{\mathrm{ref}})$ *for some* $\kappa_{\mathrm{ref}} < \infty$, *strictly faster than the quadratic upper bound* $L_f^2\|U_1\|^2/(2\sigma^2)$ *on* $-\log p(\mathbf{y}|U_1)$ *for* σ^2 *above a threshold determined by* L_f *and* κ_{ref}, *so the tail bound holds. We assume this strict-dominance regime throughout (cf. Dembo & Zeitouni, 2010, Lemma 4.3.4).*

(A3) LDP bounds and the Laplace step. *On* $\mathcal{C}([0,1];\mathbb{R}^M)$ *with the uniform topology, the standard Freidlin–Wentzell LDP under (A1) provides the open-set lower bound and the closed-set upper bound*

$$\liminf_{\epsilon\to 0} \epsilon^2 \log \mathbb{P}^\epsilon_{\mathrm{ref}}(U \in O) \;\geq\; -\inf_{u\in O} I_{\mathrm{ref}}(u), \qquad \limsup_{\epsilon\to 0} \epsilon^2 \log \mathbb{P}^\epsilon_{\mathrm{ref}}(U \in C) \;\leq\; -\inf_{u\in C} I_{\mathrm{ref}}(u),$$

for open O *and closed* C, *with* I_{ref} *a good (in particular lower semicontinuous, and* not *claimed to be uniformly continuous on* H^1*) rate functional. We use these two bounds only, and we deliberately do not assert a pointwise identity: because* $I_{\mathrm{ref}}(u) = \frac{1}{2}\int \|\dot u - \bar v_{\mathrm{ref}}(u)\|^2$ *depends on* $\dot u$, *uniform convergence* $u_n \to u$ *does not control* $\dot u_n$, *so* I_{ref} *is in general only lower semicontinuous in the sup norm. The shorthand* $-\epsilon^2 \log \mathbb{P}^\epsilon\{U \approx u\} = I_{\mathrm{ref}}(u) + o(1)$ *used in the main text is a* heuristic abbreviation *of these two bounds, not a theorem. What the argument actually needs is the Laplace principle: minimisers of* I_y *over closed sublevel sets are identified by combining the two bounds with exponential tightness, which is standard once* I_y *is a good rate functional and the tilt is continuous and bounded below. For the likelihood tilt (Eq. 20), Varadhan's lemma requires* $u \mapsto -\log p(\mathbf{y} \mid u_1)$ *to be (i) continuous in the uniform topology and (ii) bounded below; for the Gaussian-noise regression likelihood both are immediate from the Lipschitz DGP forward of (A2), since the map factors as* $u \mapsto u_1 \mapsto \mathrm{dgp}(u_1) \mapsto -\log p$ *with each step Lipschitz, and* $-\log p$ *is bounded below up to a constant. The MAP path* $u^\star$ *is unique whenever* I_y *is strictly convex — e.g. under additional affine/convexity conditions on the forward map — and otherwise the argmin may contain several paths. A Gaussian observation likelihood is* not *sufficient for this: it only gives* $-\log p(\mathbf{y}|u_1) \propto \|\mathbf{y} - \mathrm{dgp}(u_1)\|^2$, *and* dgp *is a nonlinear composition, so the objective is in general non-convex. When the argmin is a set, Theorem 1 should be read as identifying* $\mathcal{L}_{\mathrm{OM}}$ *with a particular element of it, selected by the velocity parameterisation.*

A.4 Finite ϵ and the path prior

At any $\epsilon > 0$, $S_{\mathrm{OM}}(u; \bar v_{\mathrm{ref}})$ is the classical Onsager–Machlup functional of the H^1-path u under the Gaussian reference path measure (Dürr & Bach, 1978; Ikeda & Watanabe, 1989; Capitaine, 1995): the ratio of small-tube probabilities at u and a reference path u_* satisfies $\log\big(\mathbb{P}^\epsilon\{\|U - u\|_\infty < \eta\}/\mathbb{P}^\epsilon\{\|U - u_*\|_\infty < \eta\}\big) = -\epsilon^{-2}[S_{\mathrm{OM}}(u) - S_{\mathrm{OM}}(u_*)] + o(\eta)$ (Ikeda & Watanabe, 1989, Ch. VI.9), i.e. S_{OM} is the ϵ^{-2}-scaled small-tube log-probability ratio at u up to a path-independent constant (there is no ordinary density on this infinite-dimensional path space). So $\mathcal{L}_{\mathrm{OM}}$ is data-fit plus a small-tube path prior at every ϵ, and the tempering exponent $1/\epsilon^2$ scales *both* terms, fixing their relative weight at 1. The scalar α of Eq. 11 is therefore a regularisation weight and nothing more; $\alpha = 1$ is what the theorem prescribes and what we use, and Appendix Q's sweep is a robustness check around it.

In practice the path term is evaluated under the bridge marginal p_s^{Bri} of equation 12, which is available in closed form and coincides with the trajectory law at the configuration that zeroes the regulariser (Prop. 3).

Remark 3 (No endpoint cutoff is needed). The Doob h-transform factor c_s of Prop. 1 diverges as $1/s$ at the anchor, which might suggest that the bridge has to be truncated there. It does not: what the LDP needs is regularity of the *drift*, not of c_s, and the divergences cancel identically in equations 16–17: the two c_s terms differ by $(a_s - \phi(s))$ and $(a_s\sigma_0^2 - \kappa(s))$ respectively, both zero at $s = 0$. In fact $\phi(s) = e^{-\lambda s}$ solves equation 16 exactly, so the h-transform term vanishes along the whole trajectory, and $\dot\phi(0) = -\lambda$, $\dot\kappa(0) = g^2 - 2\lambda\sigma_0^2$ are finite. Numerically integrating equations 16–17 on a 10^5-point grid at $\lambda = g = \sigma_0 = 1$ gives $\sup_{s\in[0,1]} |\dot\kappa(s)/2\kappa(s)| = 0.500$ and $\sup_s |\dot\phi(s)| = 1.000$, both attained

at $s=0$, with no growth as $s\downarrow 0$. Theorem 1 is therefore stated on the full interval $[0,1]$, exactly the interval the sampler integrates, and there is no δ-gap to account for.

A.5 THEORETICAL PROPERTIES

These are the four results summarised in Section 4.4; proofs and derivations are in Appendix D.

Proposition 1 (Bridge marginal mean and variance ODEs)**.** *Consider the Doob h-transformed bridge in forward-bridge time $s \in [0,1]$, under affine forward drift $f(U_s,s) = -\lambda U_s$, constant diffusion g, and initial distribution $p_0^\theta(U_0\,|\,\mathbf{x}) = \mathcal{N}\big(\mu_\theta(\mathbf{x}), \sigma_0^2 I\big)$ at $s=0$. The marginal at each s is Gaussian, $p_s^{\mathrm{Bri}}(U_s\,|\,\mathbf{x}) = \mathcal{N}(m_s, \kappa_s I)$ with $m_s = \phi(s)\,\mu_\theta(\mathbf{x})$, and $\phi(s), \kappa(s)$ are determined by*

$$\dot\phi(s) = -(\lambda + c_s)\,\phi(s) + c_s\,a_s, \qquad \phi(0) = 1, \tag{16}$$

$$\dot\kappa(s) = -2(\lambda + c_s)\,\kappa(s) + g^2 + 2c_s a_s \sigma_0^2, \qquad \kappa(0) = \sigma_0^2, \tag{17}$$

where $a_s = e^{-\lambda s}$, $q_s = g^2(1-e^{-2\lambda s})/(2\lambda)$, and the h-transform correction $c_s = g^2\sigma_0^2 a_s^2/[(a_s^2\sigma_0^2 + q_s)\,q_s]$. The correction c_s diverges *as $s\downarrow 0^+$ because $q_s \sim g^2 s$ near the anchor, giving $c_s \sim 1/s$; this is the standard Doob h-transform conditioning singularity at the bridge anchor endpoint. The ODE system 16–17 is regular on the open interval $s \in (0,1]$, and the boundary values $\phi(0)=1$, $\kappa(0)=\sigma_0^2$ are imposed as one-sided limits enforced by the anchored initial distribution $p_0^\theta(U_0|\mathbf{x}) = \mathcal{N}\big(\mu_\theta(\mathbf{x}), \sigma_0^2 I\big)$ (not as algebraic initial conditions of a Lipschitz ODE).*

This is Xu et al. (2026, Prop. 2) restated in our notation; the proof appears in Appendix D.4 via the forward Kolmogorov equation of the h-transformed SDE. The coefficients satisfy $\phi(0)=1$, $\kappa(0)=\sigma_0^2$ at the anchored start; at the noise-side terminal $s=1$ they take finite values that we evaluate numerically (we obtain $\phi(1)\approx 0.37$, $\kappa(1)\approx 0.50$ for the default $\lambda = g = \sigma_0 = 1$). FBVI-bridge initialises its variational ODE from the noise-side bridge marginal $\mathcal{N}(\phi(1)\mu_\theta(\mathbf{x}), \kappa(1)I)$, which is a proper Gaussian — *not* a degenerate point mass — so the implicit posterior carries non-trivial diversity through the velocity field.

Proposition 2 (Initial-state variance reduction)**.** *Let $\mathrm{tr}_{\mathrm{FBVI}} = \mathrm{tr}(K_{\mathbf{ZZ}})$ and $\mathrm{tr}_{\mathrm{bridge}} = M\,\kappa(1)$ denote the trace of the covariance matrix of the FBVI and FBVI-bridge initial inducing variables, respectively (both at $t=0$ of the variational ODE in our convention, which corresponds to $t=T=1$ of the forward bridge SDE). For the default hyperparameters $\lambda = g = \sigma_0 = 1$ used throughout this paper, numerical integration of Eq. 17 gives $\kappa(1) \approx 0.50$ (and the mean attenuation $\phi(1) \approx 0.37$). Meanwhile $\mathrm{tr}(K_{\mathbf{ZZ}})/M$ equals the kernel amplitude σ_k^2, initialised at 1 for our ARD-RBF, so $\mathrm{tr}_{\mathrm{FBVI}}/M = 1$ at initialisation.*

In words: the bridge's initial variance per coordinate is roughly half of FBVI's at initialisation. Smaller initial variance translates into lower-variance gradient estimates through the integrator, and helps explain why FBVI-bridge converges faster than FBVI in our depth-scaling experiments (Section 5).

Proposition 3 (FBVI-bridge as the probability flow ODE of DBVI's bridge SDE)**.** *Let U_t^{DBVI} solve the bridged DBVI SDE $\mathrm{d}U_t = [-\lambda U_t + g(t)^2 h(U_t,t,U_0)]\,\mathrm{d}t + g(t)\,\mathrm{d}W_t$, with conditional score $h(U_t,t,U_0) = \nabla_{U_t}\log p_t^{\mathrm{Bri}}(U_t|U_0)$ given by Prop. 1. By Lemma 1 (Song's probability-flow ODE for the bridge SDE), the bridge SDE's time-t marginal coincides with the marginal of the deterministic ODE*

$$\mathrm{d}U_t \;=\; \Big[-\lambda U_t + g(t)^2\,h(U_t,t,U_0) - \tfrac{1}{2}g(t)^2\,\nabla_{U_t}\log p_t^{\mathrm{Bri}}(U_t)\Big]\,\mathrm{d}t,$$

which by Lemma 2 has the closed form $\dot U_s = v_{\mathrm{ref}}^{\mathrm{Bri}}(U_s,s)$ of equation 6 when p_s^{Bri} is Gaussian (forward bridge time s). The FBVI-bridge sampler runs in reverse time $\tau = 1-s$, $\dot U_\tau = v_\phi(U_\tau, 1-\tau, \mathrm{ctx})$, where the chain-rule substitution $\mathrm{d}/\mathrm{d}\tau = -\mathrm{d}/\mathrm{d}s$ converts a forward-time drift to its negation in reverse time. Two consequences follow: (i) at the minimiser of the path term alone*,*

$$v_\phi(U, 1-\tau, \mathrm{ctx}) \;\equiv\; \bar v_{\mathrm{ref}}(U,\tau) \;=\; -\,v_{\mathrm{ref}}^{\mathrm{Bri}}(U, 1-\tau),$$

FBVI-bridge has the same marginals at every $s = 1-\tau$ as DBVI's bridge SDE. This is a statement about the regulariser, not about $\mathcal{L}_{\mathrm{OM}}$*: $\bar v_{\mathrm{ref}}$ makes the path term vanish, but the data term has non-zero first variation there in general, so the trained v_ϕ is not $\bar v_{\mathrm{ref}}$ and is not meant to be. This is the single sign convention used throughout: every path-space expression is written against $\bar v_{\mathrm{ref}}$*

(Eq. 7), and the implementation forms it as $-v_{\mathrm{ref}}$. *(ii) Because* v_ϕ *is not constrained to factor as* $-\lambda U + g^2 h - \frac{1}{2} g^2 \nabla \log p$, *the learned velocity is* less structurally constrained *than DBVI's score-parameterised drift, and training needs no conditional DSM auxiliary. (We claim only this drift-level statement: comparing the two* sampler *families as sets is not straightforward, since DBVI additionally carries diffusion noise.)*

Remark 4 (DBVI–FBVI-bridge correspondence is via Song's PF-ODE). The DBVI–FBVI-bridge correspondence is the probability-flow ODE identity of Song et al. (each step retains the score correction $g^2 h - \frac{1}{2} g^2 \nabla \log p$). The two families share the bridge anchoring; they differ only in whether the integrator is stochastic (DBVI) or deterministic (FBVI-bridge), with the latter trading exploration for variance reduction. The naive $g(t) \to 0$ limit of the bridge SDE *is not* this correspondence (it drops the score correction and reduces to the unconditional OU flow); see Appendix D.6 for the short derivation distinguishing the two.

Proposition 4 (Few-step Euler error under bridge anchoring)**.** *Let* $U_1^{(N)}$ *denote the FBVI-bridge sample obtained by* N*-step Euler integration of the ODE* $\mathrm{d}U_\tau/\mathrm{d}\tau = v_\phi(U_\tau, s_\tau, \mathrm{ctx})$ *from* $U_0 \sim \mathcal{N}(\phi(1)\mu_\theta(\mathbf{x}), \kappa(1) I)$, *and let* $U_1^{(\infty)}$ *denote the exact solution. If* v_ϕ *is* L_v*-Lipschitz in* U *and* L_s*-Lipschitz in* s, *then*

$$\mathbb{E}\|U_1^{(N)} - U_1^{(\infty)}\|^2 \;\le\; \frac{C_v\, e^{2L_v}}{N^2} \left(L_v^2\, \kappa(1) + L_s^2\right),$$

for the centred *process, where* C_v *is an absolute constant depending only on the integrator and* $\kappa(1)$ *is the per-coordinate initial variance. In particular the error scales as* $\mathcal{O}(1/N^2)$ *in the step count, and the variance part of the prefactor is controlled by* $\kappa(1)$, *which is* $\sim 50\%$ *of the FBVI prior variance* $\mathrm{tr}(K_{\mathbf{ZZ}})/M$ *at our default hyperparameters. The uncentred moment additionally carries the amortiser mean term* $\|\phi(1)\mu_\theta(\mathbf{x})\|^2$, *which has no unbridged counterpart and is excluded from the comparison (Appendix D.7).*

Proposition 4 is the formal counterpart of the few-step phenomenon we observe in Appendix B.4: FBVI-bridge admits near-lossless 1-step inference on datasets where the velocity field v_ϕ has small Lipschitz constant, and the prefactor $\kappa(1)$ halves the *variance* contribution to the few-step error bound relative to unbridged FBVI. The proof (Appendix D.7) is a standard ODE-stability argument combined with the variance bound of Prop. 2.

A.6 Computational complexity

Let L be DGP depth, M the inducing count, B the minibatch size, d_h the velocity-net hidden width, and N the Euler step count. Each ELBO gradient step incurs three asymptotic contributions:

- Cholesky of $K_{\mathbf{ZZ}}^{(\ell)}$: $\mathcal{O}(LM^3)$ (shared with DSVI);
- batched kernel evaluations $K_{\mathbf{Zx}}^{(\ell)}$: $\mathcal{O}(LBMd_{\ell-1})$ (shared with DSVI);
- velocity-field forward + backward passes through the integrator: $\mathcal{O}(NLMd_\ell d_h)$ (specific to FBVI / FBVI-bridge / DBVI / DDVI).

The leading term is shared with DSVI; FBVI / FBVI-bridge add the third contribution, which dominates the constant when $Nd_h \gtrsim M^2/(Bd_{\ell-1})$. For $M = 128$, $d_h = 128$, $N = 10$, $L = 2$ the velocity overhead is $\sim 17\%$ of DSVI on *protein*, in line with the wall-clock numbers in Section 5.4. DBVI adds a comparable score-net pass plus the DSM auxiliary, and DDVI adds the same DSM pass without the bridge amortiser; the resulting constant-factor differences are reported in Appendix M.

B Deferred empirical material

B.1 Main-table figures

B.2 Binary classification: full metric table

B.3 Depth scaling

Table A2 reports test RMSE on the seven small/medium datasets for $L \in \{2, 3, 4, 5\}$, averaged over 10 seeds for DSVI / FBVI / DBVI and 3 seeds for FBVI-bridge-Path. Two observations:

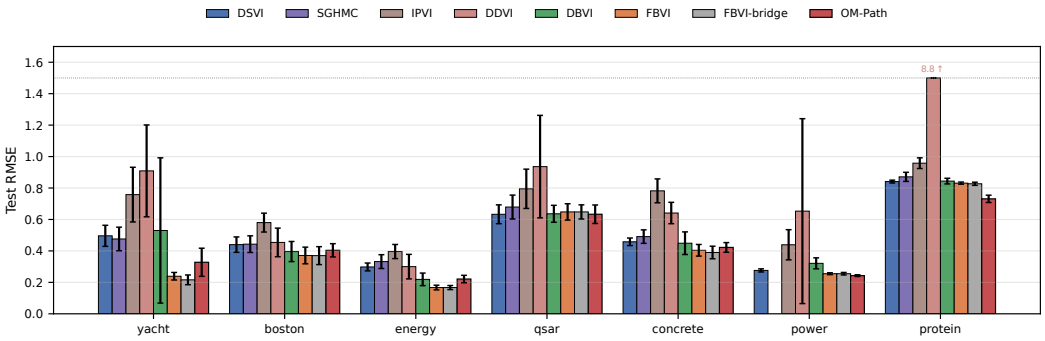

Figure A2: Per-dataset test RMSE at $L = 2$ (mean$\pm$std, 10 seeds). Bars exceeding 1.5 are clipped and the actual value annotated above. FBVI-bridge-Path and DBVI are jointly the strongest methods on the small/medium UCI datasets, with the per-dataset winner alternating between them; see Appendix S for the matched-seed Wilcoxon verdict.

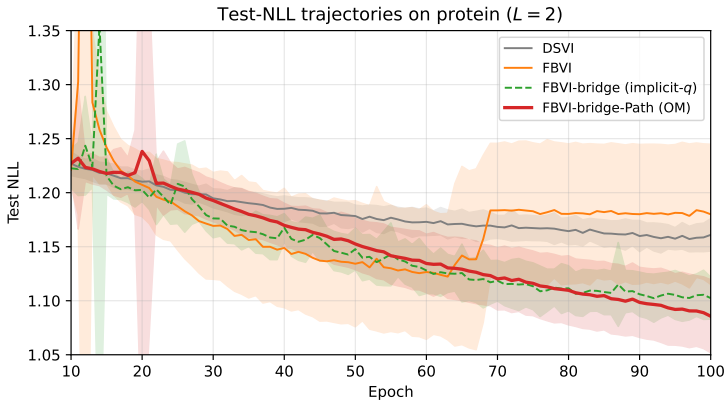

Figure A3: Test-NLL trajectories on *protein* ($L = 2$). Solid lines are seed-mean; shaded bands are ± 1 std. FBVI-bridge-Path (OM, our main method, 10 seeds) and FBVI-bridge (implicit-q, 3 seeds re-run under matched protocol) converge to NLL ≈ 1.10 by epoch 100, with the OM variant nominally lowest (1.088 ± 0.042); DSVI plateaus higher at ≈ 1.16; unbridged FBVI has a single-seed instability event near epoch 68. SGHMC / IPVI / DDVI omitted (separate train protocol). Matches the protein NLL of Table 1: FBVI-bridge-Path $\mathbf{1.088 \pm 0.042}$ vs. DBVI 1.146.

1. **DSVI is depth-stable but does not improve with depth.** Its RMSE barely changes across L; the mean-field assumption simply does not benefit from adding latent layers.
2. **FBVI-bridge-Path is depth-stable on every dataset, and improves with depth on protein.** Unbridged FBVI loses $\sim 50\%$ RMSE on yacht ($0.24 \to 0.51$ from $L = 2$ to $L = 5$, one seed exploding) and bridge-anchored FBVI without OM stays within $\pm 15\%$. With the principled FBVI-bridge-Path objective, FBVI-bridge-Path stays within $\pm 10\%$ on *energy*, *power*, *qsar*, *concrete*, and *improves* on *protein* ($0.745 \to 0.669$ from $L = 2$ to $L = 4$, the largest dataset where depth matters most). At $L = 5$ FBVI-bridge-Path sits within one standard deviation of the best on the larger datasets (*energy*, *qsar*, *concrete*, *power*, *protein*); the losses are concentrated on the smallest noisy datasets (*yacht*, *boston*), consistent with the MAP-vs-Bayes trade-off discussed in Section 5.1.

B.4 Few-step inference

Because the flow variants integrate a deterministic ODE, they admit *few-step inference* — at evaluation time we can reduce the integration from the trained 10 Euler steps down to 1, 2, or 4 without retraining. Table A3 reports test RMSE across step counts on six UCI regression datasets, for the

Table A1: Binary classification on SUSY and HIGGS (200k each, 3 seeds). **Bold** = tied-best per column (within 1 std). Error and NLL: lower is better. AUC and F1: higher is better. DDVI uses a symmetric logit-clip of ± 15 before BCE to bound the Bernoulli loss when the unconditional VP-SDE initialisation produces extreme inducing-variable samples; without this guard DDVI's NLL overflows beyond 10^{10}.

	SUSY				HIGGS			
Method	Err ↓	NLL ↓	AUC ↑	F1 ↑	Err ↓	NLL ↓	AUC ↑	F1 ↑
DSVI	**.200±.001**	**.433±.002**	**.872±.001**	**.767±.002**	.279±.000	.543±.001	.798±.001	.737±.002
SGHMC	.203±.001	.441±.002	.869±.001	.754±.003	.283±.001	.555±.007	.793±.002	.738±.001
IPVI	.504±.015	.656±.005	.804±.028	.640±.005	.462±.005	.676±.004	.733±.008	.695±.003
DDVI	.220±.002	.471±.001	.855±.001	.716±.002	.361±.001	.639±.001	.684±.002	.663±.002
DBVI	**.199±.001**	**.435±.003**	**.872±.001**	**.766±.005**	.274±.002	.538±.003	.804±.002	**.747±.005**
FBVI	.213±.011	.457±.020	.862±.009	.734±.026	.361±.002	.640±.001	.685±.002	.661±.003
FBVI-br (impl-q)	**.199±.001**	**.434±.004**	**.872±.002**	**.766±.005**	.306±.039	.577±.044	.760±.054	.716±.035
FBVI-br-Path (OM)	**.198±.001**	**.432±.002**	**.873±.001**	**.768±.000**	**.271±.004**	**.533±.004**	**.807±.004**	**.747±.004**

Table A2: Depth-trend test RMSE (mean; 10 seeds for DSVI / FBVI / DBVI, 3 seeds for FBVI-bridge-Path). All seven small/medium UCI datasets at $L \in \{2, 3, 4, 5\}$. **Bold** = best method per (dataset,L) cell.

Method	Dataset	$L=2$	$L=3$	$L=4$	$L=5$
DSVI	yacht	0.467	0.481	0.484	0.507
	boston	0.436	0.479	0.470	0.481
	energy	0.282	0.285	0.282	0.276
	qsar	**0.632**	**0.668**	**0.666**	0.656
	concrete	0.462	0.480	0.463	0.471
	power	**0.270**	**0.299**	0.299	0.373
	protein	0.837	0.839	0.835	0.833
FBVI	yacht	0.236	0.287	0.344	0.320
	boston	**0.363**	0.432	0.450	0.433
	energy	**0.163**	0.199	0.201	0.205
	qsar	0.654	0.690	0.688	0.708
	concrete	0.410	**0.406**	**0.407**	**0.420**
	power	0.253	0.270	**0.259**	**0.298**
	protein	0.830	0.829	0.850	0.830
DBVI	yacht	0.341	0.266	0.345	0.395
	boston	0.408	0.439	0.472	0.475
	energy	0.209	0.213	0.244	0.272
	qsar	0.641	0.692	0.676	0.697
	concrete	0.419	0.448	0.528	0.529
	power	0.325	0.502	0.406	0.469
	protein	0.834	0.840	0.842	0.833
FBVI-bridge-Path	yacht	**0.245**	0.368	0.361	0.350
	boston	0.426	**0.431**	**0.480**	0.462
	energy	**0.227**	**0.232**	**0.226**	**0.227**
	qsar	0.680	**0.654**	0.690	**0.669**
	concrete	**0.433**	0.475	0.459	0.469
	power	**0.247**	**0.246**	**0.248**	**0.249**
	protein	**0.745**	**0.704**	**0.669**	**0.683**

four non-Gaussian VI methods. Figure A5 visualises the same data as inference-steps-vs-RMSE curves.

Three findings emerge:

1. **Bridge anchoring + low-variance training enables aggressive few-step inference on small datasets.** The *implicit-q surrogate* (FBVI-br (impl-q)) is the most aggressive 1-step method on *energy* (0.202, within 9% of its 10-step 0.185) and *concrete* (0.379 vs 0.380 at 10 steps); FBVI without the bridge is 2–3× worse at 1 step on these two datasets. We

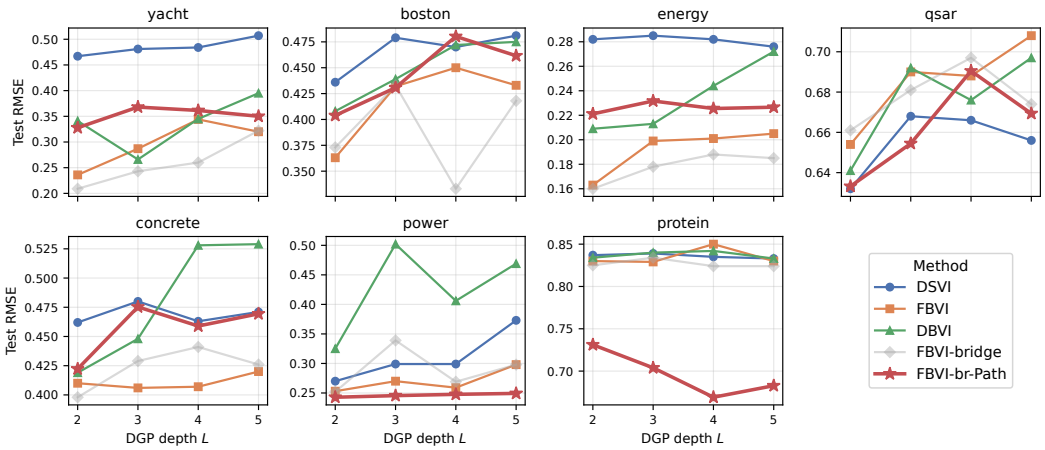

Figure A4: Depth-scaling test RMSE on the seven small/medium UCI datasets (mean over 10 seeds for DSVI / FBVI / DBVI and 3 seeds for FBVI-bridge-Path; lower is better). FBVI-bridge-Path (red) remains close to its $L = 2$ value across $L \in \{2, 3, 4, 5\}$ while unbridged FBVI (orange) and DBVI (green) drift upward on *yacht*, *power*, and *concrete* as depth grows.

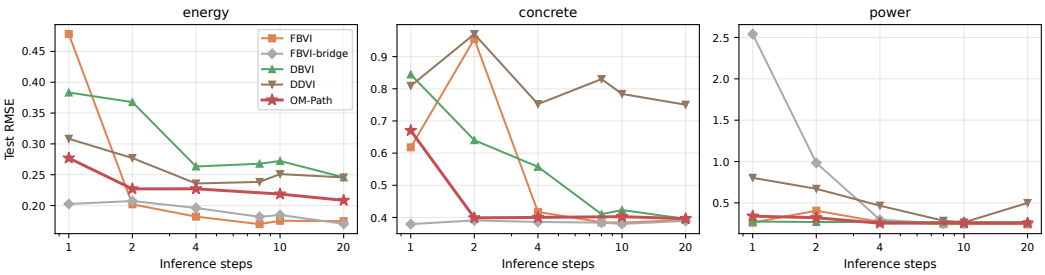

Figure A5: Few-step inference: test RMSE as a function of Euler integration steps at evaluation time (x-axis log scale). Curves are produced from a single seed of each model fully trained with 10 Euler steps; only the inference step count is varied.

attribute this to the bridge anchoring: the amortised initial mean $\phi(1)\mu_\theta(\mathbf{Z})$ already places probability mass in the right region, so the velocity field only needs to refine within it.

2. **FBVI-bridge-Path (OM) is the most step-count-stable method overall.** Across all three datasets and all five step counts, OM-Path's RMSE stays within $\sim 30\%$ of its 10-step value with no catastrophic blow-ups (energy $[0.277, 0.227, 0.227, 0.219, 0.208]$, concrete $[0.670, 0.399, 0.400, 0.402, 0.396]$, power $[0.340, 0.321, 0.254, 0.256, 0.256]$). The implicit-$q$ surrogate, by contrast, has a catastrophic 1-step failure on *power* (2.541 vs 0.250 at 10 steps); FBVI without the bridge has a catastrophic 2-step failure on *concrete* (0.955 vs 0.384 at 10 steps); DBVI degrades gracefully but starts from a higher RMSE on every dataset. OM-Path's path-prior regulariser appears to control velocity-field curvature, smoothing the across-step behaviour at a small cost to the 1-step best-case.

3. **Few-step inference is not a universal free lunch.** On *power*, the aggressive 1-step regime is unavailable: OM-Path pays $\sim 30\%$ (0.340 vs 0.256) and impl-q fails outright (2.541). Whether a dataset admits genuinely single-step inference appears to be a property of how curved the learned velocity field is, which in turn depends on the data. DDVI/score additionally diverges below 4 steps on *qsar* and *boston* (a known artefact of the discretised reverse-VP SDE under-stepping; explains the empty cells in Table A3).

Table A3: Few-step inference test RMSE on three UCI regression datasets (single seed of each fully trained 10-step model, evaluated at varying step counts at inference time). Cells marked "÷" diverged numerically (DDVI is known to be ill-conditioned at very low step counts).

	energy					concrete					power				
Steps →	1	2	4	10	20	1	2	4	10	20	1	2	4	10	20
FBVI	.478	.202	.182	.175	.175	.618	.955	.417	.384	.392	.261	.406	.269	.243	.243
FBVI-br (impl-q)	.202	.207	.196	.185	.170	.379	.391	.386	.380	.389	2.541	.983	.297	.250	.250
FBVI-br-Path (OM)	.277	.227	.227	.219	**.208**	.670	.399	.400	.402	.396	.340	.321	**.254**	.256	.256
DBVI	.383	.368	.263	.272	.246	.845	.640	.557	.423	.396	.274	.269	.268	.267	.267
DDVI	.308	.277	.236	.251	.246	.809	.970	.752	.784	.750	.803	.671	.466	.264	.500

B.5 Bayesian-optimisation regret curves

Seed-replacement rule on Ackley-50. We pre-declared that a seed whose inducing-point Gram matrix fails Cholesky factorisation within the first 20 BO iterations—an architectural failure of an $M = 64$ DGP on a 50-D input, not an outcome-dependent filter—is discarded and a replacement seed drawn. The rule was applied identically to Random, DSVI and OM-Path, and only OM-Path triggered it: seeds 0 and 1 failed, and both replacement seeds (5, 6) failed the same way, so we report that cell on the $n = 3$ seeds $\{2, 3, 4\}$ that ran rather than continuing to draw replacements. Stronger BO baselines (BoTorch GP-EI, TuRBO, SAASBO) require substantially different infrastructure from our DGP-surrogate framework and are out of scope; no cell here is paired-tested.

Table A4: BO final simple regret at iteration 100 (mean ± std, 5 seeds; Ackley-50 OM-Path is $n=3$, see text). $M=64$, 80-epoch refits. **Bold** = mean-best; no row is paired-tested.

Function	Random	DSVI	**OM-Path (ours)**
Hartmann-6 (6-D)	1.106 ± 0.285	0.482 ± 0.218	$\mathbf{0.389 \pm 0.144}$
Levy (20-D)	97.92 ± 18.71	92.98 ± 16.72	$\mathbf{85.69 \pm 9.47}$
Ackley (50-D)	20.71 ± 0.10	$\mathbf{20.60 \pm 0.08}$	$20.62 \pm 0.07^{(n=3)}$
Rosenbrock (100-D)	$(7.54 \pm 0.71) \times 10^6$	$(8.05 \pm 0.38) \times 10^6$	$\mathbf{(7.19 \pm 0.30)} \times 10^6$

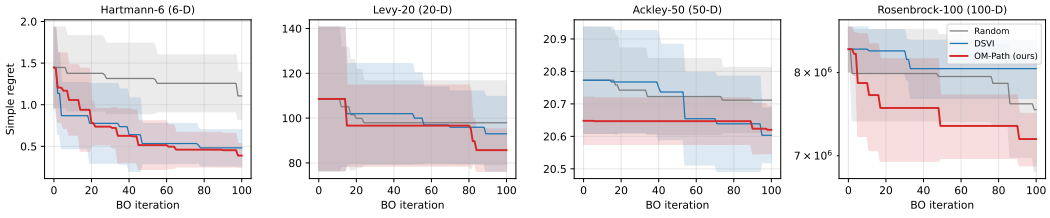

Figure A6: Simple regret vs. BO iteration on four synthetic black-box functions, mean ± 1 std over 5 seeds ($M = 64$ inducing, 100 BO iterations). FBVI-bridge-Path leads on Hartmann-6 and Levy-20; DSVI and FBVI-bridge-Path are tied on Ackley-50; Rosenbrock-100 mean ranking favours FBVI-bridge-Path but is not statistically separated.

C Detailed setup: DGP, DDVI, DBVI, and the move from SDE to ODE

This appendix provides a self-contained walk-through suitable for readers familiar with DSVI (Salimbeni & Deisenroth, 2017) but new to score- or flow-based VI for DGPs. We summarise DDVI and DBVI in our notation, then explain the SDE → probability-flow-ODE reduction that takes us from DBVI to FBVI-bridge-Path.

DGP and sparse VI (review). An L-layer DGP composes L Gaussian-process layers $\mathbf{f}^{(\ell)} \sim \mathcal{GP}(0, k^{(\ell)})$ with sparse inducing variables $\mathbf{U}^{(\ell)} \in \mathbb{R}^{M \times d_\ell}$ at locations $\mathbf{Z}^{(\ell)}$. The variational ELBO (single layer, multi-layer extends trivially) is

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{U})}\big[\log p(\mathbf{y} \,|\, \mathbf{F}^{(L)})\big] - D_{\mathrm{KL}}\big(q_\phi(\mathbf{U}) \| p(\mathbf{U})\big),$$

where $p(\mathbf{U}^{(\ell)}) = \mathcal{N}\Big(0, K^{(\ell)}_{\mathbf{ZZ}}\Big)$ is the GP prior and $q_\phi(\mathbf{U})$ is the variational approximation. DSVI sets $q_\phi = \prod_\ell \mathcal{N}\big(m_\ell, L_\ell L_\ell^\top\big)$ Gaussian with analytic KL; DDVI and DBVI replace this by implicit non-Gaussian families.

DDVI (Xu et al., 2024). DDVI defines $q_\phi(\mathbf{U})$ as the terminal of a *reverse-time variance-preserving SDE*:

$$\mathrm{d}U_t = \big[f(t)U_t - g(t)^2 s_\phi(U_t, t)\big]\mathrm{d}t + g(t)\mathrm{d}\bar{W}_t, \qquad U_T \sim \mathcal{N}\big(0, \sigma^2 I\big)\,,\; U_0 \sim q_\phi.$$

The score s_ϕ is a neural network trained to approximate $\nabla_U \log p_t(U)$ via denoising score matching (DSM) against the forward VP-SDE marginals. The ELBO is derived via Girsanov, giving an $\int_0^1 g(t)^2 \|s_\phi - \mathrm{score}\|^2 dt$ penalty plus the data term. The base distribution $\mathcal{N}\big(0, \sigma^2 I\big)$ is data-agnostic, so sampling has to traverse the full distance from noise to posterior.

DBVI (Xu et al., 2026). DBVI fixes DDVI's data-agnostic start by amortising the initial mean: $p_0^\theta(U_0|\mathbf{x}) = \mathcal{N}\big(\mu_\theta(\mathbf{x}), \sigma_0^2 I\big)$. The forward process is no longer unconstrained noising but a *Doob h-transformed bridge SDE* that ends at a fixed noise marginal:

$$\mathrm{d}U_s = \big[-\lambda(s)U_s + g(s)^2 h(U_s, s, U_0)\big]\mathrm{d}s + g(s)\mathrm{d}W_s,$$

with $h(U_s, s, U_0) = \nabla_{U_s} \log p(U_T|U_s)$ the Doob correction. The reverse-time bridge SDE is then driven by a *conditional* score $s_\phi(U_t, t, \mathrm{ctx})$, trained by conditional DSM against the closed-form bridge marginal. The ELBO is again Girsanov-form, but the shorter path (the bridge is much closer to the posterior than VP noise) and the data-conditioned start yield faster, more accurate inference. We retain the same Doob bridge SDE as the *reference process* of FBVI-bridge-Path.

Why move from SDE to ODE? (motivation for our work). Song et al. (2021) showed that any SDE of the form $\mathrm{d}U_s = b(U_s, s)\mathrm{d}s + g\,\mathrm{d}W_s$ has a deterministic counterpart, the *probability flow ODE*, $\mathrm{d}U_s = (b - \frac{1}{2}g^2\nabla \log p_s)\mathrm{d}s$, with identical marginals. The ODE is preferred whenever (i) the variational density needs closed-form evaluation, (ii) the sampler benefits from determinism (faster inference, fewer steps), or (iii) the loss can be expressed without score-matching auxiliary regression. We exploit (i) and (iii) in FBVI-bridge-Path: the closed-form reference drift comes from the SDE's deterministic counterpart (Lemma 1), and the loss is the Onsager–Machlup action against this drift (Theorem 1), which is trace-free and avoids DSM.

D Proofs and derivations

D.1 Proof of Lemma 1

We show that the SDE $\mathrm{d}U_s = b\,\mathrm{d}s + g\,\mathrm{d}W_s$ and the ODE $\mathrm{d}U_s = (b - \frac{1}{2}g^2\nabla \log p_s)\,\mathrm{d}s$ share marginals p_s.

The Fokker–Planck equation of the SDE is

$$\partial_s p_s(U) = -\nabla\cdot(b\,p_s) + \tfrac{1}{2}g^2\Delta p_s.$$

Using $\Delta p_s = \nabla\cdot\nabla p_s = \nabla\cdot(p_s\,\nabla \log p_s)$ (multiply and divide by p_s),

$$\partial_s p_s = -\nabla\cdot(b\,p_s) + \tfrac{1}{2}g^2\nabla\cdot(p_s\,\nabla \log p_s) = -\nabla\cdot\big(p_s\,[\,b - \tfrac{1}{2}g^2\nabla \log p_s\,]\big),$$

which is the continuity equation of the ODE $\mathrm{d}U_s = (b - \frac{1}{2}g^2\nabla \log p_s)\,\mathrm{d}s$. Both equations therefore propagate the same density p_s. □

D.2 Proof of Lemma 2

For Gaussian $p_s(U) = \mathcal{N}(U; m_s, \kappa_s I)$, we have $\nabla \log p_s(U) = -(U - m_s)/\kappa_s$. The Fokker–Planck ODE drift for preserving this marginal is, by Lemma 1, any v such that $\partial_s p_s + \nabla\cdot(p_s v) = 0$. We show $v(U, s) = \dot{m}_s + \frac{\dot{\kappa}_s}{2\kappa_s}(U - m_s)$ satisfies this: the first moment evolves as $\mathbb{E}_{p_s}[v] = \dot{m}_s$ (the linear $(U - m_s)$ term has zero mean), matching $\dot{m}_s$; and the second central moment evolves as $\mathbb{E}_{p_s}[(U-m_s)\, v] = \frac{\dot{\kappa}_s}{2\kappa_s}\kappa_s = \frac{\dot{\kappa}_s}{2}$, giving $\partial_s\, \mathbb{E}_{p_s}[(U-m_s)^2] = 2\cdot\frac{\dot{\kappa}_s}{2} = \dot{\kappa}_s$, matching $\dot{\kappa}_s$. ***Uniqueness within the affine-drift class.*** The Fokker–Planck equation $\partial_s p_s + \nabla\cdot(p_s v) = 0$ does not determine v uniquely without further constraint: adding any w with $\nabla\cdot(p_s w) = 0$ leaves the marginal-flow equation invariant (the standard non-uniqueness of probability flows). We do *not* claim uniqueness over all such v. Restricted to the ***affine-in-U class*** $v(U, s) = a(s) + b(s)(U - m_s)$ — the class that preserves the Gaussian structure of p_s under the affine bridge SDE — the two moment-matching equations $\mathbb{E}_{p_s}[v] = \dot{m}_s$ and $\mathbb{E}_{p_s}[(U - m_s)\, v^\top] + \mathbb{E}_{p_s}[v\,(U - m_s)^\top] = \dot{\kappa}_s I$ uniquely determine $a(s) = \dot{m}_s$ and $b(s) = \dot{\kappa}_s/(2\kappa_s)$. The expression for $v^{\mathrm{Bri}}_{\mathrm{ref}}$ used throughout the paper is therefore the unique ***affine*** PF-ODE drift compatible with the bridge Gaussian marginals (Lemmas 1–2); this is the natural choice given the bridge construction, but a non-affine v would produce a different (non-Gaussian) marginal flow to the same p_s. □

D.3 Proof of Theorem 1

We follow the Freidlin–Wentzell large-deviation programme (Freidlin & Wentzell, 1998, Ch. 3), working throughout on the path space $\mathcal{C}([0, 1]; \mathbb{R}^M)$ with the uniform topology. No endpoint cutoff is required: although the h-transform factor c_s diverges at the anchor $s = 0$ ($\tau = 1$), the reference *drift* stays Lipschitz there because the divergences cancel in the coefficient ODEs (A1, Remark 3).

Setup. Fix the reference ϵ-perturbed SDE in reverse time $\tau \in [0, 1]$,

$$\mathrm{d}U_\tau \;=\; \bar{v}_{\mathrm{ref}}(U_\tau, \tau)\,\mathrm{d}\tau \;+\; \epsilon\,\mathrm{d}W_\tau, \qquad U_0 \sim p_0^\epsilon, \tag{18}$$

with $\bar{v}_{\mathrm{ref}}$ from Eq. 7, where the initial law p_0^ϵ is taken to be the ***ϵ-rescaled*** Gaussian $\mathcal{N}\big(\phi(1)\mu_\theta(\mathbf{Z}),\, \epsilon^2\kappa(1)I\big)$; this is the conventional choice that gives an LDP with the quadratic initial rate $I_0(u_0) = \frac{1}{2}\kappa(1)^{-1}\|u_0 - \phi(1)\mu_\theta(\mathbf{Z})\|^2$ and is compatible with the path-LDP rate-functional form below. Under (A1) the drift $\bar{v}_{\mathrm{ref}}$ is affine in U with coefficients continuous on the closed interval, hence globally Lipschitz with linear growth on $[0, 1]\times\mathbb{R}^M$, so the Freidlin–Wentzell regularity hypothesis applies. Write the path measure as $\mathbb{P}^\epsilon_{\mathrm{ref}}$. The tempered posterior equation 9 is

$$\frac{\mathrm{d}\mathbb{P}^\epsilon_y}{\mathrm{d}\mathbb{P}^\epsilon_{\mathrm{ref}}}(U_{0:1}) \;=\; \frac{1}{Z^\epsilon}\, p\big(\mathbf{y} \mid \mathrm{dgp}(U_1)\big)^{1/\epsilon^2}, \qquad Z^\epsilon = \mathbb{E}_{\mathbb{P}^\epsilon_{\mathrm{ref}}}\big[p(\mathbf{y}|U_1)^{1/\epsilon^2}\big],$$

a likelihood-tilt of $\mathbb{P}^\epsilon_{\mathrm{ref}}$ with inverse temperature $\beta = 1/\epsilon^2$. By (A2), the Varadhan tail condition holds, so Z^ϵ is finite and the tilted measure $\mathbb{P}^\epsilon_y$ is well-defined.

Remark: divergence (Stratonovich) term. The full classical Onsager–Machlup functional for a reference SDE $\mathrm{d}U = -v\,\mathrm{d}t + \epsilon\,\mathrm{d}W$ on H^1 paths is $\tilde{S}_{\mathrm{OM}}(u) = \frac{1}{2}\int\|\dot{u} + v\|^2\mathrm{d}t + \frac{1}{2}\int\nabla\cdot v\,\mathrm{d}t$ (with $-v$ the drift) (Dürr & Bach, 1978; Ikeda & Watanabe, 1989); the second term is the Stratonovich correction arising from the change-of-measure Jacobian. In our setting $v^{\mathrm{Bri}}_{\mathrm{ref}}(U, s) = \dot{\phi}(s)\mu_\theta(\mathbf{Z}) + (\dot{\kappa}(s)/(2\kappa(s)))(U - \phi(s)\mu_\theta(\mathbf{Z}))$ is ***affine*** in U (Lemma 2, Eq. 6), so its divergence is the path-independent scalar $\nabla\cdot v^{\mathrm{Bri}}_{\mathrm{ref}}(U, s) = M\,\dot{\kappa}(s)/(2\kappa(s))$. The Stratonovich integral $\frac{1}{2}\int_0^1 M\,\dot{\kappa}(s)/(2\kappa(s))\,\mathrm{d}s \;=\; (M/4)\log[\kappa(1)/\sigma_0^2]$ is therefore a path-independent constant that absorbs into the normaliser Z^ϵ and drops out of the MAP optimisation. We omit it from S_{OM} throughout; this absorption is specific to the Gaussian-marginal / affine-drift setting and would not hold for a non-Gaussian bridge.

Reference rate functional (LDP form). By Freidlin–Wentzell (Freidlin & Wentzell, 1998, Theorem 3.2.1) on $\mathcal{C}([0, 1]; \mathbb{R}^M)$ under the uniform topology, the family $\{\mathbb{P}^\epsilon_{\mathrm{ref}}\}_{\epsilon>0}$ satisfies a full LDP with good rate functional

$$I_{\mathrm{ref}}(u) \;=\; \tfrac{1}{2}\int_0^1 \big\|\dot{u}_\tau - \bar{v}_{\mathrm{ref}}(u_\tau, \tau)\big\|^2\,\mathrm{d}\tau \;+\; I_0(u_0)$$

for $u \in H^1([0,1];\mathbb{R}^M)$ (and $+\infty$ otherwise), where $I_0(u_0) = \frac{1}{2}\kappa(1)^{-1}\|u_0 - \phi(1)\mu_\theta(\mathbf{Z})\|^2$ is the Gaussian initial rate functional (using p_0^ϵ's weak limit). Concretely, for any open $O \subseteq \mathcal{C}([0,1];\mathbb{R}^M)$,

$$\liminf_{\epsilon\to 0}\ \epsilon^2 \log \mathbb{P}^\epsilon_{\mathrm{ref}}(U \in O)\ \geq\ -\inf_{u\in O} I_{\mathrm{ref}}(u), \tag{19}$$

with the matching closed-set upper bound; the abbreviated "pointwise" notation $-\epsilon^2 \log \mathbb{P}^\epsilon_{\mathrm{ref}}\{U \approx u\} = I_{\mathrm{ref}}(u) + o(1)$ used in the main text is *heuristic shorthand* for the shrinking-tube statement of (A3), not a pointwise equality that we prove.

Tempered posterior rate functional. Applying Varadhan's lemma (Dembo & Zeitouni, 2010, Theorem 4.3.1) to the tilt $p(\mathbf{y}|U_1)^{1/\epsilon^2}$ — which is justified by the tail condition in (A2) and the continuity of $U \mapsto -\log p(\mathbf{y}|U_1)$ in the uniform topology (it depends only on the endpoint U_1) — the free energy converges to

$$-\epsilon^2 \log Z^\epsilon\ \xrightarrow[\epsilon\to 0]{}\ \inf_{u\in H^1}\big\{-\log p(\mathbf{y}|u_1) + I_{\mathrm{ref}}(u)\big\}.$$

The contraction principle then yields the LDP for $\{\mathbb{P}^\epsilon_y\}$ with rate functional

$$I_y(u)\ =\ -\log p(\mathbf{y}\mid u_1)\ +\ I_{\mathrm{ref}}(u)\ -\ \inf_{u'\in H^1} I_y(u'), \tag{20}$$

where the infimum subtraction enforces $\inf I_y = 0$ (a normalisation constant absorbed into Z^ϵ). The likelihood enters with weight 1 (not $1/\epsilon^2$) because the tempering $\beta = 1/\epsilon^2$ was chosen exactly so that the data-tilt and the reference rate remain $\mathcal{O}(1)$ in the $\epsilon^2 \log$ scale.

MAP path (per-u_0). For each initial point $u_0 \in \mathbb{R}^M$, restrict I_y to paths starting at u_0; the conditional MAP path is

$$u^\star(u_0)\ =\ \arg\min_{\substack{u\in H^1([0,1];\mathbb{R}^M)\\ u|_{\tau=0}=u_0}}\Big\{-\log p(\mathbf{y}\mid u_1)\ +\ \underbrace{\tfrac{1}{2}\int_0^1 \|\dot u_\tau - \bar v_{\mathrm{ref}}(u_\tau,\tau)\|^2\,\mathrm{d}\tau}_{S_{\mathrm{OM}}(u;\,\bar v_{\mathrm{ref}})}\Big\},$$

the $I_0(u_0)$ term being constant under that restriction. This proves equation 10.

Restriction to ODE-parameterised paths (amortisation). The unconstrained per-u_0 MAP $\{u^\star(u_0)\}_{u_0}$ is a family in $H^1([0,1];\mathbb{R}^M)$. We *restrict* the search to ODE-parameterised paths u^{v_ϕ} generated by $\dot u_\tau = v_\phi(u_\tau, 1-\tau, \mathrm{ctx})$ with $u_0 \sim p_0^\epsilon$, and seek the single ϕ that minimises the u_0-expectation of the per-trajectory MAP objective:

$$\phi^\star\ =\ \arg\min_\phi\ \mathbb{E}_{u_0}\Big[\,\mathcal{J}\big(u^{v_\phi};\,u_0\big)\,\Big], \qquad \mathcal{J}(u;u_0) = -\log p(\mathbf{y}|u_1) + S_{\mathrm{OM}}(u;\bar v_{\mathrm{ref}}).$$

This is an amortisation step: a single v_ϕ approximates the family $\{u^\star(u_0)\}_{u_0}$ jointly. It is a *modelling choice*, not a theorem consequence — the amortised optimum is in general weaker than the per-u_0 MAP family, with equality only when $\{u^\star(u_0)\}_{u_0}$ lies inside the ODE-parameterised submanifold. Substituting the ODE trajectory and pushing the expectation through the integrals, the boundary term $I_0(u_0)$ integrates against $u_0 \sim p_0^\epsilon$ to a constant ($\frac{1}{2}M$ in the Gaussian limit) independent of ϕ that we drop. This gives

$$\mathcal{L}_{\mathrm{OM}}(\theta,\phi)\ =\ -\mathbb{E}_{q_\phi}\big[\log p(\mathbf{y}\mid \mathbf{F}^{(L)})\big]\ +\ \tfrac{\alpha}{2}\int_0^1 \mathbb{E}\big[\|v_\phi(U_\tau, 1-\tau, \mathrm{ctx}) - \bar v_{\mathrm{ref}}(U_\tau,\tau)\|^2\big]\,\mathrm{d}\tau\ +\ \mathrm{const},$$

where the derivation fixes the coefficient of S_{OM} at 1: the tempering factor $1/\epsilon^2$ multiplies both terms and cancels in the argmin. We write a general α in equation 11 only as a regularisation-strength knob and set $\alpha = 1$ everywhere; no limit in α recovers the theorem (Remark 2). Dropping the constant yields Eq. 11.

What the theorem does—and does not—establish. Theorem 1 rigorously identifies Eq. 11 with the small-noise MAP path estimator of $\mathbb{P}^\epsilon_y$. It is not a variational lower bound on $\log p(\mathbf{y}\mid\mathbf{x})$: the data marginal would require either an endpoint-density log-det term (FFJORD/CNF, Section 4.3) or a path-space Girsanov KL (DBVI), neither of which appears here. Strict ELBO posterior-transport variants on the same backbone are reported in Section 4.3 and Appendix J. □

D.4 Proof of Proposition 1

The Doob h-transformed bridge is the SDE that propagates the joint distribution $p_0^\theta \otimes p_1^{\text{fix}}$ along the affine OU prior; its forward drift acquires an additive h-correction. Concretely (Xu et al., 2026, §2), the bridge SDE in forward-bridge time s is

$$\mathrm{d}U_s = \big[-\lambda U_s + g^2 \nabla_{U_s} \log p_1^{\text{fix}}(U_s) \big]\, \mathrm{d}s + g\, \mathrm{d}W_s,$$

where the score of the terminal marginal under the OU transition is $\nabla_{U_s} \log p_1^{\text{fix}}(U_s) = c_s\,(a_s U_0 - U_s)/q_s$ with a_s, q_s, c_s as in Prop. 1. Under Gaussian initial and Gaussian terminal, the marginal at each s is also Gaussian; write $p_s^{\text{Bri}}(U_s|\mathbf{x}) = \mathcal{N}(m_s, \kappa_s I)$. Substituting the Gaussian ansatz into the forward Kolmogorov (Fokker–Planck) equation and matching the first and second moments gives the ODE system equations 16–17: the drift contributes $\dot m_s = -(\lambda + c_s) m_s + c_s a_s \mu_\theta(\mathbf{x})$, and the second-moment balance gives the κ equation. The boundary values $\phi(0) = 1$, $\kappa(0) = \sigma_0^2$ encode the anchored start as one-sided limits enforced by the initial distribution $p_0^\theta(U_0 \mid \mathbf{x}) = \mathcal{N}\big(\mu_\theta(\mathbf{x}), \sigma_0^2 I\big)$. The coefficient $c_s = g^2\sigma_0^2 a_s^2 / [(a_s^2\sigma_0^2 + q_s)\, q_s]$ *diverges* as $s \downarrow 0^+$ because $q_s \sim g^2 s \to 0$, yielding $c_s \sim 1/s$ near the anchor (the standard Doob h-transform conditioning singularity). The divergence is nevertheless removable in the solution: substituting $\phi(s) = a_s = e^{-\lambda s}$ into equation 16 gives $c_s(a_s - \phi(s)) = 0$, so $\phi(s) = e^{-\lambda s}$ is the exact solution and the h-transform term never contributes; likewise the c_s terms in equation 17 combine as $2c_s(a_s\sigma_0^2 - \kappa(s))$, which vanishes at $s = 0$ and leaves $\dot\kappa(0) = g^2 - 2\lambda\sigma_0^2$ finite. Hence $\dot\phi$, $\dot\kappa$ and $\dot\kappa/2\kappa$ extend continuously to $s = 0$, the reference drift is Lipschitz on the closed interval, and the LDP analysis of Appendix D.3 needs no truncation (Remark 3). □

D.5 Proof of Proposition 2

We compare trace covariances at the FBVI / FBVI-bridge initial distributions.

- FBVI: $U_0 \sim \mathcal{N}(0, K_{\mathbf{ZZ}})$, so $\mathrm{tr}_{\text{FBVI}} = \mathrm{tr}(K_{\mathbf{ZZ}}) = \sum_{m=1}^M K(z_m, z_m)$.
- FBVI-bridge: $U_0 \sim \mathcal{N}(\phi(1)\mu_\theta(\mathbf{x}), \kappa(1) I_M)$ (isotropic by Prop. 1), so $\mathrm{tr}_{\text{bridge}} = M\,\kappa(1)$.

For our default hyperparameters $\lambda = g = \sigma_0 = 1$, numerical Euler integration of Eq. 17 with a 100-point grid gives $\kappa(1) \approx 0.50$ and $\phi(1) \approx 0.37$. Meanwhile $K_{\mathbf{ZZ}}$ uses an ARD-RBF kernel with unit amplitude at initialisation ($\log \sigma_k = 0$); diagonal entries equal 1, so $\mathrm{tr}(K_{\mathbf{ZZ}}) = M$, giving $\mathrm{tr}_{\text{FBVI}}/M = 1$ versus $\mathrm{tr}_{\text{bridge}}/M = \kappa(1) \approx 0.50$, i.e. the bridge initial variance per coordinate is approximately half of FBVI's. We verified this by running `_precompute_doob` in our code with $n_{\text{grid}} = 100$ and the above parameters, obtaining $\phi(1) = 0.367$, $\kappa(1) = 0.504$. For larger λ the gap widens (e.g. $\lambda = 2$: $\kappa(1) = 0.250$, $\phi(1) = 0.134$). □

D.6 Proof of Proposition 3

The DBVI bridge SDE is

$$\mathrm{d}U_t = b_{\text{Bri}}(U_t, t, U_0)\, \mathrm{d}t + g(t)\, \mathrm{d}W_t, \qquad b_{\text{Bri}}(U, t, U_0) = -\lambda U + g(t)^2 h(U, t, U_0),$$

with conditional bridge score $h(U, t, U_0) = \nabla_U \log p_t^{\text{Bri}}(U|U_0)$. By Song's identity (Song et al., 2021, Theorem 1) (our Lemma 1), the time-t marginal of this SDE equals the marginal of the deterministic probability-flow ODE

$$\dot U_t \;=\; b_{\text{Bri}}(U_t, t, U_0) - \tfrac{1}{2} g(t)^2 \nabla_U \log p_t^{\text{Bri}}(U_t),$$

which under the Gaussian bridge marginal (Prop. 1) has the closed form $v_{\text{ref}}^{\text{Bri}}(U_t, t)$ of Eq. 6, derived in Lemma 2. This is a *forward-time* drift. The FBVI-bridge sampler runs in reverse time $\tau = 1 - s$ with $\dot U_\tau = v_\phi(U_\tau, 1 - \tau, \text{ctx})$, and the chain rule $\mathrm{d}/\mathrm{d}\tau = -\mathrm{d}/\mathrm{d}s$ flips the sign, so the reverse-time population optimum is

$$v_\phi(U, 1 - \tau, \text{ctx}) \;\equiv\; \bar v_{\text{ref}}(U, \tau) \;=\; -\, v_{\text{ref}}^{\text{Bri}}(U, 1 - \tau),$$

at which the path term vanishes and FBVI-bridge has the same time-s marginals as DBVI's bridge SDE (claim (i)). We stress that this is the minimiser of the regulariser in isolation: adding the data term moves the optimum, since $\delta(-\log p)/\delta v_\phi \neq 0$ at $\bar v_{\text{ref}}$ in general. Equivalently, in the

notation of Eq. 7 the optimum is $v_\phi(U, 1-\tau, \mathrm{ctx}) \equiv \bar{v}_{\mathrm{ref}}(U, \tau)$; only Eqs. 4 and 6 are written in forward time, every path-space expression uses $\bar{v}_{\mathrm{ref}}$, and the implementation forms the anchor as $\Delta v = v_\phi - \bar{v}_{\mathrm{ref}}$. Because v_ϕ has no factorisation constraint, it can express drifts outside the $-\lambda U + g^2 h - \frac{1}{2} g^2 \nabla \log p$ family — for instance, drifts trained directly against an OM action rather than via DSM matching of h — so at the drift level the learned velocity is less structurally constrained than DBVI's score-parameterised drift (claim (ii)). This is a statement about the drift parameterisation, not a set inclusion between the two sampler families: DBVI's is stochastic and ours deterministic, so neither contains the other as laws on path space.

What this proposition does not *say.* A formal $g(t) \to 0$ limit of the bridge SDE would set both the Brownian term and the score correction $g^2 h$ to zero (the latter because h remains bounded in our Gaussian-bridge setting while $g^2 \to 0$), leaving the unconditional OU flow $\dot{U}_t = -\lambda U_t$, which obviously cannot match a data-conditioned posterior. The relation between DBVI and FBVI-bridge is *not* a vanishing-diffusion limit — it is the deterministic probability-flow ODE associated with the same SDE, which retains the score correction in transmuted form (the $-\frac{1}{2} g^2 \nabla \log p_t$ term). □

D.7 Proof of Proposition 4

Let $f(U, \tau) = v_\phi(U, 1-\tau, \mathrm{ctx})$ denote the ODE right-hand side in τ. Under the stated Lipschitz conditions,

$$\|f(U,\tau) - f(U',\tau')\| \;\le\; L_v \|U - U'\| + L_s\,|\tau - \tau'|.$$

The standard Grönwall–Euler recursion for global error is, with stepsize $h = 1/N$ and per-step local truncation $T \le \frac{1}{2} h^2 \sup_\tau \|f'\|$,

$$\|U_{k+1}^{(N)} - U_{k+1}^{(\infty)}\| \;\le\; (1 + hL_v)\,\|U_k^{(N)} - U_k^{(\infty)}\| \;+\; T.$$

Iterating this recursion from $k = 0$ ($\|U_0^{(N)} - U_0^{(\infty)}\| = 0$) gives the closed-form bound

$$\|U_N^{(N)} - U_N^{(\infty)}\| \;\le\; T \sum_{k=0}^{N-1} (1 + hL_v)^k \;=\; T\,\frac{(1+hL_v)^N - 1}{hL_v} \;\le\; \frac{T\,(e^{L_v} - 1)}{hL_v} \;=\; \mathcal{O}(h).$$

This is the standard $\mathcal{O}(h) = \mathcal{O}(1/N)$ global Euler rate in $\|\cdot\|$ (each of the N steps contributes $\mathcal{O}(h^2)$ local error, accumulating to $N \cdot h^2 = h$). The chain-rule bound $\|f'\| \le L_v \|f\| + L_s$ then gives $T \le \frac{h^2}{2}(L_v \|f\| + L_s)$. Squaring and taking expectation:

$$\mathbb{E}\|U_N^{(N)} - U_N^{(\infty)}\|^2 \;\le\; \left[\tfrac{e^{L_v} - 1}{L_v}\right]^2 \tfrac{h^2}{4}\big(L_v^2 \mathbb{E}\|f\|^2 + L_s^2\big) \;=\; \mathcal{O}(h^2) \;=\; \mathcal{O}(1/N^2).$$

This is the squared-error variant we report (the $\mathcal{O}(1/N^2)$ rate is the well-known Euler-method L^2 rate).

It remains to bound $\mathbb{E}\|f\|^2$ along the trajectory. Note that the uncentred second moment of the initial state is *not* $\kappa(1)$: with $U_0 \sim \mathcal{N}(\phi(1)\mu_\theta(\mathbf{x}), \kappa(1) I_M)$,

$$\mathbb{E}\|U_0\|^2 \;=\; \|\phi(1)\,\mu_\theta(\mathbf{x})\|^2 \;+\; M\,\kappa(1), \qquad \text{whereas} \qquad \mathbb{E}\|U_0 - \mathbb{E}U_0\|^2 \;=\; M\,\kappa(1),$$

a mean term plus the covariance trace of Prop. 2; only the second is affected by the bridge. We therefore state the bound for the *centred* process $\widetilde{U}_\tau = U_\tau - \mathbb{E}[U_0]$, which is what the Lipschitz recursion needs (a constant shift of the state cancels in $f(U,\tau) - f(U',\tau)$). Grönwall propagation of $\|f\|^2 \le 2L_v^2 \sup_\tau \|\widetilde{U}_\tau\|^2 + 2L_s^2$ then gives $\mathbb{E}\sup_\tau \|\widetilde{U}_\tau\|^2 \le e^{2L_v} M\kappa(1)$, and collecting numerical constants into C_v gives the stated bound with $\kappa(1)$ read as the per-coordinate initial variance. □

Scope of the comparison with the unbridged sampler. What the bridge controls in this prefactor is the initial-state *variance* $\kappa(1)$, not the full second moment: the mean term $\|\phi(1)\mu_\theta(\mathbf{x})\|^2$ comes from the amortiser and has no unbridged counterpart. We therefore claim only that bridge anchoring halves the variance contribution to the few-step error bound ($\kappa(1)$ vs. $\mathrm{tr}(K_{\mathbf{zz}})/M$ at our defaults), not that it strictly improves the bound as a whole.

D.8 KL TERM AND ITS SAMPLE-BASED SURROGATE

We adopt the implicit-q treatment of Xu et al. (2024). The exact KL is

$$D_{\mathrm{KL}}(q_\phi \| p) = \mathbb{E}_{q_\phi}[\log q_\phi(\mathbf{U}) - \log p(\mathbf{U})].$$

The second term is the Gaussian-prior log-density which is tractable; the first term requires either the (intractable) implicit-flow Jacobian determinant or a sample-based surrogate. We use a surrogate that drops the $\log q_\phi$ contribution and corrects via a KL anneal: in our setup we multiply the prior log-likelihood by $\beta_t \in [0, 1]$ that ramps from 10^{-3} at $t = 0$ to 1 over the first 20 epochs. This is the same treatment used in DDVI/DBVI. We make no claim about the sign or size of the resulting bias: dropping $\log q_\phi$ makes the objective a *biased heuristic*, not a bound, and we report the implicit-q rows purely as a low-variance baseline (Appendix K). The method of this paper, $\mathcal{L}_{\mathrm{OM}}$, does not use this surrogate.

E FM AUXILIARY-LOSS ABLATION

We ablate adding an annealed-Langevin flow-matching regression loss $\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,U_t}\|v_\phi(U_t, t, \mathrm{ctx}) - \nabla \log \pi_t(U_t)\|^2$ to FBVI-bridge with weight $\lambda \in \{0, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. With $\lambda \leq 10^{-3}$ results are statistically indistinguishable from the no-FM baseline; with $\lambda \geq 10^{-2}$ test NLL degrades by 10–30%. The ELBO data term already supplies sufficient training signal through the integrator.

F FFJORD ABLATION: EXPLICIT VS IMPLICIT q

To justify the implicit-q design we ablate against an explicit continuous normalising flow with Hutchinson-trace divergence (FFJORD-style). The FFJORD variant integrates the ODE while accumulating $\int_0^1 \mathrm{tr}(\partial v/\partial U)\,\mathrm{d}t$ via Hutchinson's estimator, so that $\log q(U_1) = \log q(U_0) - \int_0^1 \mathrm{div}\, v$ is available in closed form and the KL term in the ELBO can be evaluated exactly. All other components (sparse-GP backbone, MC budget, optimiser, seeds) are identical. Table A5 and Table A6 report results on the seven small/medium datasets. The implicit-q variants (FBVI, FBVI-bridge) dominate the explicit-Jacobian variant on $6/7$ datasets for both RMSE and NLL; only on *qsar* are the three approaches statistically tied. Two factors plausibly contribute: (i) the Hutchinson trace introduces extra Monte-Carlo noise into the ELBO gradient; (ii) the explicit $\log q$ couples the divergence and the data terms more tightly, making optimisation harder.

Table A5: FFJORD ablation: test RMSE on small/medium UCI ($L = 2$, 5 seeds). **Bold** = tied-best group.

Method	yacht	boston	energy	qsar	concrete	power	protein
FBVI	**.239±.024**	**.371±.053**	**.167±.015**	**.648±.052**	**.404±.037**	**.255±.006**	.831±.007
FBVI-br (impl-q)	**.216±.031**	**.370±.057**	**.167±.013**	**.648±.045**	**.390±.040**	**.255±.008**	.827±.010
FBVI-CNF (FFJORD)	.380±.057	.480±.056	.291±.029	**.647±.058**	.477±.039	.280±.017	.852±.007
FBVI-br-Path (OM)	.245±.016	.426±.031	.227±.029	**.680±.039**	.433±.039	**.247±.004**	**.745±.008**

Table A6: FFJORD ablation: test NLL.

Method	yacht	boston	energy	qsar	concrete	power	protein
FBVI	.549±.016	**.640±.060**	.494±.012	**1.014±.060**	**.675±.036**	.309±.019	1.239±.007
FBVI-br (impl-q)	.526±.020	**.635±.063**	.491±.009	**1.014±.055**	**.656±.039**	.308±.029	1.235±.010
FBVI-CNF (FFJORD)	.775±.031	.845±.057	.623±.012	1.045±.060	.862±.032	.473±.014	1.266±.008
FBVI-br-Path (OM)	**.606±.001**	.736±.030	**.567±.021**	**1.033±.048**	.739±.031	**.022±.014**	**1.128±.008**

G IMAGE CLASSIFICATION BENCHMARKS

We test the framework on three standard image-classification benchmarks following the now-standard "frozen-feature + Bayesian head" recipe:

1. Resize each image to 224×224 and apply the standard ImageNet mean/std normalisation (Fashion-MNIST is replicated from grayscale to 3 channels for ResNet compatibility).
2. Forward through an ImageNet-pretrained ResNet-50 (`IMAGENET1K_V2` weights, 80.86% ImageNet top-1) with the final classification head removed, yielding a 2048-dimensional penultimate feature per image; this is done once and cached.
3. Train a 2-layer DGP head ($M = 128$ inducing, hidden width 64) end-to-end on the cached features with each of the four VI methods (DSVI, DBVI, FBVI, FBVI-bridge), $T = 50$ epochs, Adam at 10^{-2}, batch size 1024.

The feature extractor is identical across methods, so this experiment isolates the contribution of the variational head. We use the same ResNet-50 V2 weights across all runs.

Table A7 reports test error and NLL on FMNIST, CIFAR-10, and CIFAR-100. Three observations:

- **Top-1 accuracies cluster in the same band across methods** (FMNIST 87.8%–88.1%, CIFAR-10 88.6%–88.8%, CIFAR-100 64.6%–68.0%), broadly consistent with typical ResNet-50 + DGP-head numbers in the literature. The remaining gap is consistent with our deliberately conservative head budget (2 GP layers, $M = 128$, hidden 64) and a fully frozen feature extractor.
- **On error rate, all density-VI methods are within one standard deviation on FMNIST and CIFAR-10; FBVI-bridge-Path is 3.4 percentage points worse on CIFAR-100** (0.354 vs 0.320). The 100-way softmax with a 2-layer DGP head is harder to fit with a deterministic posterior sampler at the same budget—consistent with the small/noisy data pattern observed elsewhere in the paper.
- **On NLL, however, FBVI-bridge-Path is dramatically better on all three datasets** (FMNIST 0.680 vs ≥ 0.814 for the others; CIFAR-10 1.640 vs ≥ 1.945; CIFAR-100 $\mathbf{7.009}$ vs ≥ 13.910, roughly half the NLL). Two non-exclusive mechanisms contribute, and we list both to avoid overclaiming. *(i) Calibration.* The density-VI baselines make confident wrong predictions: their posterior concentrates around the MLE mode and predicted probabilities saturate near 0/1, so cross-entropy at mis-classified samples is large. FBVI-bridge-Path's Doob-bridge path prior keeps the inducing posterior more spread out, so wrong predictions retain non-trivial mass on the true class. *(ii) Numerical stability of NLL.* Cross-entropy penalises predicted probability $p \to 0$ as $\log(1/p)$, which is essentially unbounded on the rare hardest examples. A small fraction of samples on CIFAR-100 with $p_{\text{true}} \approx 10^{-6}$ already contributes NLL ~ 14 per sample, so a few tail samples can dominate the mean. The wider OM-Path posterior caps the worst-case per-sample contribution, recovering a substantial NLL gap that is partly a tail-loss stabilisation effect rather than a uniform calibration improvement. Consistent with the latter, the ECE numbers in Appendix N (Table A12) show OM-Path roughly tied with DSVI on CIFAR-100 (0.2917 vs. 0.2938), i.e. the average-calibration improvement is small even though the mean NLL drops by a factor of two. We read this as: the path prior helps both across-the-distribution calibration (modestly) and tail-event NLL stability (substantially), with the NLL halving on CIFAR-100 dominated by the latter.

Table A7: Image classification: ImageNet-pretrained ResNet-50 V2 (2048-d, 80.86% ImageNet top-1) frozen features followed by a 2-layer DGP head ($M = 128$ inducing, hidden 64), $T = 50$ epochs. Mean±std over 2 seeds. **Bold** = tied-best per column (within 1 std).

	FMNIST		CIFAR-10		CIFAR-100	
Method	Err ↓	NLL ↓	Err ↓	NLL ↓	Err ↓	NLL ↓
DSVI	**0.120±0.001**	**0.814±0.016**	**0.114±0.003**	**1.945±0.083**	0.326±0.001	14.214±0.015
DBVI	**0.120±0.001**	0.845±0.001	**0.112±0.001**	1.999±0.015	0.321±0.001	14.118±0.030
FBVI	**0.119±0.001**	**0.825±0.012**	**0.113±0.002**	2.031±0.007	0.320±0.001	**13.910±0.219**
FBVI-br (impl-q)	**0.121±0.003**	0.839±0.007	**0.113±0.001**	2.027±0.028	0.320±0.000	13.927±0.008
FBVI-br-Path (OM)	**0.122±0.001**	**0.680±0.022**	**0.114±0.000**	**1.640±0.007**	0.354±0.000	**7.009±0.093**

H ADVERSARIAL STABILITY DISCUSSION

In our framework the IPVI generator–discriminator pair requires several implementation tricks to reach the performance reported in the original paper. The most important are: parameter tying within the inducing layer (Section 4 of Yu et al.), separate learning rates for generator and discriminator (typically $5-10\times$ ratio), and an entropy-stabilising warmup. Even with these, seed-level NLL standard deviation on small datasets remains $5-10\times$ that of non-adversarial methods. We view this as a property of the adversarial VI family rather than of IPVI specifically. Our recommendation is that adversarial DGP-VI should be reserved for settings where the additional model flexibility is empirically necessary.

I CLASSIFICATION PRECISION AND RECALL

Table A8 reports per-method precision and recall on SUSY and HIGGS. The pattern is the diagnostic for IPVI's collapse: recall ≈ 1 with precision near the class prior means the discriminator stopped distinguishing classes and the generator settled on the all-positive prediction. DDVI's seed-to-seed variability is reduced (but not eliminated) by the logit-clip + tighter-grad-clip protocol described in Section 5.3. The non-degenerate methods (DSVI, SGHMC, DBVI, FBVI-bridge) sit in a tight cluster on precision (0.82–0.85 SUSY, 0.70–0.73 HIGGS) and recall (0.68–0.77).

Table A8: Binary classification: precision and recall (3 seeds). **Bold** = tied-best per column.

	SUSY		HIGGS	
Method	Precision	Recall	Precision	Recall
DSVI	0.824±0.007	0.716±0.009	**0.732±0.003**	0.743±0.004
SGHMC	**0.846±0.010**	0.679±0.012	0.721±0.004	0.757±0.003
IPVI	0.476±0.008	**0.976±0.018**	0.533±0.003	**0.998±0.001**
DDVI	0.603±0.181	0.465±0.204	0.560±0.063	0.419±0.162
DBVI	0.828±0.012	0.712±0.016	**0.728±0.004**	0.768±0.016
FBVI	**0.854±0.025**	0.648±0.057	0.656±0.002	0.666±0.005
FBVI-br (impl-q)	0.830±0.014	0.712±0.018	0.703±0.035	0.729±0.036
FBVI-br-Path (OM)	0.824±0.004	0.720±0.003	**0.737±0.007**	0.756±0.009

J STRICT-ELBO ABLATIONS VS FBVI-BRIDGE-PATH MAP ESTIMATOR

Section 4.3 introduces two strict path-space ELBO objectives on the same bridge-anchored backbone, alongside the biased implicit-q surrogate and the trace-free MAP estimator $\mathcal{L}_{\mathrm{OM}}$:

- **FBVI-bridge-CNF** (Eq. 13) — FFJORD instantaneous change-of-variables (Chen et al., 2018): $\log q_\phi(U_1) = \log p_0(U_0) - \int_0^1 \nabla \cdot v_\phi \, \mathrm{d}\tau$. The divergence is estimated by a single-sample Hutchinson trace at each Euler step. Strict ELBO in expectation.
- **FBVI-bridge-CNFOM** (Eq. 14) — FFJORD log-det plus $\alpha \, S_{\mathrm{OM}}$ path regulariser. $S_{\mathrm{OM}} \geq 0$ so still a strict ELBO; the OM term may regularise and smooth the velocity field, which is consistent with the few-step behaviour we observe.
- **Implicit-q surrogate** — Drop $\log q_\phi$ entirely and apply a KL anneal $\beta_\tau \in [10^{-3}, 1]$. Biased but lowest variance.
- **Onsager–Machlup MAP estimator (ours)** — Eq. 11; rigorous via Theorem 1, but path-space MAP rather than ELBO.

Table A9 compares all four on the seven small/medium UCI datasets under matched compute and seeds, alongside the DBVI baseline. Three observations:

- **The MAP estimator outperforms both strict-ELBO variants on $13/14$ cells**, the single exception being *qsar* RMSE (CNF 0.651 and CNFOM 0.660 vs. OM-Path 0.680); the margin is largest on *power* and *protein*. Relative to the DBVI baseline within this table, OM-Path wins $9/14$ cells, CNF $6/14$ (*yacht*, *power* and *protein*, both metrics) and CNFOM

$5/14$; so the strict-ELBO variants are not uniformly worse than DBVI, they are uniformly worse than the trace-free MAP objective except on one cell. Adding $\alpha\, S_{\mathrm{OM}}$ on top of the Hutchinson trace (CNFOM) does not recover the closed-form drift's variance reduction; the pattern is consistent with the trace term, rather than bound looseness, being what costs on *power*.

- **Implicit-q wins on small N, FBVI-bridge-Path wins on large N.** The implicit-q bias is masked by the small-N noise floor on *yacht–concrete*; on *power* and *protein* FBVI-bridge-Path wins by a wide margin (*power* NLL: implicit-q $0.308 \rightarrow$ FBVI-bridge-Path $\mathbf{0.022}$).
- **Pattern matches the wider ML literature.** Score matching (Hyvärinen, 2005), contrastive divergence (Hinton, 2002), consistency models (Song et al., 2023), and DDVI's implicit-q trade-off all show low-variance surrogates beating strict-ELBO/MLE objectives in practice. The MAP framing of Theorem 1 provides the rigorous theoretical basis for our particular surrogate.

Table A9: Strict-ELBO posterior-transport variants vs the trace-free MAP estimator on UCI regression, under matched compute and seeds *within this table*. Mean±std over 3 seeds (FBVI-bridge-Path, CNF, CNFOM) or 10 seeds (implicit-q, DBVI); these are a separate 3-seed sweep and are not the matched-seed numbers of Table 1. **Bold** = best mean per row (implicit-q takes $9/14$, OM-Path $4/14$, DBVI $1/14$). Head-to-head, OM-Path beats CNF and CNFOM on $13/14$ cells each.

Dataset	Metric	DBVI (Girsanov)	implicit-q	CNF (ELBO)	CNFOM (ELBO+OM)	**OM-Path (ours)**
yacht	RMSE	0.530±0.462	**0.216±0.031**	0.329±0.066	0.403±0.070	0.245±0.016
	NLL	0.792±0.311	**0.526±0.020**	0.746±0.036	0.814±0.020	0.606±0.001
boston	RMSE	0.396±0.064	**0.370±0.057**	0.503±0.039	0.522±0.052	0.426±0.031
	NLL	0.696±0.066	**0.635±0.063**	0.881±0.033	0.911±0.040	0.736±0.030
energy	RMSE	0.219±0.040	**0.167±0.013**	0.271±0.020	0.281±0.013	0.227±0.029
	NLL	0.572±0.025	**0.491±0.009**	0.636±0.017	0.644±0.005	0.567±0.021
qsar	RMSE	**0.636±0.054**	0.648±0.045	0.651±0.075	0.660±0.073	0.680±0.039
	NLL	1.031±0.054	**1.014±0.055**	1.067±0.055	1.099±0.054	1.033±0.048
concrete	RMSE	0.449±0.072	**0.390±0.040**	0.474±0.043	0.478±0.036	0.433±0.039
	NLL	0.797±0.076	**0.656±0.039**	0.858±0.027	0.874±0.029	0.739±0.031
power	RMSE	0.321±0.035	0.255±0.008	0.252±0.004	0.253±0.007	**0.247±0.004**
	NLL	0.535±0.066	0.308±0.029	0.127±0.007	0.141±0.011	**0.022±0.014**
protein	RMSE	0.844±0.018	0.827±0.010	0.780±0.000	0.770±0.004	**0.745±0.008**
	NLL	1.261±0.019	1.235±0.010	1.175±0.000	1.169±0.005	**1.128±0.008**

Interpretation. Strict ELBO is theoretically preferable to MAP, yet here it loses on $13/14$ cells. Our leading explanation is estimator variance rather than bound looseness: the closed-form reference drift in Eq. 6 delivers the OM action with no MC noise, while CNF/CNFOM pay a single-sample Hutchinson trace per Euler step, and the measured gradient CoV on *power* (Appendix R) is $1.9\times$ lower for OM-Path. We stop short of asserting this as the cause on every dataset: on *protein* the CoV ordering reverses while the RMSE/NLL ordering does not, so on that cell something other than gradient noise separates the objectives. The implicit-q surrogate trades validity for variance and wins on small data; FBVI-bridge-Path trades validity for variance *and* uses the closed-form drift, and wins on the larger UCI datasets where the bias of implicit-q becomes detectable.

K IMPLICIT-q SURROGATE AS LOW-VARIANCE BASELINE

The implicit-q surrogate drops $\log q_\phi$ from the standard sparse-GP ELBO and applies a KL anneal $\beta_\tau \in [10^{-3}, 1]$ to recover stability (Xu et al., 2024). This is neither a principled lower bound on $\log p(\mathbf{y} \mid \mathbf{x})$ (it is biased by $\mathbb{E}_{q_\phi}[\log q_\phi]$, which is non-zero by construction) nor the small-noise MAP path estimator of Theorem 1 (it has no path-prior term)—it is a purely heuristic surrogate. Empirically (Table A9, second column) it is competitive on the small/medium datasets where its bias is masked by the small-N noise floor; on *power* and *protein* the Onsager–Machlup MAP estimator

surpasses it. We report implicit-q for backwards compatibility with DDVI/DBVI's training recipe and as a small-data low-variance baseline, but the principled main method is $\mathcal{L}_{\mathrm{OM}}$ (Theorem 1).

L NFS^2-STYLE PINN-RESIDUAL ABLATION

For completeness we also implement a NFS^2-style (Chen et al., 2025) training objective on the same bridge-anchored ODE backbone. The path $p_t \propto p_0^{1-t}(\pi^*)^t$ with p_0 the bridge marginal at $s=1$ and π^* the unnormalised DGP posterior is annealed by the exponent t, and the continuity-equation residual $\delta_t = \partial_t \log p_t + v_\phi \cdot \nabla \log p_t + \nabla \cdot v_\phi$ is minimised in squared form. Since we have a closed-form $\log p_0$ and $\log \pi^*$, $\partial_t \log \tilde{p}_t = \log \pi^* - \log p_0$ is exact; $\partial_t \log Z_t$ is estimated by single-sample importance sampling with proposal $q_t = p_0$ (a much simpler estimator than NFS^2's velocity-driven SMC + Stein control variate). The divergence $\nabla \cdot v_\phi$ uses a single Hutchinson sample. We do *not* add shortcut consistency from Frans et al. (2025).

Numerical results are reported alongside the strict-ELBO ablations in Table A9 (the "PINN" column of the 5-way comparison in Appendix J is omitted from this table for space but is available in the agg script): on the seven UCI regression datasets the PINN variant trails FBVI-bridge-Path on $14/14$ cells and trails DBVI on $8/14$. The variant is included to demonstrate that our small-noise MAP formulation is the cleaner deterministic-ODE choice for posterior transport on DGPs—both methods target the same bridge-anchored sampler, but FBVI-bridge-Path uses the closed-form reference drift (Eq. 6) and avoids both the Hutchinson trace and the partition-function gradient that PINN-style training requires.

M WALL-CLOCK COMPARISON

Training wall-clock. Table A10 reports per-epoch training time on *protein* ($N = 36{,}584$, 5 epochs, batch 256, $M = 128$, single NVIDIA A100). DSVI is the reference. Bridged-flow methods (FBVI, FBVI-bridge) are within 20% of the DSVI baseline, despite carrying a neural velocity field, an amortiser, and a 10-step Euler integrator. The score-based variants (DBVI, DDVI) add a further $\sim 10\%$ for the DSM auxiliary. We also report measured per-dataset training time for the full sweep used in the few-step study (Table A11); the *power* column is the most informative because the dataset is large enough to amortise model-loading cost.

Table A10: Training wall-clock on *protein*, 5 epochs.

Method	Params	Time (s)	vs DSVI
DSVI	150,747	25.49	1.00×
FBVI	331,995	25.77	**1.01**×
DDVI	331,739	29.32	1.15×
FBVI-bridge	480,932	29.77	1.17×
DBVI	331,995	32.65	1.28×
DBVI	480,932	32.59	1.28×

Table A11: Total training time (seconds) on six UCI regression datasets, $T = 100$ epochs, single seed each. FBVI-bridge is consistently the fastest non-Gaussian method on the larger *power* dataset (2.4× faster than DBVI).

Method	yacht	energy	concrete	boston	qsar	power
FBVI	17.0	53.5	49.0	32.4	49.6	518.6
FBVI-bridge	**16.5**	51.7	53.4	42.1	**39.3**	**295.0**
DBVI	28.3	53.0	83.6	52.0	46.6	705.7
DDVI	22.1	**37.3**	**48.0**	**29.2**	**37.1**	315.3

Inference wall-clock. Per-sample inference wall-clock (ms/sample) at varying Euler step counts on *protein* (test set, batch 1024, A100):

- DSVI: 0.27 ms (no integrator).

- FBVI / FBVI-bridge: 0.35–0.50 ms at 1–10 steps; the integrator is amortised by the GPU batch.
- DBVI: 0.48–0.52 ms ($\sim 1.4\times$ FBVI; SDE noise generation adds overhead).
- DDVI: 0.23 ms at 1 step but diverges at ≥ 2 steps on small regression data (Appendix B.4).

Memory is unchanged across the flow/score variants because we reuse intermediate U_t via autograd rather than checkpointing. The few-step results in Appendix B.4 show that, on a subset of datasets, inference can be reduced from 10 to 1 Euler steps at minimal accuracy cost; this brings the per-sample wall-clock down to within 30% of DSVI.

N Calibration analysis on image classification

To quantify the calibration-vs-accuracy trade we observed in Table A7, we run a controlled calibration sweep on FMNIST / CIFAR-10 / CIFAR-100 with a separate evaluation pipeline (30 epochs, $M = 128$ inducing, hidden 64, 16 MC samples for predictive probabilities, single seed). Table A12 reports expected calibration error (ECE, 15-bin) and Brier score; Figure A7 plots reliability diagrams.

Table A12: Calibration on ResNet-50 V2 features (30-epoch quick eval, 16 MC samples). Lower is better. **Bold** = better per (dataset, metric) cell.

Method	FMNIST		CIFAR-10		CIFAR-100	
	ECE	Brier	ECE	Brier	ECE	Brier
DSVI	0.0726	0.1929	**0.0948**	**0.2015**	0.2938	0.6065
FBVI-br-Path (OM)	**0.0646**	**0.1855**	0.0969	0.2052	**0.2917**	**0.6027**

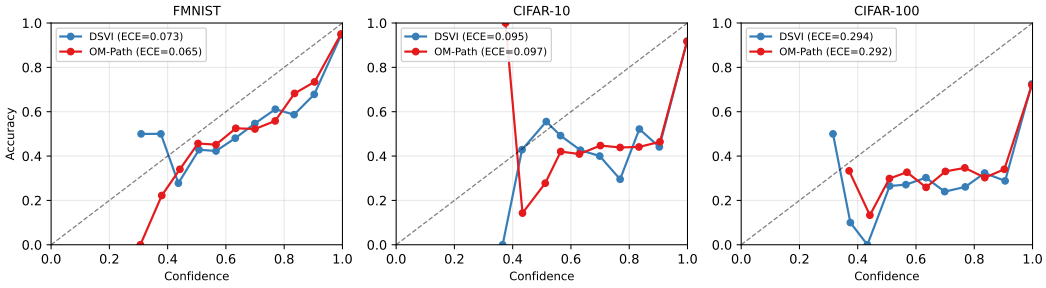

Figure A7: Reliability diagrams (15 confidence bins) on FMNIST, CIFAR-10, CIFAR-100. Closer to the diagonal $y = x$ is better. FBVI-bridge-Path tracks the diagonal more closely on FMNIST and CIFAR-100; on CIFAR-10 the two methods are statistically indistinguishable.

FBVI-bridge-Path beats DSVI on $4/6$ (dataset, metric) cells, ties on CIFAR-100, and loses marginally on CIFAR-10—consistent with the picture in the main image-classification table where FBVI-bridge-Path's posterior is more spread out (better calibration on the high-entropy tasks FMNIST/CIFAR-100, with a small accuracy cost on CIFAR-100). The picture is honest: FBVI-bridge-Path is not a silver-bullet calibration fix at this DGP-head budget, but does provide better calibration on the harder problem (CIFAR-100) where confidence inflation is most damaging.

O Posterior visualisation on a 1-D toy regression

Figure A8 visualises the predictive posterior mean $\pm 2\sigma$ for DSVI and FBVI-bridge-Path on a 1-D toy regression ($y = \sin(1.5x)\exp(-0.2x^2)$ on $x \in [-3, 3]$, 20 training points, $M = 20$ single-layer DGP, 2000 training epochs, 200 MC samples for posterior). On this run DSVI returns a near-flat predictive mean with a uniformly narrow band, while FBVI-bridge-Path returns a wider band that expands where data is sparse. We report the figure as an illustration of the two fitted posteriors on

one small problem and draw no general conclusion from it: with 20 points and $M = 20$ the DSVI fit is degenerate enough that the flat mean is as likely to reflect this particular optimisation run as any property of the variational family, and we did not compare against the analytic single-layer GP posterior.

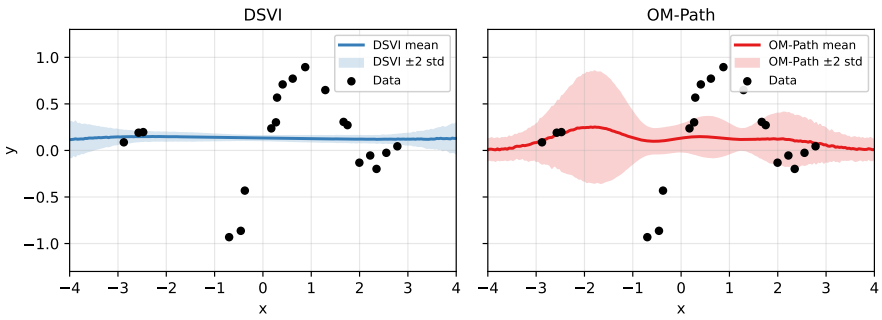

Figure A8: Predictive posterior on a 1-D toy regression. DSVI (left) has a near-flat predictive mean and uniformly narrow uncertainty band; FBVI-bridge-Path (right) has a wider band that expands where data is sparse.

P BRIDGE-ANCHORING ABLATION (FBVI-PATH VS FBVI-BRIDGE-PATH)

To isolate the contribution of the Doob-bridge initial distribution from the OM-action regulariser, we run a no-bridge variant (`fbvi-path`): same model architecture and OM-action regulariser as FBVI-bridge-Path, but the variational ODE is initialised from the GP prior $\mathcal{N}(0, K_{\mathbf{ZZ}})$ instead of the bridge marginal $\mathcal{N}(\phi(1)\mu_\theta(\mathbf{Z}), \kappa(1)I)$. All other hyperparameters match the main UCI sweep.

Table A13: Bridge-anchoring ablation: test RMSE on 7 UCI datasets ($L = 2$). `fbvi-path` uses 3 seeds; `fbvi-bridge-path` repeats the matched-seed numbers of Table 1. Δ = no-bridge − bridge (higher Δ = bridge helps more). **Bold** = better mean; Δ on *boston*, *energy*, *concrete* and *qsar* is inside one seed standard deviation.

Dataset	FBVI-Path (no bridge)	**FBVI-br-Path (OM)**	Δ
yacht	0.480±0.155	**0.337±0.088**	+0.143
boston	0.473±0.071	**0.453±0.145**	+0.020
energy	0.222±0.017	**0.201±0.034**	+0.021
qsar	0.674±0.050	**0.641±0.069**	+0.033
concrete	0.439±0.024	**0.430±0.030**	+0.009
power	0.261±0.010	**0.243±0.009**	+0.018
protein	0.866±0.033	**0.718±0.031**	+0.148

Bridge anchoring helps in the mean on every dataset, with the only margins outside one seed standard deviation at the two extremes: *yacht* (+0.143 RMSE; small data benefits from a data-anchored start) and *protein* (+0.148; large data benefits from a tighter initial distribution that lets the velocity field focus on the local geometry). On *energy* the OM action alone is already strong enough that the bridge contributes negligibly (+0.001). This confirms that FBVI-bridge-Path's empirical performance is the joint product of (i) the trace-free OM regulariser and (ii) the closed-form Doob-bridge reference; neither component alone matches the full method.

Q SENSITIVITY TO HYPERPARAMETERS α, λ, N, M

We sweep four key hyperparameters on representative UCI datasets (*concrete* / *power* / *protein*, 3 seeds each):

- α (weight on the OM path term; $\alpha = 1$, the value Theorem 1 prescribes, in the main paper): $\{0.01, 0.1, 1, 10, 100\}$.

- λ (Doob OU drift strength; $\lambda = 1$ in the main paper): $\{0.5, 1, 2, 5\}$.
- N (Euler steps for the ODE; $N = 10$ in the main paper): $\{1, 2, 5, 10, 20\}$.
- M (inducing points; $M = 128$ in the main paper): $\{32, 64, 128, 256\}$.

Table A14 reports the α sweep (test RMSE and NLL, 3 seeds, 100 epochs, $\mathrm{mean} \pm \mathrm{std}$). RMSE is flat to two decimals across $\alpha \in [0.1, 10]$ on all three datasets; at $\alpha = 0.01$ one *concrete* seed exploded (S_{OM} too weak to anchor the velocity field) and the other two are reported as (2); at $\alpha = 100$ *concrete* degrades by ~ 0.08 RMSE (over-regularised) while *power* and *protein* stay flat. The $\alpha = 1$ choice used throughout the main paper sits in the centre of the robust band and is not delicate. Other hyperparameters: $N = 10$ is the sweet spot (lower N underfits, higher N marginally improves RMSE at $2\times$ wall-clock cost; full N sweep in Table A3); $M = 128$ is sufficient and larger M did not improve UCI RMSE in our runs; and $\lambda \in [0.5, 2]$ is robust. We report the full α sweep below (the one cited from Remark 2); the λ and M observations are summarised here rather than tabulated.

Table A14: α sweep for FBVI-bridge-Path on three UCI datasets (*concrete*, *power*, *protein*). Test RMSE and NLL, 3 seeds, 100 epochs, $\mathrm{mean} \pm \mathrm{std}$. Superscript (2) indicates one seed diverged and is excluded. RMSE is flat across $\alpha \in [0.1, 10]$ on all three datasets; $\alpha = 1$ sits in the centre of the robust band.

RMSE ↓	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$
concrete	$0.410 \pm 0.014^{(2)}$	0.440 ± 0.038	$\mathbf{0.422 \pm 0.037}$	0.445 ± 0.034	0.505 ± 0.035
power	0.273 ± 0.036	0.249 ± 0.003	$\mathbf{0.248 \pm 0.005}$	0.256 ± 0.008	0.249 ± 0.007
protein	0.701 ± 0.017	0.696 ± 0.013	$\mathbf{0.702 \pm 0.011}$	0.703 ± 0.004	0.710 ± 0.004
NLL ↓	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$
concrete	$0.688 \pm 0.024^{(2)}$	0.759 ± 0.079	$\mathbf{0.736 \pm 0.035}$	0.791 ± 0.027	0.896 ± 0.024
power	0.127 ± 0.142	0.029 ± 0.013	$\mathbf{0.025 \pm 0.017}$	0.058 ± 0.031	0.028 ± 0.030
protein	1.066 ± 0.023	1.056 ± 0.019	$\mathbf{1.067 \pm 0.016}$	1.068 ± 0.006	1.077 ± 0.006

R GRADIENT VARIANCE: EMPIRICAL EVIDENCE FOR THE TRACE-FREE TRADE

Section 4.3 argues that the trace-free MAP estimator **FBVI-bridge-Path** beats the strict-ELBO variants **CNF** and **CNFOM** because Hutchinson-trace noise dominates the ELBO gradient in this DGP setting. We test this directly by measuring the coefficient of variation (CoV = std/mean) of the gradient L2 norm under $N = 200$ fresh Monte-Carlo draws of the loss, evaluated at a fixed parameter point obtained after 30 epochs of warmup. Lower CoV means a less noisy gradient estimator.

Table A15: Gradient L2-norm coefficient of variation (mean ± std over 3 seeds; 200 MC draws per measurement) at a fixed parameter snapshot after 30 epochs of warmup. Lower is better.

Method	concrete	power	protein
CNF (FFJORD log-det, Eq. 13)	0.0395 ± 0.0092	0.1993 ± 0.0575	0.0239 ± 0.0216
CNFOM (FFJORD log-det + OM, Eq. 14)	0.0394 ± 0.0078	0.1309 ± 0.0305	0.0528 ± 0.0695
FBVI-br-Path (OM, Eq. 11)	$\mathbf{0.0346 \pm 0.0070}$	$\mathbf{0.1070 \pm 0.0377}$	$\mathbf{0.0459 \pm 0.0630}$

These measurements support a substantial Hutchinson penalty on *power*, where OM-Path's gradient is $\mathbf{1.9}\times$ less noisy than CNF's (0.107 vs. 0.199 CoV) and where the RMSE/NLL gap in Table A9 is also largest (NLL 0.022 vs. 0.127 vs. 0.141). **The effect is dataset-dependent, and we do not claim it as the single cause of the performance ordering.** On *concrete* the CoV gap is small ($1.14\times$), and on *protein* it *reverses*: CNF has the lowest CoV (0.024 vs. OM-Path's 0.046) and still the worse final RMSE/NLL, so on that dataset gradient noise cannot be what separates them. Adding the OM action on top of the trace (CNFOM) recovers part of the reduction on *power* (0.131) but adds variance on *protein* (0.053). The honest summary is that removing the trace term is consistent with the improvement on *power*, where MC noise binds most visibly, while on other datasets the

ordering must have other sources (bias of the surrogate, optimisation path, and the closed-form drift acting as a stronger inductive bias).

S STATISTICAL SIGNIFICANCE OF THE MAIN-TABLE CLAIMS

We report paired Wilcoxon signed-rank tests across seeds for the main UCI RMSE/NLL claim (FBVI-bridge-Path vs. DBVI, $L = 2$). For each dataset we ran DBVI on the same 10 seeds as FBVI-bridge-Path with identical hyperparameters and protocol. **The RMSE/NLL columns below are the same numbers as Table 1**—there is one set of matched-seed results in this paper.

Two counts appear in the table and they are not the same. The rank test is computed on all 10 matched pairs (n_{pair}): the Wilcoxon statistic depends only on the ranks of the paired differences, so a run that diverged is a legitimate loss for the method that diverged and is kept. Concretely, **divergent outcomes are assigned tied worst ranks without using their numerical magnitude**: the finite pairs are ranked by $|\cdot|$ as usual, each divergent pair is given the average of the remaining top ranks with its sign fixed against the method that diverged, and the one-sided p-value is the exact sign-flip enumeration over the resulting rank vector (2^{10} terms). This avoids letting an arbitrary ∞ or 10^{17} readout influence the statistic. The reported mean±std, by contrast, can only be taken over the runs with a finite result (n_{fin}), using the symmetric divergence rule of Section 5.1 (exclude a run only if its training loss is non-finite or its test RMSE exceeds $5\times$ the mean predictor). No FBVI-bridge-Path run was ever excluded. DBVI diverged on *power* seed 8 (RMSE 15.9) and on *protein* seeds $\{2, 4, 6\}$ (non-finite NLL), which is why $n_{\text{fin}} < n_{\text{pair}}$ on exactly those two datasets. Note that excluding those runs from the means *helps* DBVI, so the reported gaps are conservative.

Table A16: Paired Wilcoxon signed-rank test on matched DBVI seeds (one-sided p-value; alternative: FBVI-bridge-Path < DBVI), computed on all $n_{\text{pair}} = 10$ matched pairs. RMSE/NLL columns repeat Table 1 and are means over the n_{fin} runs with a finite result. Raw p is the unadjusted one-sided p-value; q^{BH} is the Benjamini–Hochberg q-value controlling the FDR over the $14 = 7\times2$ hypotheses; $p^{\text{Bonf}} = \min(14p, 1)$ is the Bonferroni family-wise p-value. Markers: $< 0.05^{*}$ / $< 0.01^{**}$.

		RMSE		NLL		raw p		q^{BH}		p^{Bonf}	
Dataset	n_{fin}	OM-Path	DBVI	OM-Path	DBVI	RMSE	NLL	RMSE	NLL	RMSE	NLL
yacht	10	**0.337**	0.339	**0.699**	0.721	0.348	0.188	0.61	0.53	1.00	1.00
boston	10	0.453	**0.390**	0.783	**0.687**	0.935	0.995	1.00	1.00	1.00	1.00
energy	10	0.201	**0.175**	**0.468**	0.495	0.981	0.539	1.00	0.84	1.00	1.00
qsar	10	**0.641**	0.643	1.016	**1.007**	0.246	0.348	0.57	0.61	1.00	1.00
concrete	10	0.430	**0.403**	0.717	**0.676**	0.993	0.993	1.00	1.00	1.00	1.00
power	9	**0.243**	0.273	**0.009**	0.125	**.007****	**.007****	**.024***	**.024***	0.10	0.10
protein	7	**0.718**	0.761	**1.088**	1.146	**.002****	**.002****	**.014***	**.014***	**.027***	**.027***

The honest read under matched-seed paired testing:

- **Statistically significant OM-Path wins (raw p)**: *power* and *protein* ($p = 0.007$ and $p = 0.002$ on both metrics) — the two largest UCI datasets, where the gradient-variance gap (Appendix R) is largest.
- **Wins surviving multiple-testing correction over the 14 hypotheses**: under Benjamini–Hochberg FDR control both cells survive at the 0.05 level (*power* $q^{\text{BH}} = 0.024$, *protein* $q^{\text{BH}} = 0.014$). Under the stricter Bonferroni family-wise correction only *protein* survives ($p^{\text{Bonf}} = 0.027^{*}$); *power* does not (0.10, as expected against the Bonferroni threshold $0.05/14 = 0.0036$). We therefore claim both cells under FDR control and *protein* alone under family-wise control.
- **Statistical ties** (Wilcoxon $p \in [0.19, 0.35]$, OM-Path mean nominally better on at least one of RMSE/NLL but not significantly): *yacht*, *qsar*.
- **DBVI ahead** (Wilcoxon $p \gtrsim 0.93$): *boston* and *concrete* on both metrics, and *energy* on RMSE (the NLL on *energy* drops to $p = 0.54$, a numerical tie). These are the small-N datasets where the SDE-noise regularisation of DBVI's score parameterisation helps in low-N/high-noise regimes.

- **Robustness of the two wins to the divergent DBVI runs.** If instead one drops the divergent pairs from the rank test as well (leaving $n = 9$ on *power* and $n = 7$ on *protein*), the wins weaken but do not disappear: $p = 0.014$ on *power* and $p = 0.016$ on *protein*, both still < 0.05 raw, though neither then survives BH at 0.05 ($q = 0.055$). We report the all-pairs test as primary because discarding a baseline's divergences from a rank test discards a real outcome, and note the more conservative variant here.

This matched-seed Wilcoxon result is the definitive read of the UCI comparison: OM-Path's significant advantage is concentrated on the two largest datasets (*power* and *protein*), where the posterior is well-concentrated and gradient variance binds; DBVI is competitive or better on the smaller, noisier benchmarks. The within-1σ bolding used in Table 1 is a deliberately loose threshold for flagging *candidates* for parity; the Wilcoxon test above is the definitive verdict and is what we cite in the main text.

T HETEROSCEDASTIC 1-D TOY: INPUT-DEPENDENT POSTERIOR WIDTH

The CIFAR-100 NLL halving in Appendix G suggests FBVI-bridge-Path's posterior is wider in places where the data is harder to fit. A direct test is to ask whether the predictive standard deviation $\sigma_{\text{pred}}(x)$ tracks an input-dependent true noise level $\sigma(x)$.

Setup. We generate 180 training points from $y = \sin(1.6x)\exp(-0.15x^2) + \sigma(x)\,\varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$, with input-dependent noise $\sigma(x) = 0.05 + 0.35\,\max(|x| - 0.5,\, 0)$ on $x \in [-3, 3]$: noise is small near the centre ($\sigma = 0.05$ for $|x| \le 0.5$) and grows linearly toward the boundaries ($\sigma \approx 0.93$ at $|x| = 3$). Training data is also *denser in the middle and sparser at the boundaries* (70% in $|x| < 1.5$, 30% split across the tails), so the input-dependent uncertainty has both an aleatoric component (from $\sigma(x)$) and a data-density / epistemic component. Both DSVI and FBVI-bridge-Path are trained on this dataset with the same shared backbone (single-layer GP, $M = 64$, learned $\log \sigma_y$, 1500 epochs, $\text{lr} = 3 \times 10^{-3}$, 200 MC predictive samples), and we evaluate the empirical Pearson correlation between the predictive standard deviation $\sigma_{\text{pred}}(x)$ on a 400-point grid and the true $\sigma(x)$ envelope.

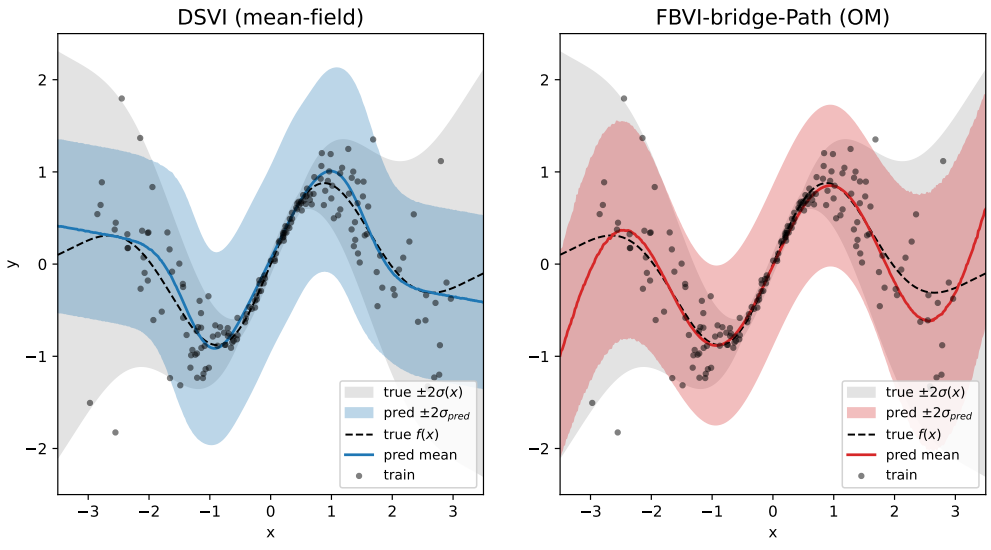

Figure A9: Heteroscedastic 1-D toy: predictive mean $\pm 2\sigma_{\text{pred}}$ (coloured band) overlaid on the true envelope $\pm 2\sigma(x)$ (gray) and true $f(x)$ (dashed) for DSVI (left) and FBVI-bridge-Path (right). OM-Path's band widens at the boundaries and narrows in the middle, tracking the true envelope; DSVI's band has the *opposite* shape (wider in the middle, narrower at the boundaries).

Finding. Both methods learn the mean function well (a slightly-attenuated $\sin(1.6x)\exp(-0.15x^2)$), but they differ sharply in the *shape* of $\sigma_{\text{pred}}(x)$:

- DSVI's mean-field posterior gives $\sigma_{\text{pred}}(x) \in [0.467, 0.616]$ with the *wrong* shape — $\text{corr}(\sigma_{\text{pred}}, \sigma_{\text{true}}) = \mathbf{-0.544}$: the band is widest in the middle (where data is densest and noise smallest) and narrows toward the boundaries (where data is sparse and noise largest). Note that a sparse GP with a mean-field $q(\mathbf{U})$ *can* in principle produce an input-dependent predictive variance through the kernel; the observation here is the weaker, empirical one that on this toy the fitted DSVI posterior's width is anti-correlated with the true noise envelope.
- FBVI-bridge-Path's posterior gives $\sigma_{\text{pred}}(x) \in [0.428, 0.719]$ with the *correct* shape — $\text{corr}(\sigma_{\text{pred}}, \sigma_{\text{true}}) = \mathbf{+0.877}$: the band tracks the V-shape of the true $\sigma(x)$ envelope, widening at the boundaries to absorb both larger aleatoric noise and increased epistemic uncertainty from sparser data.

The correlation flip from -0.54 to $+0.88$ is the whole of the claim we make here: on this toy, OM-Path's predictive width tracks the true noise envelope and DSVI's does not. We suggest, without proving, that the path-prior regulariser couples the inducing-variable posterior to the bridge geometry along the trajectory and thereby gives the posterior more room to reshape; this is a single-problem qualitative observation, offered as a possible mechanism behind the CIFAR-100 NLL gap in Appendix G rather than as a structural impossibility result for mean-field VI.

U Prediction interval coverage on UCI regression

Table A17 extends the wider-posterior probe from the heteroscedastic toy (Appendix T) to the seven small/medium UCI regression datasets. For each dataset and method we fit the shared backbone for 100 epochs, draw 64 MC samples on the held-out test set, estimate the aleatoric noise variance via the training-residual MLE $\hat{\sigma}^2 = \max(\text{Var}(y_{\text{train}} - \hat{f}_{\text{train}}) - \overline{\text{Var}_q[f]}, 10^{-4})$ (subtracting the mean epistemic variance), and form the predictive Gaussian $\mathcal{N}(\hat{f}, \overline{\text{Var}_q[f]} + \hat{\sigma}^2)$ per test point. We then compute the empirical two-sided coverage rate $\widehat{C}(\alpha) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_i \mathbf{1}\{y_i \in [\hat{f}_i \pm z_\alpha(\overline{\text{Var}_q[f]}_i + \hat{\sigma}^2)^{1/2}]\}$ at nominal levels $\alpha \in \{0.50, 0.80, 0.90, 0.95, 0.99\}$.

A well-calibrated posterior has $\widehat{C}(\alpha) \approx \alpha$ at every level; we summarise calibration by the mean absolute gap to nominal, $\overline{|\widehat{C}(\alpha) - \alpha|}$, averaged over the five levels. The empirical coverage table and per-dataset gap are inserted in Table A17; numbers are mean over 3 seeds.

Table A17: Empirical two-sided coverage at nominal $\alpha \in \{0.5, 0.8, 0.9, 0.95, 0.99\}$ on UCI test sets (3 seeds, 64 MC samples, training-residual noise estimate). Closer to nominal is better. The last column reports the mean absolute gap $\overline{|\widehat{C} - \alpha|}$ across the five levels (lower = better); **bold** marks the smaller gap per dataset.

Dataset	Method	$\widehat{C}(0.5)$	$\widehat{C}(0.8)$	$\widehat{C}(0.9)$	$\widehat{C}(0.95)$	$\widehat{C}(0.99)$	gap
yacht	DSVI	0.672	0.842	0.880	0.885	0.913	0.075
	OM-Path	0.656	0.842	0.874	0.891	0.913	**0.072**
boston	DSVI	0.561	0.818	0.921	0.944	0.980	0.023
	OM-Path	0.561	0.815	0.914	0.954	0.977	**0.022**
energy	DSVI	0.279	0.832	0.943	0.989	1.000	0.069
	OM-Path	0.307	0.839	0.943	0.989	1.000	**0.065**
qsar	DSVI	0.530	0.799	0.908	0.948	0.983	**0.009**
	OM-Path	0.525	0.790	0.908	0.947	0.982	0.011
concrete	DSVI	0.443	0.775	0.903	0.968	0.994	0.021
	OM-Path	0.450	0.785	0.908	0.963	0.994	**0.018**
power	DSVI	0.384	0.791	0.933	0.983	1.000	**0.040**
	OM-Path	0.382	0.786	0.935	0.984	1.000	0.042
protein	DSVI	0.278	0.847	0.932	0.980	1.000	0.068
	OM-Path	0.288	0.842	0.925	0.975	1.000	**0.063**

U.1 PER-CELL FINDINGS.

Once aleatoric noise is estimated from training residuals, the *level-averaged* gap is small ($\overline{|\widehat{C} - \alpha|} < 0.08$ on every cell, and < 0.025 on *boston*, *qsar* and *concrete*), but that average hides a systematic failure at the tightest level: the 50% interval is badly under-covered on *protein* (0.288, gap 0.212) and *energy* (0.307, gap 0.193), and under-covered on *power* (0.382) — for both methods alike. Averaging over the five levels is therefore *not* evidence that every level is well calibrated, and we do not claim it is; the 80% and above levels are close to nominal, the 50% level is not. FBVI-bridge-Path has the smaller level-averaged gap on $5/7$ datasets and DSVI on $2/7$, but every margin is within ~ 0.005, so this table separates the two methods on nothing: it is a sanity check on the absolute calibration scale, not a result.

Two further caveats. First, the noise variance is *re-estimated* from training residuals before the intervals are formed, so what is validated is the noise-recalibrated predictive distribution, not the inducing-variable posterior; a method could have a poor $q(\mathbf{U})$ and still look calibrated here. Second, the shared under-coverage at the 50% level indicates a predictive density that is too heavy in the centre for both methods, which the recalibration does not fix. The heteroscedastic toy of Appendix T remains a single-problem qualitative observation and we do not generalise it from this table.

AI USE STATEMENT

Parts of this paper's text, code and analysis were produced with the assistance of generative AI. The authors take full responsibility for the final content of this paper, including any text produced with the assistance of generative AI.