STRUCTURED NEURAL SDES FOR FUNCTIONAL CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural Stochastic Differential Equations (Neural SDEs) provide flexible continuous-time generative models, but generic neural drift and diffusion networks are costly to simulate on long horizons and can give unstable gradients when the training signal is a path functional rather than a pointwise observation. We introduce *SLiSDE*, a family of Neural SDE models built from structured linear stochastic layers. Parallel-in-time simulation is obtained at the layer level, while expressivity is recovered by *gated in-flow* stacking: previous-layer paths modulate the next layer's latent flow through learned gates. For functional calibration tasks in which rare paths dominate the loss, we add an optional Girsanov tilt that acts as a learned importance sampler with an exact likelihood-ratio correction. We prove well-posedness, a discretisation error bound, validity of the change of measure, and a universality result: the terminal laws of the gated stack are dense in the space of square-integrable laws. Experiments on functional calibration benchmarks show that the structured model outperforms fully neural SDE baselines while retaining parallel-time simulation and stable importance weights.

1 INTRODUCTION

Many quantities of practical interest are expectations of functionals of a stochastic path whose trajectories are never observed. In derivatives markets the available data are option quotes – expectations of terminal payoffs – and exceedance probabilities, and calibrating a model of the underlying dynamics to such quotes is the entry point of pricing, hedging and risk management (Buehler et al., 2019); the same situation, dynamics observed only through a finite family of statistics, arises in the physical sciences. We refer to this task as *functional calibration*: fitting a generative model of continuous-time stochastic dynamics to prescribed expectations of path functionals. Training such a model requires simulating it afresh at every gradient step, so the computational cost of simulation is the binding constraint, and it is the constraint that shapes the model proposed here.

Neural differential equations provide a flexible framework for learning continuous-time dynamics by parameterizing their vector fields with neural networks. Neural ODEs (Chen et al., 2018) model deterministic latent evolution, neural CDEs (Kidger et al., 2020; Walker et al., 2025) condition the dynamics on an observed driving path, and neural SDEs (Tzen and Raginsky, 2019) provide the stochastic analogue, modeling distributions on continuous-time paths via stochastic differential equations, which may be interpreted in either the Itô or Stratonovich sense.

In this work, we focus on the Itô setting, where the latent state evolves as

$$\mathrm{d}Z_t = \mu_\theta(t, Z_t)\ \mathrm{d}t + \sigma_\theta(t, Z_t)\ \mathrm{d}W_t, \qquad Z_0 \sim p_0, \tag{1}$$

where W_t is a Brownian motion and μ_θ, σ_θ unconstrained neural networks. This is a rich model class (Tzen and Raginsky, 2019; Liu et al., 2019; Li et al., 2020; Kidger et al., 2021), and in practice it is often used as a *Markovian lift*: the latent state is lifted to a higher-dimensional space in which a non-Markovian observed process can be represented as a Markov diffusion. The viewpoint is powerful, but it makes simulation expensive — every time step requires fresh nonlinear evaluations of the lifted vector field, and the recurrence is inherently sequential unless additional structure is imposed.

The central design principle of this paper is to move the non-linear part of the model from the non-linear drift and diffusion evaluated at every time step to non-linear coupling between structured linear

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.

SDE layers. After discretisation, each layer is an affine recurrence whose composition is associative, so it can be evaluated by the parallel associative scan of Section 2.1, which cuts the sequential simulation cost over a horizon of T time steps from $O(T)$ to $O(\log T)$. To recover expressivity, we stack such layers through *gated in-flow* coupling: the previous-layer path modulates the next layer's transition and offset through learned gates. The current layer remains affine in its own state, so the scan structure is preserved, while the gates make the stack a genuinely nonlinear model, universal at the level of terminal laws (Theorem 6).

For high-variance path-dependent objectives, we add an optional last-layer-only Girsanov tilt. A learned adapted controller changes only the Brownian increments driving the final layer and returns the corresponding Radon–Nikodym log-weight in closed form.

RELATED WORK.

Within the neural-SDE literature two training paradigms coexist. Generative training matches the law of the model to an empirical distribution of paths through a statistical divergence – the KL divergence of variational inference (Tzen and Raginsky, 2019; Li et al., 2020), the Wasserstein-1 distance of the GAN formulation (Kidger et al., 2021), or signature-kernel scores (Issa et al., 2023) – and requires sample paths. Calibration training, our setting, instead matches prescribed expectations of functionals, typically derivative prices, when no trajectory is observed (Gierjatowicz et al., 2020; Cuchiero et al., 2020; Cohen et al., 2021); before neural SDEs this task was handled by parametric families that hard-wire the dynamics, such as stochastic-volatility and local-volatility models in finance or mechanistic diffusion models in the physical sciences.

Structured state-space models have shown that carefully parameterised linear recurrences can achieve strong sequence-modelling performance while remaining hardware efficient (Gu et al., 2022; Smith et al., 2023; Hasani et al., 2022; Gu and Dao, 2023). Recent work also connects this line of models to controlled differential equations, viewing sequence transformations as dynamics driven by an input path (Muca Cirone et al., 2024). SLiCE (Walker et al., 2025) develops this perspective for structured linear controlled differential equations, showing how scan-compatible linear dynamics can be embedded in a CDE-style framework.

The Girsanov component is connected to importance sampling for diffusion processes and to the stochastic-control view of variance reduction (Hartmann et al., 2017; Zhang and Chen, 2022; Hartmann and Richter, 2024). In those approaches, one learns or designs a drift change that makes rare paths less rare and then corrects the resulting bias with a likelihood ratio. Our tilt is narrower: it acts only on the Brownian motion of the last SLiSDE layer. This keeps the prefix computation unchanged, and makes the proposal lightweight enough to train jointly with the backbone.

CONTRIBUTIONS.

- **Structured linear neural SDE for functional calibration.** We introduce a structured linear Neural SDE backbone whose discretized layers are affine recurrences. This gives $O(\log T)$ parallel depth in sequence length by the parallel associative scan, compared with the $O(T)$ complexity of standard neural SDE solvers.
- **Gated in-flow stacking.** We propose a gated in-flow stacking rule in which previous-layer paths modulate the next layer's affine transition and offset. Nonlinearity enters between layers only, so each layer stays affine in its own state and scan-compatible.
- **Girsanov overlay.** We add an optional last-layer-only Girsanov tilt with a closed-form likelihood-ratio correction: a learned adapted controller acts as a self-normalised importance sampler that improves performance on rare-event functional calibration with little added compute.
- **Theory.** We prove that the terminal laws generated by the gated stack are dense in $\mathcal{P}_2(\mathbb{R}^p)$ for the 2-Wasserstein distance (Theorem 6), so the structural restrictions cost nothing at the level of terminal laws.
- **Experiments.** We evaluate the model on nonlinear path-functional calibration tasks and real option-surface benchmarks, comparing against neural SDE and structured sequence-model baselines. SLiSDE attains lower calibration loss than the baselines and is faster per training step in every regime tested; a controlled study shows the tilt as a learned importance sampler for rare-event functionals.

2 MATHEMATICAL BACKGROUND

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in[0,T]}, \mathbb{P})$ be a filtered probability space supporting an m-dimensional Brownian motion $W_t = (W_t^1, \dots, W_t^m)$ on $[0, T]$. We consider a time grid $0 = t_0 < t_1 < \cdots < t_N = T$, with Brownian increments $\Delta W_k = W_{t_{k+1}} - W_{t_k} \sim \mathcal{N}(0, \Delta t_k I_m)$. Throughout, complexities are stated in terms of the horizon T, which for a fixed step size is proportional to the number of steps N. The Girsanov change of measure on which the tilt of Section 4 rests is recalled in Appendix A.1.

2.1 LINEAR SDES AND AFFINE FLOWS

A general affine linear SDE takes the form

$$\mathrm{d}Z_t = (AZ_t + b)\ \mathrm{d}t + \sum_{j=1}^{m} C^j Z_t\ \mathrm{d}W_t^j + D\ \mathrm{d}W_t, \quad Z_0 = z_0 \tag{2}$$

with $z_0 \in L^2(\Omega, \mathcal{F}_0, \mathbb{P}; \mathbb{R}^d)$, $A, C^j \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, and $D \in \mathbb{R}^{d\times m}$. On each grid interval $[t_k, t_{k+1}]$ its discretisation, by the matrix-exponential transition or by the Euler–Maruyama transition (Appendix C.1; the latter is used throughout our experiments), is an affine recurrence

$$Z_{k+1} = F_k Z_k + g_k, \tag{3}$$

where the transition matrix $F_k \in \mathbb{R}^{d\times d}$ and the offset $g_k \in \mathbb{R}^d$ are built from the coefficients and the Brownian increment ΔW_k.

Parallel associative scan. The solution of the affine recurrence Eq. 3 can be written $Z_k = \widehat{F}_k Z_0 + \widehat{g}_k$ with $(\widehat{F}_k, \widehat{g}_k) = (F_{k-1}, g_{k-1}) \circ \cdots \circ (F_0, g_0)$, where

$$(F', g') \circ (F, g) := (F'F,\ F'g + g') \tag{4}$$

is composition of affine maps. This operation is associative, so all prefixes $(\widehat{F}_k, \widehat{g}_k)_{k\le N}$ can be computed with any bracketing, in particular along a balanced binary tree of depth $O(\log T)$, and the result equals the sequential recursion exactly (Proposition 11). We call this evaluation the *parallel associative scan* and say that a recurrence is *scan-compatible* when its one-step maps are affine in the state; this is the only name used for the scan in the

Remark 1. The scan does not reduce the total arithmetic work across all time steps, but it reduces the sequential dependence from $O(T)$ recurrent updates to $O(\log T)$ parallel depth, exposing the time dimension to parallel hardware.

.

2.2 FUNCTIONAL CALIBRATION OBJECTIVE

Let Y^θ denote the output path of a model with parameters θ. The data of a functional calibration problem are a finite family Φ of path functionals φ, each with a prescribed value c_φ of its expectation; no trajectory is observed. Given n independent model paths $Y^{\theta,1}, \dots, Y^{\theta,n}$ and the Monte-Carlo average $\widehat{\mathbb{E}}_{\theta,n}[\varphi] = n^{-1}\sum_{i=1}^n \varphi(Y^{\theta,i})$, the objective is

$$\mathcal{L}_n(\theta) = \sum_{\varphi\in\Phi} w_\varphi \Big(\widehat{\mathbb{E}}_{\theta,n}[\varphi] - c_\varphi\Big)^2, \qquad w_\varphi > 0, \tag{5}$$

minimised by stochastic gradient descent with a fresh sample of n paths at every step. In this context no model path is paired with a data path: Eq. 5 is a functional of the model law rather than a samplewise error. Since $\mathbb{E}[(\widehat{\mathbb{E}}_{\theta,n}[\varphi] - c_\varphi)^2] = (\mathbb{E}_\theta[\varphi] - c_\varphi)^2 + n^{-1}\operatorname{Var}_\theta(\varphi)$, it converges almost surely to the population objective $\mathcal{L}_\infty(\theta) = \sum_\varphi w_\varphi(\mathbb{E}_\theta[\varphi] - c_\varphi)^2$, the Monte-Carlo variance entering at order $1/n$ only, so there is no incentive to collapse the model law onto a mean. Finitely many expectation constraints do not determine the path law, and $\mathcal{L}_\infty$ has no unique minimiser; this is a property of the task rather than of the objective, and the selected model is validated on functionals it was never fitted to. Where a reference dynamics is available, a relative-entropy penalty turns Eq. 5 into a selection rule that drops into the training loop unchanged. Appendix A.2 develops these points and relates the objective to divergence-based generative training.

3 THE SLISDE MODEL

A SLiSDE is built by composing structured linear stochastic layers through a stacking mechanism, followed by a learned linear readout. The optional Girsanov component is described separately in Section 4.

3.1 STRUCTURED LINEAR BASE LAYER

For latent dimension d and Brownian dimension m, each structured layer is an Itô SDE of the form

$$\mathrm{d}Z_t = \big(A_t Z_t + b_t\big)\,\mathrm{d}t + \sum_{j=1}^{m} \big(C_t^j Z_t + d_t^j\big)\,\mathrm{d}W_t^j, \quad Z_0 = z_0, \tag{6}$$

where $A_t, C_t^j \in \mathbb{R}^{d\times d}$ and $b_t, d_t^j \in \mathbb{R}^d$ and we assume that $z_0 \in L^2(\Omega, \mathcal{F}_0, \mathbb{P}; \mathbb{R}^d)$. Under coefficient bounds that hold by construction (Appendix B.1), the layer has a unique strong solution with moment bounds independent of the time grid (Theorem 4). We write $(A_t, b_t, C_t^1, \ldots, C_t^m, d_t^1, \ldots, d_t^m)$ for the time-indexed coefficient family induced by the model parameters.

These parameters contain static structured coefficients $(A, b, C^1, \ldots, C^m, d^1, \ldots, d^m) \in \Theta$, where Θ denotes the set of trainable weights, together with optional time-feature decoder weights. The matrices A and C^j are chosen from a fixed structured family: **diagonal**, **block-diagonal**, or **dense**. The additive diffusion vectors d^j may be (jointly) represented densely or through a low-rank parameterisation.

Remark 2 (Time-dependent coefficients)**.** The map $A \mapsto A_t$, and similarly for b, C^j, d^j, is optional and makes the coefficients time dependent through a small deterministic time-feature decoder (described in Appendix C.2). In both cases the layer remains linear in the current state Z_t.

Remark 3 (Noise type)**.** The noise type may be controlled by switching off parts of the diffusion. If $d_t^j = 0$ for all j, the layer has only multiplicative noise; if $C_t^j = 0$ for all j, the layer has only additive noise; and if both terms are retained, the layer has general affine diffusion.

After discretisation, the layer produces affine transition pairs (F_k, g_k) of the form of Eq. 3, with the static coefficients of the transitions of Appendix C.1 replaced by their time-dependent values at t_k.

The choice of matrix structure is important computationally: diagonal, block-diagonal, and dense matrices are closed under matrix multiplication, so products of transition matrices remain in the same family. This closure is what allows the affine maps to be composed by the parallel associative scan with $O(\log T)$ parallel depth without leaving the chosen representation. By contrast, other cheap matrix parameterization families such as fixed-rank low-rank or DPLR matrices do not satisfy this property, thus requiring sequential recursion and $O(T)$ complexity.

3.2 RESIDUAL STACKING SEQUENCE-MODEL BASELINE)

Sequence models such as Mamba (Gu and Dao, 2023) and the linear CDE SLiCE (Walker et al., 2025) build depth by *residual stacking*: a sequence block $\mathcal{S}^{(\ell)}$ is applied to a normalised hidden path and added back, $H^{(\ell)} = H^{(\ell-1)} + \mathcal{S}^{(\ell)}(\mathrm{Norm}(H^{(\ell-1)}))$, so a single hidden path is refined layer by layer. This is the natural depth mechanism to compare against. It is, however, ill-suited to our setting: each SLiSDE layer is itself a Brownian-driven SDE, so reusing the previous layer's already-Brownian path as the driver of the next would turn that layer into a *deterministic* transform of an already-stochastic input, adding compositional depth but injecting no fresh randomness per layer. We therefore do not stack SLiSDE layers residually; instead we couple them through the flow itself, via gated in-flow stacking (Section 3.3).

3.3 GATED IN-FLOW STACKING

Our main architectural contribution is *gated in-flow stacking*: a selective, path-dependent modulation of the next layer's affine flow, in the spirit of selective state-space models but adapted to stochastic latent dynamics. Layer ℓ builds its affine transition pairs $(F_k^{(\ell)}, g_k^{(\ell)})$, $k = 0, \ldots, N-1$, from its own structured coefficients and Brownian increments. Before they are scanned, causal features $x_k^{(\ell-1)}$ of the previous layer's path (its normalised state and the time features) produce, through two gated

linear units, a diagonal scale $\alpha_k^{(\ell)} \in (\mathbf{1}-\varepsilon, \mathbf{1}+\varepsilon)^d$ and an offset $o_k^{(\ell)} \in \mathbb{R}^d$; the gate architecture, the construction of the gate input and alternative parameterisations are given in Appendix C.4.

The affine pair is then modified by scaling only the diagonal entries of the transition matrix and by adding the offset correction,

$$\bar{F}_k^{(\ell)} = \mathcal{D}(\alpha_k^{(\ell)}, F_k^{(\ell)}) \qquad \text{and} \qquad \bar{g}_k^{(\ell)} = g_k^{(\ell)} + o_k^{(\ell)}, \tag{7}$$

where $\mathcal{D}$ rescales the diagonal of $F_k^{(\ell)}$ entrywise by $\alpha_k^{(\ell)}$ (Appendix C.4), and layer ℓ is evaluated by scanning the modified affine recurrence

$$Z_{k+1}^{(\ell)} = \bar{F}_k^{(\ell)} Z_k^{(\ell)} + \bar{g}_k^{(\ell)}. \tag{8}$$

The key point is that gated in-flow introduces nonlinearity through the dependence of $(\bar{F}_k^{(\ell)}, \bar{g}_k^{(\ell)})$ on the previous path $Z^{(\ell-1)}$, while preserving affine dependence on the current state $Z_k^{(\ell)}$; the recurrence is therefore evaluated exactly by the parallel associative scan (Proposition 11(b), Appendix B).

3.4 Linear decoder

After the final stacking layer, the model maps the last latent path to the observation space through a learned linear readout:

$$Y_t = \Pi_o Z_t^{(L)}, \qquad \Pi_o \in \mathbb{R}^{p\times d}, \tag{9}$$

where $Z^{(L)}$ denotes the path produced by the final layer and $\Pi_o \in \Theta$. When the initial observation is prescribed, we apply an optional shift so that the generated path satisfies the required value at $t = 0$.

3.5 Discretisation and scan

At each time step, every layer builds an affine pair (F_k, g_k) by freezing the time-dependent coefficients of Eq. 6 at t_k, using either the matrix-exponential or the Euler–Maruyama transition of Appendix C.1. The previous-layer path then modifies the pairs to $(\bar{F}_k^{(\ell)}, \bar{g}_k^{(\ell)})$ as in Eq. 7. Once the previous layers are fixed, each layer is therefore an affine recurrence in its own state, and the modified recurrence is evaluated, layer after layer, by the parallel associative scan of Section 2.1 (Proposition 11); the strong error of the resulting layerwise Euler–Maruyama scheme is of order $1/2$ for every fixed depth (Theorem 16, Appendix B).

4 Last-Layer Girsanov Tilt

Calibration losses on rare events receive little signal from plain Monte-Carlo paths under $\mathbb{P}$: an event of probability p estimated from n paths has relative standard error $\sqrt{(1-p)/(pn)}$, so more paths help only as $n^{-1/2}$. We therefore simulate under a tilted measure $\mathbb{Q}$ that visits the rare region more often and reweight by $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$; for diffusion paths Girsanov's theorem (Appendix A.1) gives the ratio in closed form for an adapted drift shift, and choosing the shift as a learned controller yields an importance sampler that acts on the constant $1/\sqrt{p}$ itself.

The shift is confined to the final layer, so that the prefix dynamics calibrated by the vanilla loss are not perturbed. The final layer is driven by a partially independent innovation ϵ, and only the innovation is tilted: under $\mathbb{Q}$ it is shifted by an adapted control $u_{\theta,k} \in \mathbb{R}^m$, a function of the innovation history and of stop-gradient features of the prefix latent path,

$$\mathrm{d}W_k^{(L)} = \rho\, \mathrm{d}W_k^{\mathrm{prefix}} + \sqrt{1-\rho^2}\, \mathrm{d}\epsilon_k, \qquad \rho \in [0,1), \tag{10}$$

$$\mathrm{d}\epsilon_k^{\mathbb{Q}} = \mathrm{d}\epsilon_k^{\mathbb{P}} - u_{\theta,k}\, \mathrm{d}t. \tag{11}$$

In a gated stack the random initial latent $z_0 \sim \mathcal{N}(0, I_d)$ carries a large share of the terminal variance, so the tilt also shifts the initial state, $z_0 = \mu_\theta + \xi$ with $\xi \sim \mathcal{N}(0, I_d)$ and a learned $\mu_\theta \in \mathbb{R}^d$. The discrete log-likelihood ratio is then

$$\log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \mu_\theta^\top z_0 - \tfrac{1}{2}\|\mu_\theta\|^2 + \sum_{k=0}^{N-1} \left(u_{\theta,k}^\top \mathrm{d}\epsilon_k^{\mathbb{P}} - \tfrac{1}{2}\|u_{\theta,k}\|^2\, \mathrm{d}t \right), \tag{12}$$

which is exact, normalised ($\mathbb{E}^{\mathbb{Q}}[\,\mathrm{d}\mathbb{P}/\,\mathrm{d}\mathbb{Q}] = 1$) and depends only on the control, the shift and the shifted innovation (Theorem 5). The tilt is a training-time device: the backbone minimises the calibration loss with the far-tail terms estimated on the tilted batch by self-normalised importance sampling, the controller is trained by the cross-entropy method under a wall on $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})$, and samples from the trained model are drawn under $\mathbb{P}$ without it. The rationale for the initial-state shift, the controller, the two objectives and the gradient convention they require are given in Appendix C.5.

5 Theoretical Results

This section states the guarantees behind the architecture; the assumptions in full, the auxiliary results and all proofs are collected in Appendix B. Three questions are answered here:

(i) whether the stacked model is a well-posed stochastic process with moment bounds that do not deteriorate with depth (Theorem 4);

(ii) whether the last-layer change of measure yields exact importance weights (Theorem 5);

(iii) what the class can represent (Theorem 6).

Two further facts are used throughout and are stated and proved in the appendix:

(iv) the parallel associative scan returns exactly the sequential iterates of every layer, gated or not, so that the depth-L stack is evaluated exactly in $O(L \log T)$ parallel depth (Proposition 11);

(v) the layerwise Euler–Maruyama scheme, with the gates evaluated on the discrete previous layer, has strong order $1/2$ for every fixed depth (Theorem 16).

Standing assumptions. The statements concern the model of Section 3 under three assumptions, stated in full in Appendix B.1. All of them hold by construction of the architecture; we list them only to fix what the theorems cover.

- *Model class* (Assumption 1): structured affine layers with the time-dependent coefficients of Appendix C.2, gated in-flow stacking with gates that read the normalised previous-layer state and the time features, arbitrary depth and widths, a linear readout, and a square-integrable initial state whose law may be degenerate.
- *Coefficients* (Assumption 2): the coefficients of every layer are bounded and regular in time. For the ungated base layer this is automatic, because the time features are smooth and the decoders are fixed networks; the gated layers inherit it from the regularity of the previous layer.
- *Gates* (Assumption 3): the gates are causal and Lipschitz in their input path, the flow gate stays within a fixed distance of one, and the offset gate is bounded for every fixed choice of its weights because its input is normalised.

The consequence that drives every constant is that the gated coefficients of a layer obey bounds that depend neither on the layer index nor on the depth of the stack.

Theorem 4 (Well-posedness)**.** *Let $q \geq 2$ and $\mathbb{E}\|z_0\|^q < \infty$.*

(a) Single layer. *Under Assumptions 1 and 2, the layer SDE Eq. 6 has a unique strong solution on $[0, T]$, and*

$$\mathbb{E} \sup_{0 \leq t \leq T} \|Z_t\|^q \leq K_q\big(1 + \mathbb{E}\|z_0\|^q\big),$$

where K_q depends only on $(a_*, b_*, c_*, d_*, m, T, q)$.

(b) Gated stack. *Under Assumptions 1–3, the depth-L stack of Section 3.3 has a unique strong solution, and the bound in (a) holds for every layer $Z^{(\ell)}$, $\ell = 1, \ldots, L$, with one constant K_q, which depends on $(a_*, b_*, c_*, d_*, m, T, q, \varepsilon, \Lambda_o)$ but neither on ℓ nor on L, for any correlation structure of the layer drivers.*

Theorem 5 (Validity of the last-layer Girsanov tilt)**.** *Assume the last-layer controller $u_{\theta,t}$ is progressively measurable with respect to the final-layer filtration and satisfies Novikov's condition, and let $\mu_\theta \in \mathbb{R}^d$ be a fixed initial-state shift, with $\mu_\theta = 0$ unless $z_0 \sim \mathcal{N}(0, I_d)$, then $\mathrm{d}\mathbb{Q}/\,\mathrm{d}\mathbb{P} = \exp\big(\mu_\theta^\top z_0 - \frac{1}{2}\|\mu_\theta\|^2\big)\,\mathcal{E}_T(u_\theta)$, with $\mathcal{E}_T$ the stochastic exponential of Eq. 14, defines an equivalent measure $\mathbb{Q} \sim \mathbb{P}$ under which $z_0 \sim \mathcal{N}(\mu_\theta, I_d)$, the prefix driver keeps its law, and*

$\epsilon_t^{\mathbb{Q}} = \epsilon_t - \int_0^t u_{\theta,s}\,\mathrm{d}s$ *is Brownian, the three being independent. For any integrable functional* Φ *of the full SLiSDE output,*

$$\mathbb{E}^{\mathbb{P}}[\Phi] = \mathbb{E}^{\mathbb{Q}}\left[\Phi\,\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right].$$

Moreover, the likelihood ratio depends only on the final-layer control, the initial-state shift and on the Brownian innovation being shifted.

Theorem 6 (Wasserstein density of terminal laws)**.** *Fix* $T > 0, m \geq 1$ *and an output dimension* $p \in \mathbb{N}$*, and let the gated layers use the RMS normalisation with a strictly positive regulariser and a gate input containing the time coordinate, as in Assumption 1. For every prescribed initial condition* $z_0 \in L^2(\Omega, \mathcal{F}_0, \mathbb{P}; \mathbb{R}^d)$*, possibly having a degenerate law, the collection of terminal laws generated by the model is dense in* $\mathcal{P}_2(\mathbb{R}^p)$ *with respect to the* 2*-Wasserstein distance. More precisely, if*

$$\mathfrak{L}_T := \left\{\mathcal{L}(Y_T^\theta) : \theta \text{ is an admissible model configuration}\right\},$$

where the union is taken over all admissible latent widths, then

$$\overline{\mathfrak{L}_T}^{W_2} = \mathcal{P}_2(\mathbb{R}^p).$$

One ungated layer followed by one gated layer suffices.

Theorem 6 is what makes the structural restrictions harmless: every target law with finite second moment is approximated arbitrarily well by the terminal law of an admissible configuration, so the selection among the diffusions consistent with the calibration constraints (Section 2.2) is never forced by a limitation of the class.

6 Experiments

We evaluate SLiSDE on three main benchmarks: a path-dependent toy dynamical system (TOY) and two option-surface datasets (DAX and SPX), where the TOY benchmark tests whether a model can match nonlinear long-horizon path statistics generated by a highly nonlinear stochastic dynamical system. Dataset construction, training protocols, and search grids are described in Appendix D. Additional results – ablations over the structural choices, the timing study, distributional recovery on TOY and the Girsanov tilt studies – are collected in Appendix E.

Models and baselines. We compare SLiSDE against two baselines: (i) a fully connected **Neural SDE** with neural drift and diffusion, and (ii) **SLiCE** (Walker et al., 2025). SLiCE is a natural structured-sequence baseline because it is a scan-compatible linear CDE and it is more expressive than Mamba (Gu and Dao, 2023); in our experiments we drive it with Brownian motion to match the stochastic input used by SLiSDE. For all model families, we report results for both two-layer and three-layer configurations.

Complexities. Table 1 reports the recurrent and scan costs for the SLiSDE transition structures. The scan reduces sequential depth from $O(T)$ to $O(\log T)$ when affine composition is available. For a generic Neural SDE, no analogous parallel associative scan is available, so simulation remains sequential in time. Dense transitions remain computationally expensive because each scan composition

	Diagonal	Block-diagonal	Dense	NeuralSDE
Recurrent cost	$O(dT)$	$O(n_b b_s^2 T)$	$O(d^2 T)$	$O(d^2 T)$
Scan cost	$O(d \log T)$	$O(n_b b_s^2 \log T)$	$O(d^3 \log T)$	—

Table 1: Computational cost of composing affine transition maps over T time steps. Here d is the latent dimension, b_s is the block size, and $n_b = d/b_s$ is the number of blocks. Typically $n_b << d$.

requires dense matrix multiplication; this can offset the practical gains from parallelisation, and we therefore exclude dense SLiSDE variants from the main experimental analysis.

Implementation and Compute. All models and experiments are implemented in JAX. The code is available at `https://anonymous.4open.science/r/SLiSDE-3B86/`. All experiments were run on a single NVIDIA RTX 6000 Ada Generation GPU.

6.1 PATH-DEPENDENT TOY BENCHMARK (TOY)

TOY is a controlled path-functional benchmark generated from a highly nonlinear stochastic dynamical system on $[0, 1]$, discretised with $N = 2048$ time steps (see Appendix D). The task is to calibrate the model law to the functional targets of Appendix D – thresholded positive parts, running maximum, squared path average and threshold-crossing probabilities at five evaluation times – through the objective of Eq. 5. In this first experiment, all models are trained without the Girsanov tilt, so the comparison isolates the effect of the architecture and stacking mechanism.

Model	Params	Loss (10^{-4})	Barrier MAE (10^{-2})	ms / epoch
Neural SDE – $L = 2$	124K	1.557 ± 0.206	1.058 ± 0.110	91.7
SLiCE – $L = 2$	142K	4.690 ± 0.423	2.168 ± 0.102	171.8
SLiSDE gated in-flow – $L = 2$	103K	$\mathbf{0.852 \pm 0.168}$	$\mathbf{0.658 \pm 0.054}$	**72.9**
Neural SDE – $L = 3$	157K	3.024 ± 0.341	1.637 ± 0.120	110.4
SLiCE – $L = 3$	213K	5.006 ± 0.914	2.250 ± 0.252	248.2
SLiSDE gated in-flow – $L = 3$	74K	$\mathbf{0.659 \pm 0.059}$	$\mathbf{0.672 \pm 0.057}$	**68.9**

Mean $\pm$ standard error over seven seeds per cell; best result per column in bold; the loss is $\mathcal{L}_n(\theta)$ of Eq. 5 on a fresh Monte-Carlo sample.

Table 2: TOY results without Girsanov tilt. L is the number of stacked layers; each row is the tuned cell of the search grid for that family and depth, retrained under the common protocol of Appendix D. We report parameter count, calibration loss, threshold-crossing error, and time per epoch.

Table 2 shows that SLiSDE substantially improves both accuracy and efficiency on the toy path-functional benchmark. The gated in-flow variant achieves the lowest total loss for both $L = 2$ and $L = 3$, while using fewer parameters than the Neural SDE and SLiCE baselines. In terms of runtime, SLiSDE remains faster than the Neural SDE baseline, reflecting the benefit of scan-compatible structured dynamics. SLiCE, which shares the parallel associative scan, is not designed for random inputs: it treats the Brownian path as one more input sequence, and its gap shows that deterministic sequence models do not transfer to a task that requires genuinely nonlinear stochastic dynamics. The recovery of the full path law beyond the calibrated targets (Kolmogorov–Smirnov and Wasserstein-1 distances to the ground truth for the terminal value and the running maximum) is reported in Appendix E.

6.2 OPTION-SURFACE BENCHMARK (DAX)

DAX is a real-data functional calibration benchmark built from historical European option quotes on the DAX index across multiple threshold levels and maturities. The task is to learn a stochastic path generator whose terminal functionals match the observed option surface. For evaluation, we hold out 25% of the threshold levels in each target group using a structured non-contiguous mask, so that the model is tested on interpolation across the surface rather than extrapolation outside the observed range.

The same protocol is applied to the SPX surface (Appendix D): there SLiSDE matches the held-out loss of SLiCE within one standard error with $8\times$ fewer parameters and $4\times$ faster epochs, attains the lowest far-tail loss with three layers, and both structured models are far ahead of the Neural SDE (Table 4, Appendix E).

Out-of-the-money (OTM) options correspond to thresholds that are far from the current level of the underlying index, so their value is determined by paths that end in the tails of the distribution.

Table 3 shows the real-data benchmark. With three layers, SLiSDE gated in-flow attains the lowest held-out loss and the lowest far-tail loss of all models, at a parameter count comparable to SLiCE and in half of its time per epoch; with two layers SLiCE is better, so on this surface the advantage of SLiSDE comes with depth. The Neural SDE is the weakest at both depths.

Girsanov tilt. With the tilt of Section 4 the far-tail calibration of the DAX model improves at a fixed path budget: over seven seeds the far-call held-out error halves relative to the vanilla model ($0.50\times$, paired $p = 0.03$) and the total held-out loss drops by 15–24%, while the budget-matched untilted model is never better on the far tails (Table 9). In a controlled rare-event study on TOY, the

Model	Params	Loss (10^{-4})	Far put (10^{-6})	ms / epoch
Neural SDE – $L = 2$	79K	1.972 ± 0.117	9.87 ± 2.47	91.4
SLiCE – $L = 2$	142K	$\mathbf{1.605 \pm 0.149}$	$\mathbf{8.14 \pm 1.69}$	169.9
SLiSDE gated in-flow – $L = 2$	90K	1.552 ± 0.070	9.40 ± 2.30	**79.7**
Neural SDE – $L = 3$	157K	1.908 ± 0.180	8.20 ± 1.42	**114.4**
SLiCE – $L = 3$	213K	1.560 ± 0.086	6.41 ± 1.33	246.9
SLiSDE gated in-flow – $L = 3$	135K	$\mathbf{1.401 \pm 0.244}$	$\mathbf{4.07 \pm 0.62}$	129.1

Mean $\pm$ standard error over seven seeds per cell; best result per column in bold; losses are $\mathcal{L}_n(\theta)$ of Eq. 5 on the held-out strikes, from a fresh Monte-Carlo sample.

Table 3: DAX results without Girsanov tilt. L is the number of stacked layers; each row is the tuned cell of the search grid for that family and depth, retrained under the common protocol of Appendix D. We report parameter count, held-out loss, held-out far-put (OTM) loss, and time per epoch.

tilt with the vanilla importance-sampling estimator reduces the relative error of plain Monte Carlo at equal budget by a factor of 2 to 4, growing with the rarity of the event (Table 10); both studies are reported in Appendix E.

Timing. Wall-clock times of full training steps for batch sizes up to 4096 and horizons up to $T = 8192$ are reported in Appendix E. At equal parameter count, SLiSDE evaluated sequentially is 5–7$\times$ faster per step than the Neural SDE at $B = 64$ and 1.2–1.4$\times$ faster at $B = 1024$; only at $B = 4096$, where the GPU is saturated by batch parallelism, is the Neural SDE step faster (154 against 242 ms).

7 Conclusion

We introduced *SLiSDE*, a family of Neural SDEs that moves nonlinearity from per-step drift and diffusion evaluations to *gated in-flow* coupling between structured linear stochastic layers. Each layer remains affine in its own state and is therefore parallel-scannable, while cross-layer modulation, time-dependent coefficients, and a learned linear decoder provide expressivity. For tail-sensitive path-functional objectives, we also use a last-layer Girsanov tilt, together with a shift of the initial latent state, with a closed-form Radon–Nikodym correction, yielding a lightweight importance-sampling overlay. We prove well-posedness, scan compatibility, discretisation error bounds, and validity of the change of measure, and that the terminal laws of the gated stack are dense in $\mathcal{P}_2$: the structural restrictions cost nothing in expressive power at the level of terminal laws.

Empirically, SLiSDE improves both accuracy and efficiency over the fully neural SDE baseline on the path-dependent TOY benchmark and on the DAX and SPX option surfaces, and over the structured-CDE baseline on TOY and on the three-layer DAX surface, while matching it on SPX with a fraction of the parameters. The tilt is a variance-reduction device: on TOY it reduces the error of rare-event estimates by factors of two to four at equal budget, growing with the rarity of the event, and on the quote-backed far tails of DAX it improves the far-tail calibration of the vanilla model, while budget-matched plain Monte Carlo remains a strong competitor where the targets are not rare. The results suggest that structured linear stochastic layers, combined with selective in-flow coupling and optional importance sampling, are a practical alternative to fully neural drift–diffusion networks for long-horizon functional calibration.

Future work. From the stochastic-analysis point of view, a natural direction is to better understand gated in-flow stacking as a path-dependent transformation of linear diffusions, including its possible connections to filtering. A second direction is to study structured Neural SDEs beyond functional calibration, in particular as continuous-time generative models such as Neural-SDE GANs.

AI DISCLOSURE

In this work, we used generative AI tools to help implement some of the ablation experiments and to edit the paper for readability. We have reviewed all AI-assisted work: the ablation code was verified and tested by the authors, and all edited text was checked against the original content.

REFERENCES

Marco Avellaneda, Craig Friedman, Richard Holmes, and Dominick Samperi. Calibrating volatility surfaces via relative-entropy minimization. *Applied Mathematical Finance*, 4(1):37–64, 1997.

Guy E. Blelloch. Prefix sums and their applications. Technical Report CMU-CS-90-190, School of Computer Science, Carnegie Mellon University, 1990.

Hans Buehler, Lukas Gonon, Josef Teichmann, and Ben Wood. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper/7892-neural-ordinary-differential-equations.

Samuel N. Cohen, Christoph Reisinger, and Sheng Wang. Arbitrage-free neural-SDE market models. *arXiv preprint arXiv:2105.11053*, 2021.

Christa Cuchiero, Wahid Khosrawi, and Josef Teichmann. A generative adversarial network approach to calibration of local stochastic volatility models. *Risks*, 8(4):101, 2020.

Patryk Gierjatowicz, Marc Sabate-Vidales, David Šiška, Lukasz Szpruch, and Žan Žurič. Robust pricing and hedging via neural SDEs. *arXiv preprint arXiv:2007.04154*, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Carsten Hartmann and Lorenz Richter. Nonasymptotic bounds for suboptimal importance sampling. *SIAM/ASA Journal on Uncertainty Quantification*, 12(2):309–346, 2024. doi: 10.1137/21M1427760. URL https://epubs.siam.org/doi/10.1137/21M1427760.

Carsten Hartmann, Lorenz Richter, Christof Schutte, and Wei Zhang. Variational characterization of free energy: Theory and algorithms. *Entropy*, 19(11):626, 2017. doi: 10.3390/e19110626. URL https://www.mdpi.com/1099-4300/19/11/626.

Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, and Daniela Rus. Liquid structural state-space models, 2022. URL https://arxiv.org/abs/2209.12951.

Zacharia Issa, Blanka Horvath, Maud Lemercier, and Cristopher Salvi. Non-adversarial training of neural SDEs with signature kernel scores. In *Advances in Neural Information Processing Systems*, 2023.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

Patrick Kidger, James Foster, Xuechen Li, Harald Oberhauser, and Terry Lyons. Neural SDEs as infinite-dimensional GANs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021. URL https://proceedings.mlr.press/v139/kidger21b.html.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, 1992. doi: 10.1007/978-3-662-12616-5.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3870–3882. PMLR, 2020. URL https://proceedings.mlr.press/v108/li20i.html.

Xuanqing Liu, Tesi Xiao, Si Si, Qin Cao, Sanjiv Kumar, and Cho-Jui Hsieh. Neural SDE: Stabilizing neural ODE networks with stochastic noise, 2019. URL https://arxiv.org/abs/1906.02355.

Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. *Advances in Neural Information Processing Systems*, 37:127226–127272, 2024.

Philip E. Protter. *Stochastic Integration and Differential Equations*. Springer, Berlin, 2 edition, 2005. doi: 10.1007/978-3-662-10061-5.

H. Smith et al. Calibrating generative models to distributional constraints. In *International Conference on Machine Learning*, 2026. TODO: complete author list and pages (cited by Reviewer X1pF as "Smith, H. et al., ICML 2026").

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *CoRR*, abs/1905.09883, 2019. URL https://arxiv.org/abs/1905.09883.

Benjamin Walker, Lingyi Yang, Nicola Muca Cirone, Cristopher Salvi, and Terry Lyons. Structured linear CDEs: Maximally expressive and parallel-in-time sequence models, 2025. URL https://arxiv.org/abs/2505.17761.

Qinsheng Zhang and Yongxin Chen. Path integral sampler: A stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_uCb2ynRu7Y.

A Mathematical Background Continued

A.1 Girsanov change of measure

This subsection states the change of measure on which the tilt of Section 4 rests.

Let u_t be a progressively measurable $\mathbb{R}^m$-valued process satisfying Novikov's condition

$$\mathbb{E}^{\mathbb{P}}\left[\exp\left(\frac{1}{2}\int_0^T \|u_t\|^2 \,\mathrm{d}t\right)\right] < \infty. \tag{13}$$

Define

$$\mathcal{E}_T(u) = \exp\left(\int_0^T u_t^\top \,\mathrm{d}W_t - \tfrac{1}{2}\int_0^T \|u_t\|^2 \,\mathrm{d}t\right). \tag{14}$$

Then $\mathcal{E}_T(u)$ is a martingale density. If $\mathrm{d}\mathbb{Q}/\,\mathrm{d}\mathbb{P} = \mathcal{E}_T(u)$, then

$$W_t^{\mathbb{Q}} = W_t - \int_0^t u_s \,\mathrm{d}s$$

is a Brownian motion under $\mathbb{Q}$. Moreover, for any integrable $\mathcal{F}_T$-measurable path functional Φ,

$$\mathbb{E}^{\mathbb{P}}[\Phi] = \mathbb{E}^{\mathbb{Q}}\left[\Phi\,\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right]. \tag{15}$$

The inverse density $\mathrm{d}\mathbb{P}/\,\mathrm{d}\mathbb{Q}$ is the importance weight that turns expectations under the tilted measure back into reference-measure expectations.

A.2 Calibration objective: population limit, non-uniqueness and model selection

Population limit. With the notation of Section 2.2, assume $\varphi(Y) \in L^2(\mathbb{P}_\theta)$ for every $\varphi \in \Phi$. Since $\widehat{\mathbb{E}}_{\theta,n}[\varphi(Y)]$ is an average of n independent copies,

$$\mathbb{E}\Big[\big(\widehat{\mathbb{E}}_{\theta,n}[\varphi(Y)] - c_\varphi\big)^2\Big] = \big(\mathbb{E}_{\mathbb{P}_\theta}[\varphi(Y)] - c_\varphi\big)^2 + \frac{1}{n}\operatorname{Var}_{\mathbb{P}_\theta}\big(\varphi(Y)\big),$$

and by the strong law of large numbers, almost surely as $n \to \infty$,

$$\mathcal{L}_n(\theta) \longrightarrow \mathcal{L}_\infty(\theta) = \sum_{\varphi\in\Phi} w_\varphi\big(\mathbb{E}_{\mathbb{P}_\theta}[\varphi(Y)] - c_\varphi\big)^2.$$

The objective is therefore a functional of the model *law*. The decomposition

$$\mathbb{E}\|X_t' - X_t\|^2 = \operatorname{Var}(X_t') + \operatorname{Var}(X_t) + \|\mathbb{E}X_t' - \mathbb{E}X_t\|^2,$$

which shows that a samplewise squared error between model and data paths rewards collapse of the model onto the data mean, does not apply: no model path is paired with a data path, and the only variance term in $\mathcal{L}_n$ carries the coefficient $1/n$ and vanishes in the population limit. In all experiments $n \geq 1024$.

No collapse. Collapse is moreover incompatible with the population objective whenever the target marginal is non-degenerate. If $Y_t = m_t$ almost surely then

$$\mathbb{E}_{\mathbb{P}_\theta}[(Y_t - K)_+] = (m_t - K)_+,$$

a piecewise affine function of K with a single change of slope at $K = m_t$, which cannot coincide on an interval with a strictly convex target map $K \mapsto c_{t,K}$ such as an arbitrage-free call surface. More generally

$$\mathbb{E}_{\mathbb{P}_\theta}[(Y_t - K)_+] = \int_K^\infty \mathbb{P}_\theta(Y_t > x)\,\mathrm{d}x,$$

so knowledge of the positive-part expectations for every K determines the marginal law of Y_t; the finite strike grid of the implementation provides a finite collection of constraints of this type, and the running-maximum and threshold-crossing functionals constrain path-dependent features in addition. We do not claim that a finite collection of such constraints determines the entire path law.

Non-uniqueness and model risk. Finitely many expectation constraints are satisfied by infinitely many diffusions, so $\mathcal{L}_\infty$ has no unique minimiser. *This is a property of functional calibration rather than of the objective: the data are finitely many expectations, so no objective built on them can identify the path law.* In derivatives markets one typically has nothing beyond quoted prices, and it is standard practice to work with – and deliberately explore – the multiplicity of models consistent with the same quotes in order to quantify model risk; Gierjatowicz et al. (2020), for example, use randomised calibrated neural SDEs to obtain price *bounds* over the calibrated set. Enforcing uniqueness through an auxiliary term would not remove the ambiguity but shift it to the choice of the reference measure. Theorem 6 guarantees that the model class itself never restricts what is reachable, so the selection among constraint-consistent processes is driven by the data, the initialisation and the training procedure, never by a hidden limitation of the class. Operationally we validate the selection where a selection can be validated, on functionals the model was never fitted to: the held-out-strike transfer on DAX and SPX, and the distributional-recovery study on TOY (Kolmogorov–Smirnov and Wasserstein-1 distances to the ground-truth simulator, Appendix E).

Model selection within the calibrated set. There are settings, physics in particular, where prior information *is* available: a reference dynamics, an equilibrium law, known symmetries. There, an additive penalty turns Eq. 5 into a principled selection rule – a Kullback–Leibler divergence to the available reference (Smith et al., 2026), a minimal-relative-entropy criterion in the spirit of Avellaneda et al. (1997) for volatility surfaces, or a context-specific penalty – and such terms drop into the training loop unchanged. The framework already computes and uses exactly this quantity where a reference measure genuinely exists: in the Girsanov layer, the controller is trained with a $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})$ anchor (Appendix C.8), the pathwise log Radon–Nikodym exponent that our machinery evaluates, which penalises tilts that move too far from the untilted reference model.

Relation to generative training. Training against a fixed family of statistics can be viewed as a maximum-mean-discrepancy objective with a non-characteristic kernel, a point discussed by Kidger et al. (2021), who contrast it with their adversarial formulation using a learned discriminator; the variational formulation of Tzen and Raginsky (2019) and the signature-kernel scores of Issa et al. (2023) likewise address a different, generative learning objective that requires sample paths. The method belongs instead to the calibration strand (Gierjatowicz et al., 2020; Cuchiero et al., 2020; Cohen et al., 2021), in which prescribed functional expectations, such as derivative prices, are matched.

B Theoretical Results: Assumptions, Auxiliary Results and Proofs

This appendix states the standing assumptions in full, establishes the two auxiliary results used throughout the paper (exactness of the parallel associative scan and the strong discretisation error of the stacked scheme) and proves every result of Section 5. It is organised as follows.

- Appendix B.1 states the assumptions and proves Lemma 8: a gated layer inherits the coefficient bounds of its base layer with constants that do not depend on the depth. Every later constant rests on this lemma.
- Appendix B.2 proves Theorem 4, first for one layer with random coefficients (Lemma 9) and then by induction over the layers.
- Appendix B.3 proves Proposition 11, the exactness of the scan.
- Appendix B.4 proves Theorem 16 in four steps: one-step moments (Lemma 13), discrete stability (Lemma 14), the Euler–Maruyama error of one layer (Lemma 15) and an induction over the layers that carries a doubled moment order at every step down the stack.
- Appendices B.5 and B.6 prove Theorems 5 and 6.

Notation. $\|\cdot\|$ denotes the Euclidean norm of a vector and the operator norm of a matrix, $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm and $\|x\|_\infty$ the largest modulus of the entries of x. For a coefficient tuple $\vartheta = (A, b, C^1, \dots, C^m, d^1, \dots, d^m)$ we write

$$\|\vartheta\| := \|A\| + \|b\| + \sum_{j=1}^{m} \big(\|C^j\| + \|d^j\|\big). \tag{16}$$

For a coefficient process $\vartheta_t = (A_t, b_t, C_t^1, \dots, C_t^m, d_t^1, \dots, d_t^m)$ we set

$$\mu(t,z) := A_t z + b_t, \qquad \sigma^j(t,z) := C_t^j z + d_t^j, \qquad \sigma(t,z) := \big[\sigma^1(t,z), \dots, \sigma^m(t,z)\big] \in \mathbb{R}^{d\times m}, \tag{17}$$

so that the layer SDE Eq. 6 reads

$$\mathrm{d}Z_t = \mu(t, Z_t)\ \mathrm{d}t + \sigma(t, Z_t)\ \mathrm{d}W_t.$$

The letter C denotes a constant that may change from line to line and depends only on the bounds in the assumptions, on T, on m and on the moment order under consideration; where a dependence matters it is made explicit. We use the Burkholder–Davis–Gundy inequality (Protter, 2005, Ch. IV) in the form

$$\mathbb{E}\sup_{s\le t}\Big\|\int_0^s H_u\ \mathrm{d}W_u\Big\|^q \le c_q\,\mathbb{E}\Big(\int_0^t \|H_u\|_{\mathrm{F}}^2\ \mathrm{d}u\Big)^{q/2}, \qquad q \ge 2, \tag{18}$$

valid for every progressively measurable $\mathbb{R}^{d\times m}$-valued process H with finite right-hand side. Finally,

$$m_r := \mathbb{E}|\xi|^r, \qquad \xi \sim \mathcal{N}(0,1),$$

denotes the absolute moments of the standard normal law.

B.1 Standing assumptions

Assumption 1 (Admissible model class)**.** Fix $T > 0$, a Brownian dimension $m \ge 1$ and an output dimension $p \in \mathbb{N}$. Let $W = (W^1, \dots, W^m)$ be an m-dimensional Brownian motion on the filtered probability space of Section 2 and fix $z_0 \in L^2(\Omega, \mathcal{F}_0, \mathbb{P}; \mathbb{R}^{d_0})$ for some $d_0 \in \mathbb{N}$, without any non-degeneracy assumption on its law. For every $d \ge d_0$ set

$$\iota_d(z_0) := (z_0, 0, \dots, 0) \in \mathbb{R}^d.$$

An admissible model configuration consists of a finite number L of stacked layers of a common latent width $d \ge d_0$, each driven by a Brownian motion $W^{(\ell)}$ adapted to the common filtration (shared, independent or correlated across layers; the implementation shares one driver across the prefix layers and uses the correlated driver of Eq. 10 for the last one), such that:

1. each base layer is the structured affine Itô SDE Eq. 6 with $Z_0 = \iota_d(z_0)$, whose coefficients $A_t, C_t^j \in \mathbb{R}^{d\times d}$ and $b_t, d_t^j \in \mathbb{R}^d$ are the static parameters A, C^j, b, d^j perturbed by the time-feature decoder of Appendix C.2, which adds deterministic, smooth-in-time corrections to the free coefficient entries and respects the chosen diagonal, block-diagonal or dense structure;

2. successive layers are connected by gated in-flow stacking (Section 3.3) through the gates of Eqs. 81 and 82, with fixed $\varepsilon \in (0,1)$, and the gate input is

$$x_t^{(\ell-1)} = \big[\Pi(Z_t^{(\ell-1)}),\, \tau(t)\big], \qquad \Pi(z) := \gamma \odot \frac{z}{\sqrt{\rho + d^{-1}\|z\|^2}}, \tag{19}$$

   where Π is the RMS normalisation with gain $\gamma \in \mathbb{R}^d$ and regulariser $\rho > 0$, and τ is the vector of deterministic time features, which contains the time coordinate t and is Lipschitz on $[0,T]$; all gate weights and biases are trainable, and the optional causal convolution is disabled;

3. the number of layers, the latent width and the hidden widths of the time decoders and gates may be increased arbitrarily, and their affine weights and biases range freely over the corresponding Euclidean spaces;

4. the output is a linear readout of the final layer,

$$Y_t^\theta = \Pi_o Z_t^{(L)}, \qquad \Pi_o \in \mathbb{R}^{p\times d},$$

   shifted to match a prescribed initial observation y_0 when one is given.

Remark 7 (Discrete and continuous gating)**.** The implementation gates the discrete affine pair,

$$\bar F_k = \mathcal{D}(\alpha_k, F_k), \qquad \bar g_k = g_k + o_k, \tag{20}$$

with no explicit Δt on the offset, where $\mathcal{D}(\alpha, F)$ rescales the i-th diagonal entry of F by α_i. The continuous-time model analysed below carries the gated coefficients

$$\alpha_t^{(\ell)} \odot A_t^{(\ell)} \quad \text{in place of} \quad A_t^{(\ell)}, \qquad b_t^{(\ell)} + o_t^{(\ell)} \quad \text{in place of} \quad b_t^{(\ell)}, \tag{21}$$

where $\alpha \odot A$ rescales the i-th diagonal entry of A by α_i, while the diffusion coefficients are not gated. The two conventions are related by the dictionary

$$o_k = o_{t_k}\,\Delta t, \qquad \alpha_k = \exp\big(\delta_{t_k}\Delta t\big) \text{ entrywise}, \tag{22}$$

where δ denotes the diagonal drift modulation. The first identity is exact because the linear branch of the offset gate is homogeneous in its weights, so that rescaling W_{io}, c_{io} by Δt stays within the family. The second is realisable on every fixed grid because α_k ranges over a neighbourhood of $\mathbf{1}$.

Assumption 2 (Coefficients)**.** For every layer, the coefficient process ϑ_t is $(\mathcal{F}_t)$-progressively measurable and uniformly bounded,

$$\|A_t\| \le a_*, \qquad \|b_t\| \le b_*, \qquad \|C_t^j\| \le c_*, \qquad \|d_t^j\| \le d_*, \tag{23}$$

almost surely, for all $t \in [0,T]$ and $j \le m$, and $\frac{1}{2}$-Hölder in time in every L^q: for each $q \ge 2$ there is $L_q < \infty$ with

$$\mathbb{E}\|\vartheta_t - \vartheta_s\|^q \le L_q^q\,|t-s|^{q/2}, \qquad s,t \in [0,T]. \tag{24}$$

For the ungated base layer the coefficients are deterministic and Lipschitz in t with a constant L_τ, because the time features are smooth and the decoders are fixed networks. Eq. 24 then holds with

$$L_q = L_\tau T^{1/2},$$

so Assumption 2 holds by construction. For the gated layers it is established in Lemma 8 below. The bounds Eq. 23 imply, for the maps of Eq. 17, the Lipschitz and linear-growth bounds

$$\begin{aligned} \|\mu(t,z) - \mu(t,z')\| + \sum_{j=1}^m \|\sigma^j(t,z) - \sigma^j(t,z')\| &\le \Lambda'\|z - z'\|, \\ \|\mu(t,z)\| + \|\sigma(t,z)\|_{\mathrm{F}} &\le \Lambda'\big(1 + \|z\|\big), \end{aligned} \tag{25}$$

almost surely for all t, z and z', with

$$\Lambda' := a_* + b_* + m\,(c_* + d_*). \tag{26}$$

Assumption 3 (Gates)**.** For each layer $\ell \geq 2$ the gate map sends a continuous path $z = (z_t)_{t\in[0,T]}$ of the previous layer to the gate processes,

$$z \longmapsto \big(\alpha_t^{(\ell)}(z),\, o_t^{(\ell)}(z)\big)_{t\in[0,T]}.$$

It is causal, in that $(\alpha_t^{(\ell)}, o_t^{(\ell)})$ depends on z only through $z_{[0,t]}$ (for the gates of Eqs. 81 and 82 with the input Eq. 19, only through z_t and t), and jointly measurable, so that applied to a continuous adapted process it yields progressively measurable processes. Moreover, with ρ_d and ρ_x defined in (a):

(a) *(Normalised input.)* For all $z, z' \in \mathbb{R}^d$,

$$\|\Pi(z)\| \leq \sqrt{d}\,\|\gamma\|_\infty =: \rho_d, \qquad \|\Pi(z) - \Pi(z')\| \leq L_\Pi \|z - z'\|, \qquad L_\Pi := \frac{\|\gamma\|_\infty}{\sqrt{\rho}}, \quad (27)$$

so that the gate input satisfies

$$\|x_t^{(\ell-1)}\| \leq \rho_d + \sup_{t\leq T} \|\tau(t)\| =: \rho_x.$$

(b) *(Bounded flow gate.)* Surely, for every input path and every choice of the weights,

$$\|\alpha_t^{(\ell)} - \mathbf{1}\|_\infty \leq \varepsilon.$$

(c) *(Offset bounded for fixed weights.)* Surely,

$$\|o_t^{(\ell)}\| \leq \|W_{io}^{(\ell)}\|\,\rho_x + \|c_{io}^{(\ell)}\| =: \Lambda_o < \infty; \quad (28)$$

Λ_o is finite for every fixed choice of the weights and free in magnitude.

(d) *(Lipschitz gate.)* There is $L_g < \infty$ such that, for all continuous paths z, z' and all $s \leq t \leq T$,

$$\|\alpha_t^{(\ell)}(z) - \alpha_t^{(\ell)}(z')\|_\infty + \|o_t^{(\ell)}(z) - o_t^{(\ell)}(z')\| \leq L_g \sup_{r\leq t} \|z_r - z'_r\|, \quad (29)$$

$$\|\alpha_t^{(\ell)}(z) - \alpha_s^{(\ell)}(z)\|_\infty + \|o_t^{(\ell)}(z) - o_s^{(\ell)}(z)\| \leq L_g\big(\|z_t - z_s\| + |t - s|\big). \quad (30)$$

All four properties hold for the implemented gates.

Property (a). Write

$$s(z) := \sqrt{\rho + d^{-1}\|z\|^2}, \qquad f(z) := \frac{z}{s(z)},$$

so that $\Pi(z) = \gamma \odot f(z)$ and $\|f(z)\| \leq \sqrt{d}$. The Jacobian of f is the symmetric matrix

$$Df(z) = s(z)^{-1} I - s(z)^{-3} d^{-1} z z^\top.$$

Its eigenvalue on $z^\perp$ is $s(z)^{-1}$, and its eigenvalue in the direction of z is

$$s(z)^{-1} - s(z)^{-3} d^{-1}\|z\|^2 = s(z)^{-3}\rho;$$

all eigenvalues therefore lie in $(0, \rho^{-1/2}]$. Hence $\|\Pi(z)\| \leq \|\gamma\|_\infty \sqrt{d}$, and Π is Lipschitz with constant $\|\gamma\|_\infty \rho^{-1/2}$.

Properties (b) and (c). They hold because $|\sigma| \leq 1$ and $|\tanh| \leq 1$ in Eq. 81, and because $|\sigma| \leq 1$ in Eq. 82 while the gate input lies in the ball of radius ρ_x.

Property (d). The gates are compositions of Π, affine maps and the Lipschitz activations σ and $\tanh$, all restricted to the bounded input set, together with the Lipschitz time features. The product structure of Eq. 82 is Lipschitz on that bounded set because both factors are bounded and Lipschitz there.

Lemma 8 (Gated layers satisfy Assumption 2 with depth-uniform constants)**.** *Let Assumptions 1 and 3 hold, let $\ell \geq 2$, and let the previous layer $Z^{(\ell-1)}$ be a continuous adapted process such that, for every $q \geq 2$,*

$$\mathbb{E}\sup_{t\leq T}\|Z_t^{(\ell-1)}\|^q < \infty, \qquad \mathbb{E}\|Z_t^{(\ell-1)} - Z_s^{(\ell-1)}\|^q \leq K_q' \,|t-s|^{q/2}, \qquad s,t\in[0,T]. \tag{31}$$

Then the gated coefficient process of layer ℓ,

$$\vartheta_t^{(\ell)} := \big(\alpha_t^{(\ell)} \odot A_t^{(\ell)},\ b_t^{(\ell)} + o_t^{(\ell)},\ C_t^{(\ell),1},\dots,C_t^{(\ell),m},\ d_t^{(\ell),1},\dots,d_t^{(\ell),m}\big), \tag{32}$$

with $(\alpha^{(\ell)}, o^{(\ell)})$ evaluated on $Z^{(\ell-1)}$, is progressively measurable and satisfies Assumption 2 with the bounds

$$\|\alpha_t^{(\ell)} \odot A_t^{(\ell)}\| \leq (1+\varepsilon)a_*, \qquad \|b_t^{(\ell)} + o_t^{(\ell)}\| \leq b_* + \Lambda_o, \qquad \|C_t^{(\ell),j}\| \leq c_*, \qquad \|d_t^{(\ell),j}\| \leq d_*, \tag{33}$$

and with Hölder constants

$$L_q^q = 3^{q-1}\Big(\big((1+a_*)L_g\big)^q\big(K_q' + T^{q/2}\big) + \big((2+\varepsilon)L_\tau\big)^q T^{q/2}\Big). \tag{34}$$

Consequently the maps of Eq. 17 built from $\vartheta^{(\ell)}$ satisfy Eq. 25 with the constant

$$\Lambda_\star' := (1+\varepsilon)a_* + b_* + \Lambda_o + m\,(c_* + d_*), \tag{35}$$

which depends neither on ℓ, nor on L, nor on the previous layer.

Proof. Step 1: measurability. By Assumption 3 the gate map is causal and jointly measurable, so its evaluation on the continuous adapted process $Z^{(\ell-1)}$ is progressively measurable. The base coefficients are deterministic and continuous in t. Hence $\vartheta^{(\ell)}$ is progressively measurable.

Step 2: bounds. Rescaling the i-th diagonal entry of $A_t^{(\ell)}$ by $\alpha_{t,i}^{(\ell)}$ adds a diagonal matrix to $A_t^{(\ell)}$,

$$\alpha_t^{(\ell)} \odot A_t^{(\ell)} = A_t^{(\ell)} + \operatorname{diag}\big((\alpha_{t,i}^{(\ell)} - 1)A_{t,ii}^{(\ell)}\big)_{i\leq d}, \qquad |A_{t,ii}^{(\ell)}| \leq \|A_t^{(\ell)}\|.$$

Assumption 3(b) therefore gives

$$\|\alpha_t^{(\ell)} \odot A_t^{(\ell)}\| \leq \|A_t^{(\ell)}\| + \max_i |\alpha_{t,i}^{(\ell)} - 1|\,|A_{t,ii}^{(\ell)}| \leq (1+\varepsilon)\,a_*. \tag{36}$$

By Assumption 3(c),

$$\|b_t^{(\ell)} + o_t^{(\ell)}\| \leq b_* + \Lambda_o,$$

and the diffusion coefficients are not gated. This proves Eq. 33. Eq. 25 with the constant $\Lambda_\star'$ follows exactly as Eq. 26 follows from Eq. 23.

Step 3: time regularity. Fix $s \leq t$ and abbreviate $\alpha_t = \alpha_t^{(\ell)}$, $o_t = o_t^{(\ell)}$ and $Z = Z^{(\ell-1)}$. The argument of Eq. 36 gives

$$\|\alpha_s \odot B\| \leq (1+\varepsilon)\|B\| \qquad \text{for every matrix } B,$$

and the base coefficients are Lipschitz in t with constant L_τ. Therefore

$$\begin{aligned}\|\vartheta_t^{(\ell)} - \vartheta_s^{(\ell)}\| &\leq \|(\alpha_t - \alpha_s)\odot A_t^{(\ell)}\| + \|\alpha_s \odot (A_t^{(\ell)} - A_s^{(\ell)})\| + \|o_t - o_s\| \\ &\quad + \|b_t^{(\ell)} - b_s^{(\ell)}\| + \sum_{j=1}^m \big(\|C_t^{(\ell),j} - C_s^{(\ell),j}\| + \|d_t^{(\ell),j} - d_s^{(\ell),j}\|\big) \\ &\leq a_*\|\alpha_t - \alpha_s\|_\infty + \|o_t - o_s\| + (2+\varepsilon)L_\tau|t-s|.\end{aligned} \tag{37}$$

By Eq. 30,

$$a_*\|\alpha_t - \alpha_s\|_\infty + \|o_t - o_s\| \leq (1+a_*)L_g\big(\|Z_t - Z_s\| + |t-s|\big).$$

Raising to the power q and using $(x+y+z)^q \leq 3^{q-1}(x^q + y^q + z^q)$,

$$\|\vartheta_t^{(\ell)} - \vartheta_s^{(\ell)}\|^q \leq 3^{q-1}\Big(\big((1+a_*)L_g\big)^q\|Z_t - Z_s\|^q + \big((1+a_*)L_g\big)^q|t-s|^q + \big((2+\varepsilon)L_\tau\big)^q|t-s|^q\Big).$$

Taking expectations and using Eq. 31 together with $|t-s|^q \leq T^{q/2}|t-s|^{q/2}$,

$$\mathbb{E}\|\vartheta_t^{(\ell)} - \vartheta_s^{(\ell)}\|^q \leq L_q^q\,|t-s|^{q/2}$$

with the constant L_q of Eq. 34, which is Eq. 24. □

B.2 WELL-POSEDNESS: PROOF OF THEOREM 4

We first treat one layer with a general random coefficient process; the time regularity Eq. 24 is not needed here.

Lemma 9 (One layer with random coefficients)**.** *Let ϑ be a progressively measurable coefficient process satisfying the bounds Eq. 23, let $q \geq 2$, and let Z_0 be $\mathcal{F}_0$-measurable with $\mathbb{E}\|Z_0\|^q < \infty$. Then the SDE*

$$\mathrm{d}Z_t = \mu(t, Z_t)\ \mathrm{d}t + \sigma(t, Z_t)\ \mathrm{d}W_t, \qquad Z_0 \text{ given},$$

has a unique strong solution on $[0, T]$, and

$$\mathbb{E}\sup_{t\leq T}\|Z_t\|^q \leq K_q\big(1 + \mathbb{E}\|Z_0\|^q\big), \tag{38}$$

$$\mathbb{E}\sup_{s\leq r\leq t}\|Z_r - Z_s\|^q \leq K'_q\big(1 + \mathbb{E}\|Z_0\|^q\big)\,|t-s|^{q/2}, \qquad 0 \leq s \leq t \leq T, \tag{39}$$

with constants K_q, K'_q depending only on (Λ', T, m, q).

Proof. Step 1: existence and uniqueness. By Eq. 25 the coefficients are progressively measurable, globally Lipschitz in z and of linear growth, uniformly in (t, ω). Existence and uniqueness of a strong solution for such random Lipschitz coefficients is classical (Protter, 2005, Ch. V, Thm. 7).

Step 2: moment bound. Fix $n \in \mathbb{N}$ and the localising time

$$\tau_n := \inf\{t \geq 0 : \|Z_t\| \geq n\} \wedge T.$$

For $t \leq T$, the integral form of the equation gives

$$Z_{t\wedge\tau_n} = Z_0 + \int_0^{t\wedge\tau_n} \mu(s, Z_s)\ \mathrm{d}s + \int_0^t \mathbf{1}_{\{s<\tau_n\}}\sigma(s, Z_s)\ \mathrm{d}W_s.$$

Using $\|x + y + z\|^q \leq 3^{q-1}(\|x\|^q + \|y\|^q + \|z\|^q)$,

$$\sup_{r\leq t}\|Z_{r\wedge\tau_n}\|^q \leq 3^{q-1}\Big[\|Z_0\|^q + \Big(\int_0^{t\wedge\tau_n}\|\mu(s, Z_s)\|\ \mathrm{d}s\Big)^q + \sup_{r\leq t}\Big\|\int_0^r \mathbf{1}_{\{s<\tau_n\}}\sigma(s, Z_s)\ \mathrm{d}W_s\Big\|^q\Big]. \tag{40}$$

On $\{s < \tau_n\}$ we have $Z_s = Z_{s\wedge\tau_n}$. By Jensen's inequality and Eq. 25,

$$\Big(\int_0^{t\wedge\tau_n}\|\mu(s, Z_s)\|\ \mathrm{d}s\Big)^q \leq T^{q-1}\int_0^t \|\mu(s, Z_{s\wedge\tau_n})\|^q\ \mathrm{d}s \leq T^{q-1}\,2^{q-1}\Lambda'^{\,q}\int_0^t \big(1 + \|Z_{s\wedge\tau_n}\|^q\big)\,\mathrm{d}s.$$

By Eq. 18, Jensen's inequality (for $q/2 \geq 1$) and Eq. 25,

$$\begin{aligned}\mathbb{E}\sup_{r\leq t}\Big\|\int_0^r \mathbf{1}_{\{s<\tau_n\}}\sigma(s, Z_s)\ \mathrm{d}W_s\Big\|^q &\leq c_q\,T^{q/2-1}\,\mathbb{E}\int_0^t \mathbf{1}_{\{s<\tau_n\}}\|\sigma(s, Z_s)\|_{\mathrm{F}}^q\ \mathrm{d}s \\ &\leq c_q\,T^{q/2-1}\,2^{q-1}\Lambda'^{\,q}\int_0^t \big(1 + \mathbb{E}\|Z_{s\wedge\tau_n}\|^q\big)\,\mathrm{d}s.\end{aligned}$$

Set

$$\phi_n(t) := \mathbb{E}\sup_{r\leq t}\|Z_{r\wedge\tau_n}\|^q,$$

which is finite because $\|Z_{r\wedge\tau_n}\| \leq \max(n, \|Z_0\|)$. Taking expectations in Eq. 40,

$$\phi_n(t) \leq 3^{q-1}\,\mathbb{E}\|Z_0\|^q + C_1\int_0^t \big(1 + \phi_n(s)\big)\,\mathrm{d}s, \qquad C_1 := 3^{q-1}2^{q-1}\Lambda'^{\,q}\big(T^{q-1} + c_q T^{q/2-1}\big),$$

and Grönwall's lemma gives

$$\phi_n(T) \leq \big(3^{q-1}\mathbb{E}\|Z_0\|^q + C_1 T\big)e^{C_1 T},$$

uniformly in n. Since the paths are continuous, $\tau_n \uparrow T$ almost surely and

$$\sup_{r\leq T}\|Z_{r\wedge\tau_n}\| \uparrow \sup_{r\leq T}\|Z_r\|,$$

so monotone convergence yields Eq. 38 with

$$K_q := \big(3^{q-1} + C_1 T\big) e^{C_1 T}.$$

Step 3: increments. Fix $s \le t$. For $r \in [s,t]$,

$$Z_r - Z_s = \int_s^r \mu(u, Z_u)\ \mathrm{d}u + \int_s^r \sigma(u, Z_u)\ \mathrm{d}W_u.$$

The same two inequalities, now without localisation because the moments are finite by Step 2, give

$$\begin{aligned}\mathbb{E} \sup_{s\le r\le t} \|Z_r - Z_s\|^q &\le 2^{q-1}\Big[(t-s)^{q-1}\,\mathbb{E}\int_s^t \|\mu(u,Z_u)\|^q\ \mathrm{d}u \\ &\qquad + c_q\,(t-s)^{q/2-1}\,\mathbb{E}\int_s^t \|\sigma(u,Z_u)\|_{\mathrm{F}}^q\ \mathrm{d}u\Big] \\ &\le 2^{q-1}\,2^{q-1}\Lambda'^{\,q}\big(1+\mathbb{E}\sup_{u\le T}\|Z_u\|^q\big)\big[(t-s)^q + c_q (t-s)^{q/2}\big].\end{aligned}$$

Since $(t-s)^q \le T^{q/2}(t-s)^{q/2}$, Eq. 38 gives

$$\mathbb{E} \sup_{s\le r\le t} \|Z_r - Z_s\|^q \le 4^{q-1}\Lambda'^{\,q}\big(1+K_q\big)\big(T^{q/2}+c_q\big)\big(1+\mathbb{E}\|Z_0\|^q\big)\,(t-s)^{q/2},$$

which is Eq. 39. □

Proof of Theorem 4. (a) Single layer. The base layer satisfies Assumption 2 by construction. Lemma 9 with $Z_0 = \iota_d(z_0)$ gives existence, uniqueness and Eq. 38 with

$$K_q = K_q(\Lambda', T, m, q),$$

where Λ' is the function Eq. 26 of (a_*, b_*, c_*, d_*, m).

(b) Gated stack. Let $\Lambda'_\star \ge \Lambda'$ be the constant Eq. 35, and write

$$K_q^\star := K_q(\Lambda'_\star, T, m, q), \qquad K_q'^\star := K'_q(\Lambda'_\star, T, m, q),$$

for the constants of Lemma 9 at this value; they depend on $(a_*, b_*, c_*, d_*, m, T, q, \varepsilon, \Lambda_o)$ only. We prove by induction on ℓ that layer ℓ has a unique strong solution and that, for every $q \ge 2$ with $\mathbb{E}\|z_0\|^q < \infty$,

$$\begin{aligned}\mathbb{E}\sup_{t\le T}\|Z_t^{(\ell)}\|^q &\le K_q^\star\big(1+\mathbb{E}\|z_0\|^q\big), \\ \mathbb{E}\|Z_t^{(\ell)} - Z_s^{(\ell)}\|^q &\le K_q'^\star\big(1+\mathbb{E}\|z_0\|^q\big)\,|t-s|^{q/2}, \qquad s,t\in[0,T].\end{aligned} \tag{41}$$

The base layer is (a), since $\Lambda' \le \Lambda'_\star$.

Let $\ell \ge 2$ and assume Eq. 41 for layer $\ell - 1$. Then $Z^{(\ell-1)}$ satisfies Eq. 31 for every such q. By Lemma 8, the gated coefficient process $\vartheta^{(\ell)}$ of Eq. 32 is progressively measurable and satisfies the bounds Eq. 33, hence Eq. 25 with the constant $\Lambda'_\star$. Lemma 9 applied to layer ℓ, with $Z_0 = \iota_d(z_0)$ and $\Lambda'_\star$ in place of Λ', gives existence, uniqueness and Eq. 41 for layer ℓ. The constants do not depend on ℓ, on L or on the previous layer, which closes the induction.

Only adaptedness of the drivers to the common filtration was used, so the conclusion holds for any correlation structure of the layer drivers. □

Remark 10 (The normalisation is what decouples the depth cascade)**.** The single point where the architecture enters is that the offset, though of unbounded form, is a bounded forcing *because its input is normalised*. If the gate read the raw state, the offset would satisfy only

$$\|o_t^{(\ell)}\| \lesssim \|W_{io}\|\,\|Z_t^{(\ell-1)}\|,$$

a linear growth in the previous layer. The moment bounds $M_\ell := \mathbb{E}\sup_t \|Z_t^{(\ell)}\|^q$ would then obey a recursion of the form

$$M_\ell \le K\big(1 + \|W_{io}\|^q M_{\ell-1}\big),$$

and the bound would grow geometrically in L unless $K\|W_{io}\|^q \le 1$. Uniformity in L is therefore a consequence of the normalisation, not of the algebraic form of the gate.

B.3 EXACTNESS OF THE PARALLEL ASSOCIATIVE SCAN

Proposition 11 (Scan exactness). *Let $\mathcal{M} \subseteq \mathbb{R}^{d\times d}$ be closed under multiplication with $I \in \mathcal{M}$ (the diagonal, block-diagonal or dense family), and let $\circ$ be the composition rule of Eq. 4.*

(a) $\circ$ is associative; the prefix products

$$(\widehat{F}_k, \widehat{g}_k) := (F_{k-1}, g_{k-1}) \circ \cdots \circ (F_0, g_0), \qquad (\widehat{F}_0, \widehat{g}_0) := (I, 0), \tag{42}$$

satisfy

$$Z_k = \widehat{F}_k Z_0 + \widehat{g}_k \qquad \textit{for the recursion} \qquad Z_{k+1} = F_k Z_k + g_k;$$

hence any associative-scan evaluation order returns exactly the sequential iterates, in $O(\log T)$ parallel depth, and $\widehat{F}_k \in \mathcal{M}$ whenever all $F_k \in \mathcal{M}$.

(b) Fix a realisation of the previous-layer path. The gated pairs

$$(\bar{F}_k^{(\ell)}, \bar{g}_k^{(\ell)}) = \big(\mathcal{D}(\alpha_k^{(\ell)}, F_k^{(\ell)}),\ g_k^{(\ell)} + o_k^{(\ell)}\big) \tag{43}$$

of Eq. 7 are then fixed affine data with $\bar{F}_k^{(\ell)} \in \mathcal{M}$, so (a) applies within each layer, and the depth-L stack is evaluated by L sequential scans in $O(L \log T)$ depth.

Proof. (a) To an affine pair associate the augmented matrix

$$\iota(F, g) := \begin{pmatrix} F & g \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{44}$$

The map ι is injective, and a direct multiplication gives

$$\iota(F', g')\,\iota(F, g) = \begin{pmatrix} F'F & F'g + g' \\ 0 & 1 \end{pmatrix} = \iota\big((F', g') \circ (F, g)\big). \tag{45}$$

Hence, for any three pairs,

$$\begin{aligned} \iota\Big(\big((F_c, g_c) \circ (F_b, g_b)\big) \circ (F_a, g_a)\Big) &= \iota(F_c, g_c)\,\iota(F_b, g_b)\,\iota(F_a, g_a) \\ &= \iota\Big((F_c, g_c) \circ \big((F_b, g_b) \circ (F_a, g_a)\big)\Big), \end{aligned}$$

and injectivity of ι gives associativity of $\circ$.

Next, write $\xi_k := (Z_k^\top, 1)^\top$. The recursion reads

$$\xi_{k+1} = \iota(F_k, g_k)\,\xi_k,$$

so by induction and Eq. 45

$$\xi_k = \iota(F_{k-1}, g_{k-1}) \cdots \iota(F_0, g_0)\,\xi_0 = \iota(\widehat{F}_k, \widehat{g}_k)\,\xi_0, \qquad \text{that is,} \qquad Z_k = \widehat{F}_k Z_0 + \widehat{g}_k.$$

Since every bracketing of an associative product agrees, any order of evaluation of the prefixes Eq. 42 returns the same pairs, and therefore the same iterates Z_k. The work-efficient parallel prefix algorithm (Blelloch, 1990) evaluates all of them along a balanced binary tree of depth logarithmic in the number of time steps, that is, $O(\log T)$ with the convention of Section 2. Finally

$$\widehat{F}_k = F_{k-1} \cdots F_0 \in \mathcal{M}$$

because $\mathcal{M}$ is closed under multiplication. The diagonal and block-diagonal implementations are structured representations of the same products: diagonal multiplication is elementwise and block-diagonal multiplication is independent multiplication inside each block.

(b) The gates $\alpha_k^{(\ell)}$ and $o_k^{(\ell)}$ depend only on the previous layer up to time t_k and on deterministic time features, never on the current state $Z_k^{(\ell)}$; conditioning on the previous path fixes them as deterministic data. Rescaling the diagonal entries of a matrix in the diagonal, block-diagonal or dense family leaves its sparsity pattern intact, so $\bar{F}_k^{(\ell)} \in \mathcal{M}$, and the offset only shifts $g_k^{(\ell)}$. Thus (a) applies inside layer ℓ. The layers run sequentially because the data of layer ℓ depend on $Z^{(\ell-1)}$, which gives L scans of depth $O(\log T)$ each. □

Remark 12. The gate makes the coefficients depend on the *other* layer, never on the current state, so affinity within a layer, and with it exactness of the scan, is untouched. This is the precise sense in which gated in-flow stacking preserves the parallel associative scan.

B.4 STRONG DISCRETISATION ERROR

Throughout this subsection the coefficient process ϑ of the layer under consideration is progressively measurable and satisfies the bounds Eq. 23; the time regularity Eq. 24 is invoked where it is needed. On the uniform grid

$$t_k = k\Delta t, \qquad k = 0, \dots, N, \qquad \Delta t = T/N \le 1,$$

write

$$\kappa(t) := \max\{t_k : t_k \le t\}, \qquad \Delta W_k := W_{t_{k+1}} - W_{t_k}.$$

The Euler–Maruyama step of Eq. 78 is the affine map

$$\begin{aligned} \bar{Z}_{k+1} &= \bar{Z}_k + \mu(t_k, \bar{Z}_k)\,\Delta t + \sigma(t_k, \bar{Z}_k)\,\Delta W_k = F_k \bar{Z}_k + g_k, \\ F_k = I + M_k, \qquad M_k &:= A_{t_k}\Delta t + \sum_{j=1}^{m} C^j_{t_k}\Delta W^j_k, \qquad g_k := b_{t_k}\Delta t + \sum_{j=1}^{m} d^j_{t_k}\Delta W^j_k, \end{aligned} \tag{46}$$

started at $\bar{Z}_0 = Z_0$. Its continuous interpolation is

$$\hat{Z}_t := \bar{Z}_k + \mu(t_k, \bar{Z}_k)\,(t - t_k) + \sigma(t_k, \bar{Z}_k)\,(W_t - W_{t_k}), \qquad t \in [t_k, t_{k+1}], \tag{47}$$

which satisfies $\hat{Z}_{t_k} = \bar{Z}_k$ and

$$\mathrm{d}\hat{Z}_t = \mu\big(\kappa(t), \hat{Z}_{\kappa(t)}\big)\,\mathrm{d}t + \sigma\big(\kappa(t), \hat{Z}_{\kappa(t)}\big)\,\mathrm{d}W_t. \tag{48}$$

All estimates hold for any correlation structure of the layer drivers, since no conditioning on other layers is used.

Lemma 13 (One-step conditional moments)**.** *For every $r \ge 1$ there is a constant C_r, depending only on $(a_*, b_*, c_*, d_*, m, r)$, such that for $\Delta t \le 1$ and every k,*

$$\begin{aligned} \big\|\mathbb{E}[M_k \mid \mathcal{F}_{t_k}]\big\| &\le a_*\Delta t, \qquad \big\|\mathbb{E}[g_k \mid \mathcal{F}_{t_k}]\big\| \le b_*\Delta t, \\ \mathbb{E}\big[\|M_k\|^r \,\big|\, \mathcal{F}_{t_k}\big] &\le C_r \Delta t^{r/2}, \qquad \mathbb{E}\big[\|g_k\|^r \,\big|\, \mathcal{F}_{t_k}\big] \le C_r \Delta t^{r/2}. \end{aligned} \tag{49}$$

Proof. ϑ_{t_k} is $\mathcal{F}_{t_k}$-measurable and bounded by Eq. 23, while ΔW_k is independent of $\mathcal{F}_{t_k}$ with

$$\mathbb{E}[\Delta W_k] = 0, \qquad \mathbb{E}|\Delta W^j_k|^r = m_r\,\Delta t^{r/2}.$$

The first two bounds follow. For the third,

$$\|M_k\| \le a_*\Delta t + c_* \sum_{j=1}^{m} |\Delta W^j_k|,$$

so, by

$$(x_0 + \cdots + x_m)^r \le (m+1)^{r-1}\big(x_0^r + \cdots + x_m^r\big) \qquad \text{and} \qquad \Delta t \le 1,$$

$$\mathbb{E}\big[\|M_k\|^r \,\big|\, \mathcal{F}_{t_k}\big] \le (m+1)^{r-1}\big(a_*^r \Delta t^r + m\, c_*^r\, m_r\, \Delta t^{r/2}\big) \le (m+1)^{r-1}\big(a_*^r + m\, c_*^r\, m_r\big)\Delta t^{r/2}.$$

The bound for g_k is identical with (b_*, d_*) in place of (a_*, c_*). □

Lemma 14 (Discrete stability)**.** *Let $q \ge 2$ be even and $\mathbb{E}\|Z_0\|^q < \infty$. There is $C_q = C_q(\Lambda', T, m, q)$ such that, for $\Delta t \le 1$,*

$$\sup_{0 \le k \le N} \mathbb{E}\|\bar{Z}_k\|^q \le C_q\big(1 + \mathbb{E}\|Z_0\|^q\big), \tag{50}$$

$$\sup_{t \le T} \mathbb{E}\|\hat{Z}_t\|^q \le C_q\big(1 + \mathbb{E}\|Z_0\|^q\big), \qquad \mathbb{E}\|\hat{Z}_t - \hat{Z}_{\kappa(t)}\|^q \le C_q\big(1 + \mathbb{E}\|Z_0\|^q\big)\,\Delta t^{q/2}, \qquad t \le T, \tag{51}$$

uniformly in N. Moreover $\mathbb{E}\sup_{t \le T}\|\hat{Z}_t\|^q < \infty$.

Proof. Step 1: proof of Eq. 50. Write $q = 2p$ with $p \in \mathbb{N}$ and

$$\xi_k := M_k \bar{Z}_k + g_k, \qquad \eta_k := 2\langle \bar{Z}_k, \xi_k \rangle + \|\xi_k\|^2,$$

so that $\bar{Z}_{k+1} = \bar{Z}_k + \xi_k$ and

$$\|\bar{Z}_{k+1}\|^2 = \|\bar{Z}_k\|^2 + \eta_k.$$

By the binomial theorem,

$$\|\bar{Z}_{k+1}\|^{2p} = \|\bar{Z}_k\|^{2p} + p\,\|\bar{Z}_k\|^{2p-2}\eta_k + \sum_{j=2}^{p} \binom{p}{j} \|\bar{Z}_k\|^{2(p-j)} \eta_k^j. \tag{52}$$

Since $\bar{Z}_k$ is $\mathcal{F}_{t_k}$-measurable, Lemma 13 gives, for every $r \geq 1$,

$$\mathbb{E}\big[\|\xi_k\|^r \mid \mathcal{F}_{t_k}\big] \leq 2^{r-1}\Big(\|\bar{Z}_k\|^r\,\mathbb{E}\big[\|M_k\|^r \mid \mathcal{F}_{t_k}\big] + \mathbb{E}\big[\|g_k\|^r \mid \mathcal{F}_{t_k}\big]\Big) \leq 2^{r-1} C_r \Delta t^{r/2} \big(1 + \|\bar{Z}_k\|\big)^r. \tag{53}$$

For the linear term of Eq. 52, Lemma 13 and Eq. 53 with $r = 2$ give

$$\begin{aligned}\mathbb{E}[\eta_k \mid \mathcal{F}_{t_k}] &= 2\big\langle \bar{Z}_k, \mathbb{E}[M_k \mid \mathcal{F}_{t_k}]\bar{Z}_k + \mathbb{E}[g_k \mid \mathcal{F}_{t_k}]\big\rangle + \mathbb{E}\big[\|\xi_k\|^2 \mid \mathcal{F}_{t_k}\big] \\ &\leq 2\Delta t\big(a_*\|\bar{Z}_k\|^2 + b_*\|\bar{Z}_k\|\big) + 2C_2\Delta t\big(1 + \|\bar{Z}_k\|\big)^2 \leq C\Delta t\big(1 + \|\bar{Z}_k\|\big)^2.\end{aligned}$$

For the higher terms,

$$|\eta_k| \leq 2\|\bar{Z}_k\|\,\|\xi_k\| + \|\xi_k\|^2,$$

so for $j \geq 2$, by Eq. 53 with $r = j$ and $r = 2j$,

$$\begin{aligned}\mathbb{E}\big[|\eta_k|^j \mid \mathcal{F}_{t_k}\big] &\leq 2^{j-1}\Big(2^j\|\bar{Z}_k\|^j\,\mathbb{E}\big[\|\xi_k\|^j \mid \mathcal{F}_{t_k}\big] + \mathbb{E}\big[\|\xi_k\|^{2j} \mid \mathcal{F}_{t_k}\big]\Big) \\ &\leq C\big(\Delta t^{j/2} + \Delta t^j\big)\big(1 + \|\bar{Z}_k\|\big)^{2j} \leq C\Delta t\big(1 + \|\bar{Z}_k\|\big)^{2j},\end{aligned}$$

because $\Delta t \leq 1$ and $j \geq 2$. Taking conditional expectations in Eq. 52 and using

$$\|\bar{Z}_k\|^{2(p-j)}\big(1 + \|\bar{Z}_k\|\big)^{2j} \leq \big(1 + \|\bar{Z}_k\|\big)^{2p} \leq 2^{2p-1}\big(1 + \|\bar{Z}_k\|^{2p}\big),$$

we obtain

$$\mathbb{E}\big[\|\bar{Z}_{k+1}\|^{2p} \mid \mathcal{F}_{t_k}\big] \leq \|\bar{Z}_k\|^{2p} + C\Delta t\big(1 + \|\bar{Z}_k\|^{2p}\big).$$

With $u_k := \mathbb{E}\|\bar{Z}_k\|^{2p}$ this reads

$$u_{k+1} \leq (1 + C\Delta t)\,u_k + C\Delta t,$$

and by induction

$$u_k \leq (1+C\Delta t)^k\,u_0 + C\Delta t\sum_{i=0}^{k-1}(1+C\Delta t)^i \leq e^{Ck\Delta t}\big(u_0 + Ck\Delta t\big) \leq e^{CT}\big(u_0 + CT\big), \qquad k \leq N,$$

which is Eq. 50.

Step 2: proof of Eq. 51. For $t \in [t_k, t_{k+1}]$, Eq. 47 gives

$$\hat{Z}_t - \bar{Z}_k = \mu(t_k, \bar{Z}_k)\,(t - t_k) + \sigma(t_k, \bar{Z}_k)\,(W_t - W_{t_k}).$$

Conditionally on $\mathcal{F}_{t_k}$ the second term is a centred Gaussian vector, and

$$\|\sigma(t_k, \bar{Z}_k)(W_t - W_{t_k})\| \leq \|\sigma(t_k, \bar{Z}_k)\|_{\mathrm{F}}\,\|W_t - W_{t_k}\|, \qquad \mathbb{E}\|W_t - W_{t_k}\|^q \leq m^{q/2} m_q\,\Delta t^{q/2}.$$

Hence, by Eq. 25,

$$\mathbb{E}\big[\|\hat{Z}_t - \bar{Z}_k\|^q \mid \mathcal{F}_{t_k}\big] \leq 2^{q-1}\Lambda'^{\,q}\big(1 + \|\bar{Z}_k\|\big)^q\big(\Delta t^q + m^{q/2} m_q \Delta t^{q/2}\big) \leq C\big(1 + \|\bar{Z}_k\|\big)^q \Delta t^{q/2}.$$

Taking expectations and using Eq. 50 gives the second bound in Eq. 51; the first follows from

$$\|\hat{Z}_t\|^q \leq 2^{q-1}\big(\|\bar{Z}_k\|^q + \|\hat{Z}_t - \bar{Z}_k\|^q\big).$$

Finally,

$$\sup_{t \leq T}\|\hat{Z}_t\| \leq \max_{k<N}\Big(\|\bar{Z}_k\| + \Lambda'\big(1 + \|\bar{Z}_k\|\big)\big(\Delta t + \sup_{t \in [t_k, t_{k+1}]}\|W_t - W_{t_k}\|\big)\Big)$$

is a maximum over finitely many random variables with finite q-th moments. Hence

$$\mathbb{E}\sup_{t \leq T}\|\hat{Z}_t\|^q < \infty.$$

□

Lemma 15 (Euler–Maruyama error for one layer)**.** *Let $q \geq 2$ be even, $\Delta t \leq 1$ and $\mathbb{E}\|Z_0\|^{2q} < \infty$. Let Z solve*

$$\mathrm{d}Z_t = \mu(t, Z_t)\ \mathrm{d}t + \sigma(t, Z_t)\ \mathrm{d}W_t$$

and let $\hat{Z}$ be the interpolated scheme Eq. 47, with the same Z_0 and the same driver. Then

$$\mathbb{E}\sup_{t\leq T}\|Z_t - \hat{Z}_t\|^q \leq C_q\Big(\Delta t^{q/2} + \sup_{s\leq T}\big(\mathbb{E}\|\vartheta_s - \vartheta_{\kappa(s)}\|^{2q}\big)^{1/2}\Big), \tag{54}$$

where C_q depends only on (Λ', T, m, q) and on $\mathbb{E}\|Z_0\|^{2q}$. If in addition ϑ satisfies the time regularity Eq. 24, then

$$\begin{aligned} &\mathbb{E}\sup_{t\leq T}\|Z_t - \hat{Z}_t\|^q \leq C_q\big(1 + L_{2q}^q\big)\Delta t^{q/2}, \\ &\textit{in particular} \qquad \max_{0\leq k\leq N}\mathbb{E}\|Z_{t_k} - \bar{Z}_k\|^q \leq C_q\big(1 + L_{2q}^q\big)\Delta t^{q/2}: \end{aligned} \tag{55}$$

the scheme has strong order $\frac{1}{2}$ in every L^q (Kloeden and Platen, 1992, Thm. 10.2.2).

Proof. Set $e_t := Z_t - \hat{Z}_t$, so that $e_0 = 0$ and, by Eq. 48,

$$e_t = \int_0^t \big[\mu(s, Z_s) - \mu(\kappa(s), \hat{Z}_{\kappa(s)})\big]\,\mathrm{d}s + \int_0^t \big[\sigma(s, Z_s) - \sigma(\kappa(s), \hat{Z}_{\kappa(s)})\big]\,\mathrm{d}W_s. \tag{56}$$

Step 1: decomposition of the coefficient differences. For the drift,

$$\mu(s, Z_s) - \mu(\kappa(s), \hat{Z}_{\kappa(s)}) = A_s e_s + \big(A_s - A_{\kappa(s)}\big)\hat{Z}_s + \big(b_s - b_{\kappa(s)}\big) + A_{\kappa(s)}\big(\hat{Z}_s - \hat{Z}_{\kappa(s)}\big),$$

and, for each j,

$$\sigma^j(s, Z_s) - \sigma^j(\kappa(s), \hat{Z}_{\kappa(s)}) = C_s^j e_s + \big(C_s^j - C_{\kappa(s)}^j\big)\hat{Z}_s + \big(d_s^j - d_{\kappa(s)}^j\big) + C_{\kappa(s)}^j\big(\hat{Z}_s - \hat{Z}_{\kappa(s)}\big).$$

Define

$$D_s := \|\vartheta_s - \vartheta_{\kappa(s)}\|\big(1 + \|\hat{Z}_s\|\big), \qquad I_s := \|\hat{Z}_s - \hat{Z}_{\kappa(s)}\|. \tag{57}$$

By Eq. 23 and the definition Eq. 16 of $\|\vartheta\|$,

$$\begin{aligned} \|\mu(s, Z_s) - \mu(\kappa(s), \hat{Z}_{\kappa(s)})\| &\leq a_*\|e_s\| + D_s + a_* I_s, \\ \|\sigma(s, Z_s) - \sigma(\kappa(s), \hat{Z}_{\kappa(s)})\|_{\mathrm{F}} &\leq m\, c_*\|e_s\| + D_s + m\, c_* I_s. \end{aligned} \tag{58}$$

Step 2: a Grönwall inequality. Let

$$\phi(t) := \mathbb{E}\sup_{r\leq t}\|e_r\|^q,$$

which is finite by Lemma 9 and Lemma 14. From Eq. 56, $\|x + y\|^q \leq 2^{q-1}(\|x\|^q + \|y\|^q)$, Jensen's inequality for the time integral, Eq. 18 with Jensen's inequality for the stochastic integral, and Eq. 58,

$$\begin{aligned} \phi(t) &\leq 2^{q-1}\Big[T^{q-1}\,\mathbb{E}\int_0^t \|\mu(s, Z_s) - \mu(\kappa(s), \hat{Z}_{\kappa(s)})\|^q\ \mathrm{d}s \\ &\qquad + c_q T^{q/2-1}\,\mathbb{E}\int_0^t \|\sigma(s, Z_s) - \sigma(\kappa(s), \hat{Z}_{\kappa(s)})\|_{\mathrm{F}}^q\ \mathrm{d}s\Big] \\ &\leq C_2\int_0^t \big(\mathbb{E}\|e_s\|^q + \mathbb{E}D_s^q + \mathbb{E}I_s^q\big)\,\mathrm{d}s\ \leq\ C_2\int_0^t \phi(s)\ \mathrm{d}s + C_2\int_0^T \big(\mathbb{E}D_s^q + \mathbb{E}I_s^q\big)\,\mathrm{d}s, \end{aligned} \tag{59}$$

with $C_2 = C_2(\Lambda', T, m, q)$. Grönwall's lemma gives

$$\phi(T) \leq C_2 e^{C_2 T}\int_0^T \big(\mathbb{E}D_s^q + \mathbb{E}I_s^q\big)\,\mathrm{d}s. \tag{60}$$

Step 3: the two forcing terms. By the Cauchy–Schwarz inequality and Eq. 51 at the exponent $2q$,

$$\begin{aligned} \mathbb{E}D_s^q &\leq \big(\mathbb{E}\|\vartheta_s - \vartheta_{\kappa(s)}\|^{2q}\big)^{1/2}\big(\mathbb{E}(1 + \|\hat{Z}_s\|)^{2q}\big)^{1/2} \\ &\leq C\big(1 + \mathbb{E}\|Z_0\|^{2q}\big)^{1/2}\sup_{s\leq T}\big(\mathbb{E}\|\vartheta_s - \vartheta_{\kappa(s)}\|^{2q}\big)^{1/2}, \end{aligned}$$

and by Eq. 51 at the exponent q,

$$\mathbb{E}I_s^q \le C\big(1+\mathbb{E}\|Z_0\|^q\big)\Delta t^{q/2}.$$

Inserting both bounds into Eq. 60 proves Eq. 54.

Step 4: the rate under time regularity. Under Eq. 24,

$$\mathbb{E}\|\vartheta_s-\vartheta_{\kappa(s)}\|^{2q} \le L_{2q}^{2q}\,|s-\kappa(s)|^q \le L_{2q}^{2q}\,\Delta t^q,$$

so the supremum in Eq. 54 is at most $L_{2q}^q\Delta t^{q/2}$, which gives Eq. 55. The grid bound follows because $\hat{Z}_{t_k}=\bar{Z}_k$. □

Theorem 16 (Stacked discretisation error)**.** *Let Assumptions 1–3 hold, assume $\mathbb{E}\|z_0\|^r<\infty$ for every $r\ge 2$, and apply the Euler–Maruyama scheme of Eq. 8 layerwise, the gates of the discrete layer ℓ being evaluated on the discrete previous layer at the grid points. Let $\hat{Z}^{(\ell)}$ denote the interpolation Eq. 47 of the discrete layer ℓ, so that $\hat{Z}^{(\ell)}_{t_k}=\bar{Z}^{(\ell)}_k$. Then for each fixed depth L and every even $q\ge 2$ there is a constant $C_{L,q}$, depending on the coefficient and gate bounds, on T, m, q, L and on the moments of z_0 but not on Δt, such that for $\Delta t\le 1$*

$$\begin{aligned}&\max_{1\le\ell\le L}\ \mathbb{E}\sup_{t\le T}\big\|Z_t^{(\ell)}-\hat{Z}_t^{(\ell)}\big\|^q \le C_{L,q}\,\Delta t^{q/2},\\ &\textit{in particular}\qquad \max_{1\le\ell\le L}\ \max_{0\le k\le N}\ \mathbb{E}\big\|Z^{(\ell)}_{t_k}-\bar{Z}^{(\ell)}_k\big\|^q \le C_{L,q}\,\Delta t^{q/2}.\end{aligned} \tag{61}$$

The scheme therefore has strong order $\frac{1}{2}$ for every fixed depth.

Proof. Fix L and an even $q\ge 2$, and set

$$q_\ell := q\,2^{\,L-\ell},\qquad \hat{e}_\ell := \mathbb{E}\sup_{t\le T}\big\|Z_t^{(\ell)}-\hat{Z}_t^{(\ell)}\big\|^{q_\ell},\qquad \ell=1,\dots,L, \tag{62}$$

so that $q_L=q$ and $q_{\ell-1}=2q_\ell$: higher moments are carried at the shallow layers because each inter-layer step consumes one Cauchy–Schwarz inequality. We prove by induction on ℓ that

$$\hat{e}_\ell \le C\,\Delta t^{q_\ell/2},\qquad \ell=1,\dots,L, \tag{63}$$

with constants that do not depend on Δt. Eq. 61 then follows: at $\ell=L$ directly, and at $\ell<L$ by Jensen's inequality,

$$\mathbb{E}\sup_{t\le T}\big\|Z_t^{(\ell)}-\hat{Z}_t^{(\ell)}\big\|^q \le \big(\hat{e}_\ell\big)^{q/q_\ell} \le C^{q/q_\ell}\,\Delta t^{q/2};$$

the grid bound follows from $\hat{Z}^{(\ell)}_{t_k}=\bar{Z}^{(\ell)}_k$.

Base case. Layer 1 is ungated: its coefficients are deterministic and Lipschitz in t, so Assumption 2 holds by construction, and Lemma 15 at the exponent q_1 (which requires $\mathbb{E}\|z_0\|^{2q_1}<\infty$) gives

$$\hat{e}_1 \le C\Delta t^{q_1/2}.$$

Induction step: the frozen coefficients. Let $\ell\ge 2$ and assume Eq. 63 for $\ell-1$. The step compares three processes: the exact layer $Z^{(\ell)}$, an intermediate SDE $\check{Z}$ whose gates are frozen on the grid and evaluated on the discrete previous layer, and the discrete layer $\hat{Z}^{(\ell)}$. The distance from $\check{Z}$ to $\hat{Z}^{(\ell)}$ is a one-layer Euler–Maruyama error (Lemma 15); the distance from $Z^{(\ell)}$ to $\check{Z}$ is a coefficient perturbation controlled by the induction hypothesis. Let ϑ denote the exact-gate coefficient process of layer ℓ, Eq. 32, with the gates evaluated on the exact previous layer $Z^{(\ell-1)}$, and let $\bar{\vartheta}$ denote the discrete-gate coefficients, with the gates evaluated on the discrete previous layer and frozen on each grid interval:

$$\begin{aligned}\bar{\vartheta}_{t_k} &:= \big(\alpha^{(\ell)}_{t_k}(\hat{Z}^{(\ell-1)})\odot A^{(\ell)}_{t_k},\ b^{(\ell)}_{t_k}+o^{(\ell)}_{t_k}(\hat{Z}^{(\ell-1)}),\ C^{(\ell),j}_{t_k},\ d^{(\ell),j}_{t_k}\big),\\ \bar{\vartheta}_t &:= \bar{\vartheta}_{\kappa(t)}.\end{aligned} \tag{64}$$

Because the gates of Eqs. 81 and 82 read only the current value of their input path, $\alpha_{t_k}^{(\ell)}(\hat{Z}^{(\ell-1)})$ and $o_{t_k}^{(\ell)}(\hat{Z}^{(\ell-1)})$ are the gates evaluated on

$$\hat{Z}_{t_k}^{(\ell-1)} = \bar{Z}_k^{(\ell-1)},$$

that is, exactly the gates used by the discrete layer ℓ. The process $\bar{\vartheta}$ is progressively measurable, and it satisfies the bounds Eq. 33 because Assumption 3(b)–(c) hold surely for every input path. Let $\check{Z}$ be the solution of the intermediate SDE with the frozen coefficients and the same driver and initial state,

$$\mathrm{d}\check{Z}_t = \bar{\mu}(t, \check{Z}_t)\ \mathrm{d}t + \bar{\sigma}(t, \check{Z}_t)\ \mathrm{d}W_t^{(\ell)}, \qquad \check{Z}_0 = \iota_d(z_0), \tag{65}$$

where $\bar{\mu}, \bar{\sigma}$ are the maps Eq. 17 built from $\bar{\vartheta}$; it exists, is unique and has finite moments of all orders by Lemma 9. The Euler–Maruyama scheme Eq. 46 for $\check{Z}$ uses the coefficients $\bar{\vartheta}_{t_k}$, which are exactly the coefficients of the discrete layer ℓ; hence its iterates are $\bar{Z}_k^{(\ell)}$ and its interpolation is $\hat{Z}^{(\ell)}$. We split

$$Z^{(\ell)} - \hat{Z}^{(\ell)} = \big(Z^{(\ell)} - \check{Z}\big) + \big(\check{Z} - \hat{Z}^{(\ell)}\big). \tag{66}$$

Induction step (i): $\check{Z}$ versus $\hat{Z}^{(\ell)}$. Lemma 15 applied to $\check{Z}$ at the exponent q_ℓ, whose hypotheses hold with $\Lambda'_\star$ in place of Λ', gives Eq. 54. Since $\bar{\vartheta}_s = \bar{\vartheta}_{\kappa(s)}$ for every s, the supremum in Eq. 54 vanishes, and

$$\mathbb{E}\sup_{t\le T}\big\|\check{Z}_t - \hat{Z}_t^{(\ell)}\big\|^{q_\ell} \le C\Delta t^{q_\ell/2}. \tag{67}$$

Induction step (ii): $Z^{(\ell)}$ versus $\check{Z}$. Both are continuous SDEs with the same driver and initial state. Set $e_t := Z_t^{(\ell)} - \check{Z}_t$; then $e_0 = 0$ and

$$e_t = \int_0^t \big[\mu(s, Z_s^{(\ell)}) - \bar{\mu}(s, \check{Z}_s)\big]\,\mathrm{d}s + \int_0^t \big[\sigma(s, Z_s^{(\ell)}) - \bar{\sigma}(s, \check{Z}_s)\big]\,\mathrm{d}W_s^{(\ell)},$$

where μ, σ are built from ϑ. Writing $\vartheta_s = (A_s, b_s, \dots)$ and $\bar{\vartheta}_s = (\bar{A}_s, \bar{b}_s, \dots)$,

$$\mu(s, Z_s^{(\ell)}) - \bar{\mu}(s, \check{Z}_s) = A_s e_s + (A_s - \bar{A}_s)\check{Z}_s + (b_s - \bar{b}_s),$$

so that

$$\|\mu(s, Z_s^{(\ell)}) - \bar{\mu}(s, \check{Z}_s)\| \le (1+\varepsilon)a_*\|e_s\| + \|\vartheta_s - \bar{\vartheta}_s\|\big(1 + \|\check{Z}_s\|\big),$$
$$\|\sigma(s, Z_s^{(\ell)}) - \bar{\sigma}(s, \check{Z}_s)\|_{\mathrm{F}} \le m\, c_*\|e_s\| + \|\vartheta_s - \bar{\vartheta}_s\|\big(1 + \|\check{Z}_s\|\big).$$

The argument of Step 2 of the proof of Lemma 15, with the forcing $\|\vartheta_s - \bar{\vartheta}_s\|(1 + \|\check{Z}_s\|)$ in place of $D_s + \Lambda' I_s$, gives

$$\mathbb{E}\sup_{t\le T}\|e_t\|^{q_\ell} \le C\int_0^T \mathbb{E}\Big[\|\vartheta_s - \bar{\vartheta}_s\|^{q_\ell}\big(1 + \|\check{Z}_s\|\big)^{q_\ell}\Big]\,\mathrm{d}s \le C\sup_{s\le T}\big(\mathbb{E}\|\vartheta_s - \bar{\vartheta}_s\|^{2q_\ell}\big)^{1/2}, \tag{68}$$

where the last step is the Cauchy–Schwarz inequality together with the moment bound Eq. 38 for $\check{Z}$ at the exponent $2q_\ell$. The coefficient difference splits into a time-regularity part and a gate-perturbation part,

$$\|\vartheta_s - \bar{\vartheta}_s\| \le \|\vartheta_s - \vartheta_{\kappa(s)}\| + \|\vartheta_{\kappa(s)} - \bar{\vartheta}_{\kappa(s)}\|. \tag{69}$$

For the first part, the exact previous layer satisfies Eq. 31 for every moment order by Theorem 4, so Lemma 8 gives Eq. 24 for ϑ at the exponent $2q_\ell$, and

$$\big(\mathbb{E}\|\vartheta_s - \vartheta_{\kappa(s)}\|^{2q_\ell}\big)^{1/2} \le L_{2q_\ell}^{q_\ell}\,|s - \kappa(s)|^{q_\ell/2} \le L_{2q_\ell}^{q_\ell}\,\Delta t^{q_\ell/2}. \tag{70}$$

For the second part, at the grid time $t_k = \kappa(s)$ the two coefficient tuples differ only through the gates, evaluated on the paths $Z^{(\ell-1)}$ and $\hat{Z}^{(\ell-1)}$ up to time t_k. By Eq. 29 and the argument of Eq. 36,

$$\|\vartheta_{t_k} - \bar{\vartheta}_{t_k}\| \le a_*\big\|\alpha_{t_k}^{(\ell)}(Z^{(\ell-1)}) - \alpha_{t_k}^{(\ell)}(\hat{Z}^{(\ell-1)})\big\|_\infty + \big\|o_{t_k}^{(\ell)}(Z^{(\ell-1)}) - o_{t_k}^{(\ell)}(\hat{Z}^{(\ell-1)})\big\|$$
$$\le (1 + a_*)L_g \sup_{r\le T}\big\|Z_r^{(\ell-1)} - \hat{Z}_r^{(\ell-1)}\big\|,$$

so that, by the induction hypothesis and $2q_\ell = q_{\ell-1}$,

$$\begin{aligned}\big(\mathbb{E}\|\vartheta_{\kappa(s)} - \bar{\vartheta}_{\kappa(s)}\|^{2q_\ell}\big)^{1/2} &\le \big((1+a_*)L_g\big)^{q_\ell}\,\big(\hat{e}_{\ell-1}\big)^{1/2} \\ &\le \big((1+a_*)L_g\big)^{q_\ell}\,C^{1/2}\,\Delta t^{q_{\ell-1}/4} = C\,\Delta t^{q_\ell/2}.\end{aligned} \tag{71}$$

Inserting Eqs. 70 and 71 into Eq. 68 through Eq. 69 and $(x+y)^{2q_\ell} \le 2^{2q_\ell-1}(x^{2q_\ell}+y^{2q_\ell})$,

$$\mathbb{E}\sup_{t\le T}\big\|Z_t^{(\ell)} - \check{Z}_t\big\|^{q_\ell} \le C\Delta t^{q_\ell/2}. \tag{72}$$

Conclusion. By Eq. 66 and $\|x+y\|^{q_\ell} \le 2^{q_\ell-1}(\|x\|^{q_\ell}+\|y\|^{q_\ell})$,

$$\hat{e}_\ell \le 2^{q_\ell-1}\Big(\mathbb{E}\sup_{t\le T}\big\|Z_t^{(\ell)} - \check{Z}_t\big\|^{q_\ell} + \mathbb{E}\sup_{t\le T}\big\|\check{Z}_t - \hat{Z}_t^{(\ell)}\big\|^{q_\ell}\Big) \le C\Delta t^{q_\ell/2}$$

by Eqs. 67 and 72, which is Eq. 63 for ℓ and closes the induction. □

Remark 17 (Dependence on the depth)**.** The constant $C_{L,q}$ depends on L in two ways: the induction carries the moment order $q\,2^{L-\ell}$ at layer ℓ, so the constants of Lemmas 14 and 15 enter at exponents that grow with the depth, and each inter-layer step multiplies the propagated error by a factor involving the gate Lipschitz constant L_g through Eq. 71. The theorem is therefore a statement at fixed depth: the rate $\frac{1}{2}$ holds for every L, but no uniformity of the constant in L is claimed. For the depths used in practice ($L = 2, 3$) the exponents involved are $4q$ at most.

Remark 18 (Strong order $\frac{1}{2}$, not 1)**.** Order 1 would require the iterated Itô integrals

$$\int_{t_k}^{t_{k+1}}\int_{t_k}^{s} \mathrm{d}W_u^i\,\mathrm{d}W_s^j, \qquad i \ne j,$$

and their Lévy areas, which the scheme does not simulate. We therefore claim only order $\frac{1}{2}$, without assuming that the C_t^j commute; if they do, the exponential transition of Appendix C.1 attains order 1 and, in the diagonal case, is an exact one-step solution operator.

B.5 Proof of Theorem 5

Proof. Throughout, ϵ is the m-dimensional Brownian innovation of Eq. 10, independent of the prefix driver W^{prefix} and of $\mathcal{F}_0$, and $u_{\theta,t}$ is progressively measurable with respect to the final-layer filtration and satisfies Novikov's condition Eq. 13. The stochastic exponential of Eq. 14 is taken with respect to ϵ,

$$\mathcal{E}_t(u_\theta) := \exp\Big(\int_0^t u_{\theta,s}^\top\,\mathrm{d}\epsilon_s - \tfrac{1}{2}\int_0^t \|u_{\theta,s}\|^2\,\mathrm{d}s\Big), \qquad t \in [0,T]. \tag{73}$$

Step 0: the initial-state shift. Under $\mathbb{P}$, $z_0 \sim \mathcal{N}(0, I_d)$ is independent of $(W^{\mathrm{prefix}}, \epsilon)$. The factor

$$\exp\big(\mu_\theta^\top z_0 - \tfrac{1}{2}\|\mu_\theta\|^2\big) = \frac{\mathrm{d}\mathcal{N}(\mu_\theta, I_d)}{\mathrm{d}\mathcal{N}(0, I_d)}(z_0)$$

is $\mathcal{F}_0$-measurable and has $\mathbb{P}$-expectation one. Being independent of $\mathcal{E}_T(u_\theta)$ under $\mathbb{P}$, its product with $\mathcal{E}_T(u_\theta)$ is again a probability density, under $\mathbb{Q}$ the initial state has law $\mathcal{N}(\mu_\theta, I_d)$, and Steps 1–3 below, which argue conditionally on $\mathcal{F}_0$, apply verbatim. In the implementation

$$z_0 = \mu_\theta + \xi, \qquad \xi \sim \mathcal{N}(0, I_d),$$

and the term $\mu_\theta^\top z_0 - \frac{1}{2}\|\mu_\theta\|^2$ is added to the discrete log-ratio Eq. 76; its expectation under $\mathbb{Q}$,

$$\mathbb{E}^{\mathbb{Q}}\big[\mu_\theta^\top z_0 - \tfrac{1}{2}\|\mu_\theta\|^2\big] = \tfrac{1}{2}\|\mu_\theta\|^2,$$

is the initial-state share of $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})$. When $\mu_\theta = 0$ nothing changes.

Step 1: $\mathbb{Q}$ is a probability measure and $\epsilon^{\mathbb{Q}}$ is a $\mathbb{Q}$-Brownian motion. Under Novikov's condition, $(\mathcal{E}_t(u_\theta))_{t\le T}$ is a strictly positive martingale (Protter, 2005, Ch. III), so

$$\mathbb{E}^{\mathbb{P}}\big[\mathcal{E}_T(u_\theta)\big] = 1.$$

Hence

$$\mathbb{Q}(A) := \mathbb{E}^{\mathbb{P}}\big[\mathcal{E}_T(u_\theta)\mathbf{1}_A\big], \qquad A \in \mathcal{F}_T,$$

is a probability measure equivalent to $\mathbb{P}$, with $\mathrm{d}\mathbb{Q}/\,\mathrm{d}\mathbb{P} = \mathcal{E}_T(u_\theta)$. By Girsanov's theorem applied to the $2m$-dimensional Brownian motion $(W^{\mathrm{prefix}}, \epsilon)$, whose density process involves only the components of ϵ, the process

$$\Big(W_t^{\mathrm{prefix}},\ \epsilon_t^{\mathbb{Q}}\Big)_{t\le T}, \qquad \epsilon_t^{\mathbb{Q}} := \epsilon_t - \int_0^t u_{\theta,s}\ \mathrm{d}s, \tag{74}$$

is a $2m$-dimensional Brownian motion under $\mathbb{Q}$: the innovation is shifted by the control, while the prefix driver keeps its law and remains independent of the shifted innovation.

Step 2: the change-of-measure identity. For every $\mathcal{F}_T$-measurable functional Φ with $\mathbb{E}^{\mathbb{P}}|\Phi| < \infty$, the definition of $\mathbb{Q}$ gives

$$\mathbb{E}^{\mathbb{P}}[\Phi] = \mathbb{E}^{\mathbb{Q}}\Big[\Phi\,\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big], \qquad \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \mathcal{E}_T(u_\theta)^{-1}. \tag{75}$$

Substituting $\mathrm{d}\epsilon_s = \mathrm{d}\epsilon_s^{\mathbb{Q}} + u_{\theta,s}\,\mathrm{d}s$ in Eq. 73,

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \exp\Big(-\int_0^T u_{\theta,s}^\top\,\mathrm{d}\epsilon_s^{\mathbb{Q}} - \tfrac{1}{2}\int_0^T \|u_{\theta,s}\|^2\,\mathrm{d}s\Big).$$

In particular

$$\mathbb{E}^{\mathbb{Q}}\Big[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] = \mathbb{P}(\Omega) = 1:$$

the importance weights are exactly normalised.

Step 3: the discrete implementation is exact. In the implementation the innovation is simulated on the grid. Under $\mathbb{P}$ the increments

$$\Delta\epsilon_k^{\mathbb{P}}, \qquad k = 0, \dots, N-1,$$

are independent $\mathcal{N}(0, \Delta t I_m)$ vectors, and the control $u_{\theta,k}$ is $\mathcal{F}_{t_k}$-measurable, being a function of $\{\Delta\epsilon_j^{\mathbb{P}}\}_{j<k}$ and of prefix features up to t_k. Simulation under $\mathbb{Q}$ draws $\Delta\epsilon_k^{\mathbb{Q}} \sim \mathcal{N}(0, \Delta t I_m)$ and sets

$$\Delta\epsilon_k^{\mathbb{P}} = \Delta\epsilon_k^{\mathbb{Q}} + u_{\theta,k}\Delta t,$$

which is Eq. 11. Conditionally on $\mathcal{F}_{t_k}$, the $\mathbb{P}$-law of $\Delta\epsilon_k^{\mathbb{P}}$ has the density

$$\varphi_{\Delta t}(x) \propto \exp\Big(-\frac{\|x\|^2}{2\Delta t}\Big),$$

and its $\mathbb{Q}$-law has the density $\varphi_{\Delta t}(x - u_{\theta,k}\Delta t)$. Their ratio is

$$\frac{\varphi_{\Delta t}(x - u_{\theta,k}\Delta t)}{\varphi_{\Delta t}(x)} = \exp\Big(u_{\theta,k}^\top x - \tfrac{1}{2}\|u_{\theta,k}\|^2\Delta t\Big).$$

Multiplying the conditional density ratios over k gives the likelihood ratio of the two laws of the whole increment sequence,

$$\begin{aligned}
\log\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} &= \sum_{k=0}^{N-1}\Big(u_{\theta,k}^\top\Delta\epsilon_k^{\mathbb{P}} - \tfrac{1}{2}\|u_{\theta,k}\|^2\Delta t\Big),\\
\log\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} &= -\sum_{k=0}^{N-1} u_{\theta,k}^\top\Delta\epsilon_k^{\mathbb{Q}} - \tfrac{1}{2}\sum_{k=0}^{N-1}\|u_{\theta,k}\|^2\Delta t,
\end{aligned} \tag{76}$$

which is Eq. 12: on the grid the weight is an exact density ratio, not an approximation of Eq. 75. It is also exactly normalised, because for a fixed $\mathcal{F}_{t_k}$-measurable $u_{\theta,k}$ and an independent Gaussian increment

$$\mathbb{E}^{\mathbb{P}}\Big[\exp\big(u_{\theta,k}^\top\Delta\epsilon_k^{\mathbb{P}} - \tfrac{1}{2}\|u_{\theta,k}\|^2\Delta t\big)\ \Big|\ \mathcal{F}_{t_k}\Big] = 1,$$

and the tower property gives

$$\mathbb{E}^{\mathbb{P}}\Big[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\Big] = 1$$

without any integrability condition on the control beyond measurability.

Step 4: locality of the ratio. The prefix layers are functionals of z_0 and of the prefix driver. By Steps 0–1 the law of W^{prefix} is the same under $\mathbb{P}$ and $\mathbb{Q}$, the law of z_0 changes only through the shift μ_θ, and both are independent of ϵ: the prefix computation is the same map under both measures, evaluated at a shifted starting point. The final layer is a functional of the prefix path and of ϵ, and Eq. 75 together with Step 0 shows that the likelihood ratio is a functional of the control, of z_0 and of the shifted innovation only. Hence the ratio depends only on the final-layer control, the initial-state shift and on the Brownian innovation being shifted. □

B.6 Proof of Theorem 6

Proof. Let $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ and let $\eta > 0$.

Step 1: construction of an atomless scalar feature. Choose a latent width d, to be specified below, and configure the first layer as

$$\mathrm{d}Z_t^{(1)} = \beta e_1 \, \mathrm{d}W_t^1, \qquad \beta \neq 0,$$

where e_1 is the first coordinate vector. Thus

$$Z_t^{(1)} = z_0 + \beta e_1 W_t^1.$$

In particular,

$$S_T := \langle e_1, Z_T^{(1)} \rangle = \langle e_1, z_0 \rangle + \beta W_T^1.$$

Since W_T^1 is independent of $\mathcal{F}_0$, the law of S_T is the convolution of the law of $\langle e_1, z_0 \rangle$ with a nondegenerate Gaussian law. Consequently, S_T has a smooth density, irrespective of whether the law of z_0 is degenerate.

Write

$$\mathrm{RMSNorm}_\rho(z) := \Gamma \frac{z}{\sqrt{\rho + d^{-1}|z|^2}}, \qquad \rho > 0,$$

and choose the first gain γ_1 to be 1. Set

$$R_t := \left[\mathrm{RMSNorm}_\rho\big(Z_t^{(1)}\big) \right]_1 .$$

Conditionally on z_0, R_T is a strictly monotone function of W_T^1. Indeed, for fixed z_0, the map

$$u \longmapsto \gamma_1 \frac{u}{\sqrt{\rho + d^{-1}\left(u^2 + \sum_{j=2}^d z_{0,j}^2\right)}}$$

has derivative

$$\gamma_1 \frac{\rho + d^{-1}\sum_{j=2}^d z_{0,j}^2}{\left[\rho + d^{-1}\left(u^2 + \sum_{j=2}^d z_{0,j}^2\right)\right]^{3/2}},$$

which never vanishes. Hence R_T is atomless. Moreover, R has continuous paths and is bounded.

Step 2: reduction to a finitely supported target law. Finitely supported probability measures are dense in $\mathcal{P}_2(\mathbb{R}^p)$. We may therefore choose

$$\nu = \sum_{i=1}^N p_i \delta_{y_i}, \qquad p_i > 0, \qquad \sum_{i=1}^N p_i = 1,$$

such that

$$W_2(\mu, \nu) < \frac{\eta}{3}.$$

Let F_R be the distribution function of R_T, and define

$$P_i := \sum_{j=1}^i p_j, \qquad i = 1, \ldots, N-1.$$

Since R_T is atomless, F_R is continuous. We may therefore choose

$$q_1 < \cdots < q_{N-1}$$

such that

$$F_R(q_i) = P_i.$$

With $q_0 = -\infty$ and $q_N = +\infty$, define

$$H(r) := y_i \quad \text{whenever} \quad q_{i-1} < r \le q_i.$$

It follows that

$$\mathcal{L}\big(H(R_T)\big) = \nu.$$

Equivalently,

$$H(r) = y_1 + \sum_{i=1}^{N-1} (y_{i+1} - y_i)\mathbf{1}_{\{r > q_i\}}.$$

For $a > 0$, define the sigmoid approximation

$$H_a(r) := y_1 + \sum_{i=1}^{N-1} (y_{i+1} - y_i)\sigma\big(a(r - q_i)\big).$$

Since

$$\sigma\big(a(r - q_i)\big) \longrightarrow \mathbf{1}_{\{r > q_i\}}$$

for every $r \neq q_i$, and $\mathbb{P}(R_T = q_i) = 0$, we have

$$H_a(R_T) \longrightarrow H(R_T) \qquad \text{almost surely.}$$

The family $\{H_a(R_T) : a > 0\}$ is uniformly bounded. Hence, by dominated convergence,

$$\|H_a(R_T) - H(R_T)\|_{L^2} \longrightarrow 0.$$

Choose $a > 0$ such that

$$\|H_a(R_T) - H(R_T)\|_{L^2} < \frac{\eta}{3}.$$

Step 3: realisation by the gated layer. Choose $d \ge N$. Disable the scale gate by taking $W_{iF}^{(2)} = 0$, together with a zero bias in that branch, so that

$$\alpha_t^{(2)} = \mathbf{1},$$

and set all diffusion coefficients of the second layer equal to zero. Choose its diagonal drift matrix to be

$$A^{(2)} = -\lambda I_d, \qquad \lambda > 0.$$

Use one coordinate $V^{0,\lambda}$ to generate the constant feature and $N - 1$ coordinates $V^{i,a,\lambda}$ to generate the shifted sigmoid features:

$$\mathrm{d}V_t^{0,\lambda} = -\lambda V_t^{0,\lambda}\,\mathrm{d}t + \lambda\frac{t}{T}\,\mathrm{d}t,$$
$$\mathrm{d}V_t^{i,a,\lambda} = -\lambda V_t^{i,a,\lambda}\,\mathrm{d}t + \lambda\frac{t}{T}\sigma\left(aR_t - aq_i\frac{t}{T}\right)\mathrm{d}t, \qquad i = 1, \ldots, N-1.$$

These dynamics are contained in the stated offset-gate parameterisation. Indeed, because the gate input contains both R_t and t, suitable rows of the gate matrices give

$$W_{go,i}^{(2)} x_t^{(1)} = aR_t - aq_i\frac{t}{T}, \qquad W_{io,i}^{(2)} x_t^{(1)} = \lambda\frac{t}{T}.$$

For the constant coordinate, take

$$W_{go,0}^{(2)} x_t^{(1)} = 0, \qquad W_{io,0}^{(2)} x_t^{(1)} = 2\lambda\frac{t}{T},$$

and use $\sigma(0) = 1/2$.

We claim that, as $\lambda \to \infty$,

$$V_T^{0,\lambda} \longrightarrow 1,$$
$$V_T^{i,a,\lambda} \longrightarrow \sigma\big(a(R_T - q_i)\big), \qquad i = 1, \dots, N-1,$$

in L^2. To see this, let ϕ be any bounded process which is continuous in L^2 at T. The solution of

$$\mathrm{d}U_t^\lambda = -\lambda U_t^\lambda \, \mathrm{d}t + \lambda \phi_t \, \mathrm{d}t$$

satisfies

$$U_T^\lambda = e^{-\lambda T} U_0^\lambda + \int_0^T \lambda e^{-\lambda(T-s)} \phi_s \, \mathrm{d}s.$$

Consequently,

$$\begin{aligned} \big\| U_T^\lambda - \phi_T \big\|_{L^2} \le e^{-\lambda T} \left(\big\| U_0^\lambda \big\|_{L^2} + \|\phi_T\|_{L^2} \right) \\ + \int_0^T \lambda e^{-\lambda(T-s)} \, \|\phi_s - \phi_T\|_{L^2} \, \mathrm{d}s, \end{aligned}$$

which converges to zero. Applying this observation with

$$\phi_t = \frac{t}{T}$$

and

$$\phi_t = \frac{t}{T} \sigma \left(a R_t - a q_i \frac{t}{T} \right)$$

proves the claim.

Choose the terminal readout so that

$$Y_T^{a,\lambda} = y_1 V_T^{0,\lambda} + \sum_{i=1}^{N-1} (y_{i+1} - y_i) V_T^{i,a,\lambda}.$$

Then

$$Y_T^{a,\lambda} \longrightarrow H_a(R_T) \qquad \text{in } L^2$$

as $\lambda \to \infty$. Choose λ sufficiently large that

$$\left\| Y_T^{a,\lambda} - H_a(R_T) \right\|_{L^2} < \frac{\eta}{3}.$$

Finally, using the coupling inequality

$$W_2\big(\mathcal{L}(X), \mathcal{L}(X')\big) \le \|X - X'\|_{L^2},$$

we obtain

$$\begin{aligned} W_2\big(\mathcal{L}(Y_T^{a,\lambda}), \mu\big) \le \left\| Y_T^{a,\lambda} - H_a(R_T) \right\|_{L^2} \\ + \|H_a(R_T) - H(R_T)\|_{L^2} + W_2(\nu, \mu) < \eta. \end{aligned}$$

Since μ and η were arbitrary, the result follows.

Step 4: discretised implementation. Once a and λ have been fixed, the configuration of Steps 1–3 is a fixed two-layer model, and its discretisation on a grid π converges. The first layer has constant coefficients and additive noise only, so its Euler–Maruyama iterates are exact,

$$\bar{Z}_k^{(1)} = \iota_d(z_0) + \beta e_1 W_{t_k}^1,$$

and the discrete gate of the second layer therefore reads the exact feature R_{t_k} at the grid points. The second layer has no diffusion, so, pathwise, it is the linear ordinary differential equation

$$\mathrm{d}U_t^\lambda = -\lambda U_t^\lambda \, \mathrm{d}t + \lambda \phi_t \, \mathrm{d}t$$

of Step 3, with a forcing ϕ that is bounded and continuous in t. Its discretisation is the explicit Euler scheme, which converges to U^λ uniformly on $[0,T]$ along every path as $|\pi| \to 0$, and which is bounded by $\|z_0\| + \lambda T \sup_t |\phi_t|$ once $|\pi| \le 1/\lambda$. By dominated convergence, the discretised terminal output $Y_T^{a,\lambda,\pi}$ satisfies

$$\left\| Y_T^{a,\lambda,\pi} - Y_T^{a,\lambda} \right\|_{L^2} \longrightarrow 0 \qquad \text{as} \qquad |\pi| \longrightarrow 0.$$

When z_0 has finite moments of all orders this is also a special case of Theorem 16. Then the same density conclusion holds for the discretised architecture when the mesh is allowed to vary. Indeed,

$$W_2\left(\mathcal{L}(Y_T^{a,\lambda,\pi}), \mathcal{L}(Y_T^{a,\lambda})\right) \le \left\| Y_T^{a,\lambda,\pi} - Y_T^{a,\lambda} \right\|_{L^2},$$

and the discretisation is refined only after the finite parameters a and λ have been selected. □

C Implementation Details

This appendix expands the implementation choices summarised in the main text. All reported SLiSDE results use the gated in-flow stack (Section 3.3) with structured block-diagonal base layers, an always-on RMS normalisation of the gate input, optional causal convolutions, and time-feature augmentation; residual stacking is not used. The Girsanov tilt described in Section 4 is applied only to the final layer, with the correlated-driver construction of Eq. 10.

C.1 Discretisation of a linear layer

On the interval $[t_k, t_{k+1}]$, the affine linear SDE Eq. 2 admits the exponential approximation

$$Z_{k+1} = F_k Z_k + g_k, \qquad g_k = b\Delta t + D\Delta W_k,$$
$$F_k = \exp\left(\left(A - \frac{1}{2}\sum_{j=1}^{m}(C^j)^2\right)\Delta t + \sum_{j=1}^{m} C^j\,\Delta W_k^j\right). \tag{77}$$

The correction is the Itô correction. If $A, C^1, \ldots, C^m$ commute pairwise, then F_k is the exact flow of the homogeneous Itô equation; otherwise it is an exponential approximation. The affine term g_k remains a first-order approximation.

To avoid computing matrix exponentials, we use the Euler–Maruyama transition

$$Z_{k+1} = \left(I + A\Delta t + \sum_{j=1}^{m} C^j\,\Delta W_k^j\right) Z_k + b\Delta t + D\Delta W_k, \tag{78}$$

which preserves the affine recurrence structure and is used throughout our experiments. With the time-dependent coefficients of Appendix C.2, the static coefficients are replaced by their values at t_k in both transitions; the strong error of the Euler–Maruyama scheme is analysed in Appendix B.4.

C.2 Time-dependent coefficient map

Each base layer (Section 3.1) stores static structured coefficients $(A, b, C^1, \ldots, C^m, d^1, \ldots, d^m)$ and can be made time-dependent through a small deterministic decoder. For each grid time t_k we form

$$\tau(t_k) = \big(1,\, t_k,\, t_k^2,\, \sin(\omega t_k),\, \cos(\omega t_k)\big), \qquad h(t_k) = \tanh\big(U\tau(t_k) + c\big),$$

and add structure-aware corrections to the free coefficient entries:

$$A_{t_k} = A + \Delta A\big(h(t_k)\big), \quad b_{t_k} = b + \Delta b\big(h(t_k)\big), \quad C^j_{t_k} = C^j + \Delta C^j\big(h(t_k)\big).$$

For additive diffusion we use a per-Brownian-channel time gate on the dense diffusion matrix, equivalently producing $d^j_{t_k}$ from d^j and $h(t_k)$. The decoders respect the chosen matrix structure: diagonal layers decode only diagonal entries, block-diagonal layers decode only block entries, and dense layers decode full matrices. Decoders are zero-initialised, so the model starts from the time-homogeneous structured SDE and learns time variation only when useful.

C.3 Feature conditioning in gated layers

The gated layer consumes the previous layer path at the same time grid as the current layer. The simplest gate input is $\mathrm{Norm}(Z_k^{(\ell-1)})$, where $Z_k^{(\ell-1)}$ is the model output after $\ell - 1$ layers and the normalisation is RMSNorm or LayerNorm (RMSNorm by default, applied at every gated layer). Note that normalisation controls the *scale* of the gate input but does not by itself bound the paths; the offset gate is nevertheless bounded for every fixed choice of its weights, which is what the analysis uses (Appendix B.1). In the implementation this signal can be enriched in two ways.

Causal convolution (for gated-in-flow stacking only). Before normalisation, the previous-layer path can be passed through a left-padded depthwise causal convolution along the time axis. The value at time t_k depends only on previous-layer states up to t_k, so the gate is adapted and does not leak future information. The convolution is wrapped in a residual update and its kernel is small at initialisation, making the initial behaviour close to the no-convolution case while still allowing the gate to learn short temporal summaries.

Time features. The deterministic vector

$$\tau_{\mathrm{gate}}(t) = (t, t^2, \sin(\omega t), \cos(\omega t)) \tag{79}$$

can be concatenated to the gate input after convolution and after normalisation. Appending time features after normalisation preserves their scale, while keeping the convolution focused on the stochastic path rather than on deterministic batch-constant signals. With both refinements, the gate input is

$$x_k = \big[\mathrm{Norm}\left(\mathrm{CausalConv}(H^{(\ell-1)})_k\right), t_k, t_k^2, \sin(\omega t_k), \cos(\omega t_k)\big]. \tag{80}$$

C.4 Gated in-flow implementation

The paper focuses on the gated in-flow stacking. Given the conditioned feature $x_k^{(\ell-1)}$ from Eq. 80, the layer constructs a bounded diagonal scale

$$\alpha_k^{(\ell)} = \mathbf{1} + \varepsilon\, \sigma\big(W_{gF}^{(\ell)} x_k^{(\ell-1)}\big) \odot \tanh\big(W_{iF}^{(\ell)} x_k^{(\ell-1)}\big), \tag{81}$$

and an affine residual offset

$$o_k^{(\ell)} = \sigma\big(W_{go}^{(\ell)} x_k^{(\ell-1)}\big) \odot \big(W_{io}^{(\ell)} x_k^{(\ell-1)} + c_{io}^{(\ell)}\big). \tag{82}$$

All gate maps carry trainable biases, omitted from the displays except $c_{io}^{(\ell)}$. The $\tanh$ bound gives $\alpha_k^{(\ell)} \in (\mathbf{1} - \varepsilon, \mathbf{1} + \varepsilon)$ coordinatewise. The scale $\alpha_k^{(\ell)}$ is applied only to the diagonal entries of the structured drift matrix $F_k^{(\ell)}$, and the offset $o_k^{(\ell)}$ is added to the affine intercept $g_k^{(\ell)}$. Gate kernels are initialised at small scale, so initially $\alpha_k^{(\ell)} \approx \mathbf{1}$ and $o_k^{(\ell)} \approx \mathbf{0}$: the stacked model starts close to an unmodulated structured SDE and learns cross-layer coupling gradually. We fix $\varepsilon = 0.1$ throughout.

The offset $o_k^{(\ell)}$ is an *unbounded* GLU: the gate $\sigma(W_{go}^{(\ell)} x_k^{(\ell-1)})$ is bounded, while the value branch $W_{io}^{(\ell)} x_k^{(\ell-1)} + c_{io}^{(\ell)}$ is linear in the gate input. Because the gate input is normalised and therefore lies in a ball of radius ρ_d, $\|o_k^{(\ell)}\| \le \|W_{io}^{(\ell)}\|\rho_d + \|c_{io}^{(\ell)}\|$ for every fixed choice of weights, which is the boundedness used in Appendix B; the magnitude itself is free, which is what the proof of Theorem 6 exploits. The drift scale $\alpha_k^{(\ell)}$ is applied only to the diagonal entries of the structured transition matrix $F_k^{(\ell)}$ (a diagonal, row-wise Hadamard action), while $o_k^{(\ell)}$ is added directly to the affine intercept $g_k^{(\ell)}$ with no additional Δt factor. Unless a prior is learned, the shared initial state is $Z_0 \sim \mathcal{N}(0, I_d)$, sampled once and reused by every layer.

C.5 Last-layer Girsanov tilt: design details

This subsection collects the design choices behind Section 4.

Why the initial state is tilted. The innovation of the final layer is not the only source of randomness the rare event depends on. In a gated stack the random initial latent $z_0 \sim \mathcal{N}(0, I_d)$ carries a large share of the terminal variance – on the TOY model of Section 6 the final-layer innovation explains only about a quarter of $\mathrm{Var}(Y_T)$ – and a controller acting on the innovation alone cannot reach the rare region: every such variant we trained kept the hit rate of the tilted paths at the plain Monte-Carlo level for $p \le 10^{-3}$. Under $\mathbb{Q}$, $z_0 = \mu_\theta + \xi$ with $\xi \sim \mathcal{N}(0, I_d)$ and a learned $\mu_\theta \in \mathbb{R}^d$ (one per controller), which contributes the closed-form factor $\mu_\theta^\top z_0 - \frac{1}{2}\|\mu_\theta\|^2$ of Eq. 12 to the log-likelihood ratio and $\frac{1}{2}\|\mu_\theta\|^2$ to $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})$. The prefix driver W^{prefix} keeps its law, so the prefix layers are computed by the same map under both measures; only the distribution of their starting point moves.

Controller. The control $u_{\theta,k}$ is an adapted function of the reference innovation history $\{\,\mathrm{d}\epsilon_j^{\mathbb{P}}\}_{j<k}$ and of stop-gradient features of the prefix latent path $\{L_j^{(L-1)}\}_{j\le k}$; it is parameterised by the diagonal state-space recurrence of Appendix C.6, driven by both inputs, with recurrence rates constrained to $(0,1)$ and zero drift at initialisation.

Objectives and gradient convention. Backbone and controller are trained on different objectives with separate optimisers. The backbone minimises the calibration loss, in which each far-tail term is estimated on the tilted batch by self-normalised importance sampling (Appendix C.7), combined with the plain estimate on the reference batch in proportion to the two effective sample sizes. The controller is trained by the cross-entropy method: it minimises a Monte-Carlo estimate of $\mathrm{KL}(\mathbb{Q}^\star\|\mathbb{Q}_\theta)$, where $\mathbb{Q}^\star \propto |\Phi|\,\mathbb{P}$ is the zero-variance proposal of the rare-event functional, under a quadratic wall on $\mathrm{KL}(\mathbb{Q}_\theta\|\mathbb{P})$ that keeps the proposal in the regime where the weights remain usable (Appendix C.8). The gradient convention matters: the cross-entropy objective is a score-function estimator and requires the log-ratio Eq. 12 to be differentiated with the sampled path held fixed, whereas Monte-Carlo functionals of the weights, such as the KL wall, require the reparameterised form in which the tilted noise is held fixed; the two forms have the same value and different fixed points, and using the reparameterised one for the cross-entropy objective drives the tilt away without bound.

Which estimator. Because the ratio is normalised, the vanilla importance-sampling estimator $n_q^{-1}\sum_i w_i\Phi(X^{(i)})$ is unbiased. For the calibration loss we use the self-normalised estimator (Appendix C.7): the weights are computed pathwise in log-space, self-normalisation is invariant to a common log-shift, so a max-shift can be applied before exponentiation and no single extreme path can dominate through overflow, the ratio form cancels weight fluctuations shared by numerator and denominator, and its $O(1/n_q)$ bias is negligible at our operating point. For rare-event probabilities the vanilla estimator is the better one: the initial-state shift gives large weights to tilted paths that do not reach the event, and those paths enter only the normalisation of the self-normalised estimator. Both estimators are compared in Appendix E.

C.6 Girsanov controller

The controller returns an adapted sequence $u_{\theta,k}\in\mathbb{R}^m$ that enters the Radon–Nikodym formula Eq. 12. It is a diagonal state-space recurrence with two branches, one driven by the tilted innovation and one by the stop-gradient prefix latent path,

$$s^W_{k+1} = \mathrm{diag}(r_W)\,s^W_k + U_W\,\Delta\epsilon^{\mathbb{Q}}_k, \qquad s^Z_{k+1} = \mathrm{diag}(r_Z)\,s^Z_k + U_Z\,L^{(L-1)}_k,$$
$$u_{\theta,k} = W_{\mathrm{mix}}\big[s^W_k, s^Z_k\big] + c,$$

so that $u_{\theta,k}$ depends on the increments before step k only. The recurrence is itself affine and can be scanned in parallel.

Implementation details. Three details proved necessary.

(i) The recurrence rates of both branches are parameterised as $\exp(-\,\mathrm{softplus}(-r))\in(0,1)$, initialised at 0.9997: an unconstrained rate that drifts above one overflows over 512 steps, which is how the earlier controllers produced non-finite weights.

(ii) The mixing weights of the latent-path branch are initialised at zero, so that the untrained controller is the identity tilt with $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})=0$; with a small random initialisation the branch, which integrates the $O(1)$ prefix latent over the whole horizon, gives the untrained controller a divergence of about 8 nats per controller, i.e. a collapsed proposal before any training.

(iii) The initial-state shift μ_θ is a free vector per controller initialised at zero.

Controller parameters are trained by Adam with learning rate $2\cdot10^{-3}$ and gradient clipping at norm one, separately from the backbone optimiser: with a single optimiser and a shared global-norm clip, the cross-entropy gradient, three orders of magnitude larger than the calibration gradient, scaled the backbone updates to zero after the controller was activated.

C.7 Self-normalised importance sampling and ESS

Given tilted samples $\{X^{(i)}\}_{i=1}^{n_q}$ drawn under $\mathbb{Q}$ with unnormalised weights

$$w_i = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X^{(i)}),$$

the self-normalised importance-sampling (SNIS) estimator of $\mathbb{E}^{\mathbb{P}}[\Phi]$ is

$$\widehat{\mathbb{E}}^{\mathbb{P},\mathrm{SNIS}}[\Phi] = \frac{\sum_{i=1}^{n_q} w_i \Phi(X^{(i)})}{\sum_{i=1}^{n_q} w_i}. \tag{83}$$

The estimator is biased at finite n_q but consistent. We monitor the quality of the tilted sample through the effective sample size

$$\mathrm{ESS} = \frac{\left(\sum_{i=1}^{n_q} w_i\right)^2}{\sum_{i=1}^{n_q} w_i^2} \in [1, n_q]. \tag{84}$$

$\mathrm{ESS} = n_q$ iff all weights are equal; $\mathrm{ESS} = 1$ iff one weight dominates. In practice we compute the log-weights $\log w_i = \log(\,\mathrm{d}\mathbb{P}/\,\mathrm{d}\mathbb{Q})(X^{(i)})$ directly from the discrete Radon–Nikodym density Eq. 12 and normalise them with a numerically stable log-sum-exp, so that the SNIS estimator Eq. 83 and the ESS ratio remain well behaved even when a few tilted paths dominate. In training logs we report the ratio $\mathrm{ESS}/n_q \in [1/n_q, 1]$ for comparability across batch sizes; a controller that drives ESS/n_q toward $1/n_q$ signals weight collapse and is held back by the KL wall of Appendix C.8. For rare-event probabilities we report the vanilla estimator $n_q^{-1}\sum_i w_i\Phi(X^{(i)})$ as the primary one: with the initial-state shift, tilted paths that do not reach the event can carry large weights, which affect only the denominator of Eq. 83 and make the self-normalised estimator markedly worse (Table 11). Its truncated variant caps the weights at $\sqrt{n_q}\,\mathbb{E}^{\mathbb{Q}}[w] = \sqrt{n_q}$, using the known mean rather than the sample mean, which the very weights to be tamed would inflate.

C.8 Training objectives of backbone and controller

Backbone. The backbone parameters minimise the calibration loss Eq. 5. Terms that involve a rare-event functional Φ are estimated on the tilted batch by the self-normalised estimator Eq. 83, and, in the combined variant, averaged with the plain estimate on the reference batch of size n_p with the weights $n_p/(n_p + \mathrm{ESS})$ and $\mathrm{ESS}/(n_p + \mathrm{ESS})$, held fixed for the gradient; all other terms are plain Monte-Carlo estimates on the reference batch. The backbone receives no gradient through the weights.

Controller. The controller parameters $\theta_c = (u_\theta, \mu_\theta)$ minimise the cross-entropy objective with a KL wall,

$$\begin{aligned}\widehat{\mathcal{J}}(\theta_c) &= -\sum_{i=1}^{n_q} \bar{w}_i^\star \,\log \frac{\mathrm{d}\mathbb{Q}_{\theta_c}}{\mathrm{d}\mathbb{P}}(X^{(i)}) + \lambda_{\mathrm{KL}}\,\widehat{\mathrm{KL}}(\mathbb{Q}_{\theta_c}\|\mathbb{P}) + \big(\widehat{\mathrm{KL}}(\mathbb{Q}_{\theta_c}\|\mathbb{P}) - \kappa\big)_+^2, \\ \bar{w}_i^\star &\propto |\Phi(X^{(i)})|\,\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}_{\theta_c}}(X^{(i)}),\end{aligned} \tag{85}$$

where the self-normalised target weights $\bar{w}_i^\star$ are held fixed for the gradient (their log-ratios clipped at ± 5 so that no single path can dominate the update), κ is the wall in nats (6 on DAX; $-\log p + 2$ for an event of nominal probability p on TOY), and $\widehat{\mathrm{KL}}$ is the energy form $n_q^{-1}\sum_i\big(\frac{1}{2}\sum_k \|u_{\theta,k}\|^2\Delta t + \frac{1}{2}\|\mu_\theta\|^2\big)$, whose $\mathbb{Q}$-expectation equals $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})$ exactly. The first term is the Monte-Carlo estimate of $\mathrm{KL}(\mathbb{Q}^\star\|\mathbb{Q}_{\theta_c})$ up to a constant, with $\mathbb{Q}^\star \propto |\Phi|\,\mathbb{P}$ the zero-variance proposal; its minimiser is the moment match $\mathbb{E}^{\mathbb{Q}^\star}$ of the sufficient statistics of the tilt (the conditional mean of z_0 and of the innovation increments given the rare event). The wall replaces the weak penalty $\lambda_{\mathrm{KL}} = 10^{-4}$ of the earlier revision, which does not prevent the collapse of the proposal onto a handful of paths once the target weights degenerate.

Gradient convention. Eq. 12 can be differentiated in two ways with the same value. In the score form the sampled path is held fixed: $\partial_{u_{\theta,k}} \log(\,\mathrm{d}\mathbb{Q}/\,\mathrm{d}\mathbb{P}) = \Delta\epsilon_k^{\mathbb{Q}}$ and $\partial_{\mu_\theta} = z_0 - \mu_\theta$, the scores of the tilted law, which is what the cross-entropy term requires. In the reparameterised form the tilted noise is held fixed and the derivative is $\Delta\epsilon_k^{\mathbb{P}}$ (respectively z_0), which is the correct total derivative of a $\mathbb{Q}$-expectation of a function of the weights, such as the KL wall or the effective sample size. Using the reparameterised form inside the cross-entropy term gives an update whose fixed point is not the moment match and which pushes the tilt outward without bound; the earlier revision had this defect.

D EXPERIMENTAL SETUP AND HYPERPARAMETER GRIDS

TOY DATASET

TOY is a controlled synthetic benchmark designed to separate path fitting from path-functional calibration. The ground truth is a three-component nonlinear Itô SDE on $[0, T]$ with $T = 1$, state $X_t = (X_t^1, X_t^2, X_t^3)$ and diagonal state-dependent diffusion:

$$\begin{aligned}
\mathrm{d}X_t^1 &= -X_t^1\big((X_t^1)^2 - 1\big)\,\mathrm{d}t + \sqrt{|X_t^1| + 1}\,\sqrt[3]{X_t^1 + 5}\,\mathrm{d}W_t^1, && X_0^1 = 0.5,\\
\mathrm{d}X_t^2 &= \sin(X_t^2)\big(2 - (X_t^2)^2\big)\,\mathrm{d}t + \sqrt[3]{1 + (X_t^2)^2}\,\mathrm{d}W_t^2, && X_0^2 = 0,\\
\mathrm{d}X_t^3 &= \big(X_t^3 - (X_t^3)^3/3\big)\,\mathrm{d}t + \exp\big(-(X_t^3)^2/4\big)\,X_t^3\;\mathrm{d}W_t^3, && X_0^3 = 1,
\end{aligned}$$

with mutually independent driving Brownians. The observable is the linear combination

$$Y_t = 0.40\,X_t^1 + 0.35\,X_t^2 + 0.25\,X_t^3.$$

We simulate $2{\times}10^5$ Monte-Carlo paths via Euler–Maruyama on a uniform grid of $N = 2048$ steps. From the simulated paths we compute four classes of supervised path-functional targets at evaluation times $\mathcal{T} = \{0.1,\ 0.25,\ 0.5,\ 0.75,\ 1.0\}$.

These targets are:

- **Thresholded positive-part functionals** at levels $\mathcal{K} = \{-0.5, -0.25, 0, 0.25, 0.5, 0.75, 1.0\}$:
$$\mathbb{E}[\max(Y_T - K, 0)].$$
- **Running maximum**:
$$\mathbb{E}[\max_{0\le t\le T} Y_t].$$
- **Squared path average**:
$$\mathbb{E}\big[\tfrac{1}{T}\int_0^T Y_t^2\;\mathrm{d}t\big],$$
computed on the discrete grid.
- **Threshold-crossing probabilities**
$$\mathbb{P}(\max_{0\le t\le T_i} Y_t \ge H_j)$$
at three threshold levels H_j chosen as the 80%, 90%, and 95% empirical quantiles of the terminal running maximum; in the objective the indicator is replaced by a smooth approximation (a sigmoid of the excess over the threshold with a fixed temperature), so that Eq. 5 is differentiable.

The threshold-crossing and high-level positive-part targets are deliberately tail-sensitive path functionals. Errors on typical-path statistics and on tail functionals need not agree: a model may fit typical trajectories well while still missing the rare excursions or extrema that determine these calibration targets. Thus TOY provides a controlled setting for evaluating whether a model can match both ordinary path behaviour and rare path-dependent statistics.

DAX DATASET

DAX is a real-data functional calibration benchmark built from historical European option quotes on the DAX index. The purpose of the benchmark is to test whether a stochastic path generator can match a large collection of expectations of nonlinear functions of its terminal value. Each option quote is treated as a supervised target of the form

$$\mathbb{E}[\varphi(Y_T; K)],$$

where Y is the generated one-dimensional path, T is the evaluation time, K is a threshold level, and φ is a payoff function. All thresholds and target values are normalised by the spot level $S_0 = 25{,}280$, so that the generated process starts from $Y_0 = 1$ and $K_{\mathrm{norm}} = K/S_0$.

Option payoffs (for the non-specialist). A European option on the index with maturity T and strike K pays a fixed function of the terminal value Y_T alone. The two basic instruments are the *call* and the *put*, with payoffs

$$\varphi_{\text{call}}(Y_T; K) = (Y_T - K)_+, \qquad \varphi_{\text{put}}(Y_T; K) = (K - Y_T)_+, \qquad (x)_+ := \max(x, 0). \tag{86}$$

Under the normalised, discounting-free pricing convention used here, the quoted price of an option is the expectation

$$c_{T,K} = \mathbb{E}\big[\varphi(Y_T; K)\big], \tag{87}$$

which is exactly the supervised target above. A strike is *at-the-money* when $K \approx S_0$ (i.e. $K_{\text{norm}} \approx 1$), *in-the-money* when the option would pay out if exercised at the current level, and *out-of-the-money* (OTM) otherwise, that is, a call with $K > S_0$ or a put with $K < S_0$. OTM prices are small and depend only on the tails of the terminal law, which is why the far-tail groups are the hardest part of the benchmark. The threshold-indicator targets approximate digital payoffs through finite differences of nearby put prices in the strike,

$$\mathbb{P}(Y_T \le K) = \partial_K \, \mathbb{E}\big[(K - Y_T)_+\big] \approx \frac{\mathbb{E}\big[(K + h - Y_T)_+\big] - \mathbb{E}\big[(K - h - Y_T)_+\big]}{2h}, \tag{88}$$

and analogously $\mathbb{P}(Y_T > K) = -\partial_K \, \mathbb{E}[(Y_T - K)_+]$ from call prices.

The dataset is intentionally dense: at each available maturity we include several groups of terminal functionals covering both typical and tail regions of the path distribution:

- 25 positive-part upper-tail targets, $K_{\text{norm}} \in [0.88, 1.12]$, corresponding to $\mathbb{E}[(Y_T - K)_+]$;
- 25 positive-part lower-tail targets, $K_{\text{norm}} \in [0.88, 1.12]$, corresponding to $\mathbb{E}[(K - Y_T)_+]$;
- 20 far lower-tail targets, $K_{\text{norm}} \in [0.60, 0.88]$;
- 15 far upper-tail targets, $K_{\text{norm}} \in [1.08, 1.35]$;
- 12 threshold-indicator targets, synthesised from differences of nearby lower-tail positive-part targets, $K_{\text{norm}} \in [0.65, 0.98]$.

The target threshold levels are evenly spaced on each interval and are matched to the nearest observed threshold in the raw quote table. Maturity indices are recomputed on the model time grid with $T = 1$ and $N = 1024$ steps.

Quote-backed targets. The dense grids are filled by linear interpolation of the quotes available at each expiry. Beyond the deepest quote of a side the interpolation is constant, so every deeper grid strike would receive the deepest quote's price, and a band without quotes would receive zero: such entries are not prices, and no arbitrage-free model can match a price that does not decrease with the depth of the strike. A grid entry is therefore marked valid only if its strike lies inside the quoted strike range of its side at that expiry (both legs for a digital spread) and its target is positive, and training and held-out losses are computed on valid entries only. On DAX this keeps $114/125$ calls, $120/125$ puts, $59/100$ far puts, $22/75$ far calls and $43/60$ digitals. The Girsanov tables of this revision (Table 9 and Appendix E) use these masks. Under a calibrated model the far strikes that remain have exceedance probabilities between 0.3% and 16%, i.e. the DAX far tails are moderately rare rather than rare events at a batch of $1\,024$ paths.

The far-tail groups are the most challenging part of the benchmark because they depend on rare terminal events under the generated path distribution. When the Girsanov overlay is used for DAX, **we use two separate controllers**: one targets the far lower-tail functionals and the other targets the far upper-tail functionals. This lets each controller focus on a different rare region of path space, rather than requiring a single tilt to cover both tails simultaneously.

Train/eval split. Evaluation uses a structured hold-out over strike levels. For each group of targets, one quarter of the strikes are held out for evaluation and the remaining entries are used for training. The held-out strikes are selected non-contiguously, so that the training and evaluation sets cover overlapping ranges. This tests interpolation across the surface rather than extrapolation outside the observed range.

SPX DATASET

SPX is a second real-data option-surface benchmark, constructed identically to DAX but from historical European option quotes on the S&P 500 index. As for DAX, each quote is a supervised target $\mathbb{E}[\varphi(Y_T; K)]$ with the call and put payoffs defined above; all strikes and target values are normalised by the spot level so that $Y_0 = 1$ and $K_{\text{norm}} = K/S_0$, and maturities are recomputed on the model grid with $T = 1$ and $N = 1024$ steps. The surface is organised into the same five target groups (near-tail calls and puts, far lower- and upper-tail targets, and threshold-indicator targets) spanning comparable normalised-strike ranges, and evaluation uses the same structured non-contiguous 25% strike hold-out, testing interpolation across the surface. When the Girsanov overlay is used, we again employ two controllers, one per far tail. SPX serves as an out-of-sample check that the architecture and calibration procedure transfer across underlyings; results are reported in Table 4 of Appendix E.

TRAINING PROTOCOL

All models use the same optimiser and learning-rate schedule unless otherwise stated: AdamW with $\beta = (0.9, 0.999)$, peak learning rate 10^{-3}, 100 warm-up steps followed by cosine decay, gradient clipping at 1, batch size 1024, evaluation batch size 2048, 1 000 epochs and seven seeds per configuration; every reported number is a mean $\pm$ standard error over the seven seeds of the cell in question. The models are simulated on $N = 512$ time steps. SLiSDE uses the offset gate of Eq. 82 with the RMS-normalised gate input. The grid searches the width $d \in \{32, 64\}$ and the depth $L \in \{2, 3\}$ for every family with every other choice fixed (Appendix D), and the best cell per family and depth is reported. The selected widths are: on TOY, $d = 64$ for the Neural SDE and for the two-layer SLiSDE and $d = 32$ for the three-layer SLiSDE; on DAX, $d = 32$ for the Neural SDE and $d = 64$ for SLiSDE; on SPX, $d = 32$ for SLiSDE and $d = 32$ (two layers) or $d = 64$ (three layers) for the Neural SDE; SLiCE selects $h = 64$ throughout.

For Girsanov runs the controller is kept inactive for the first 200 epochs. After activation each training step simulates 1 024 reference paths and two tilted batches of 512 paths, one per controller (far puts, far calls), with $\rho = 0$. Each far-tail term is estimated by self-normalised importance sampling on its tilted batch, combined with the plain estimate on the reference batch in proportion to the effective sample sizes (the 'tilted batch only' variant omits the reference-batch estimate); all other terms are evaluated under $\mathbb{P}$ on the reference batch. The controllers minimise Eq. 85 with $\lambda_{\text{KL}} = 10^{-4}$, a 6-nat wall, target log-weights clipped at ± 5 and Adam with learning rate $2 \cdot 10^{-3}$ on their own parameters; the backbone keeps the optimiser and schedule of the vanilla runs. The tilt studies use seven seeds and 1 000 epochs; the tilted arms cost $1.3\times$ the epoch time of the budget-matched untilted arm (303 against 229 ms).

SEARCH GRIDS

One grid is shared by the three benchmarks. Every structural choice is held fixed across the model families, so that the comparison isolates the architecture, and only depth and width are searched.

- Searched for every family: depth $L \in \{2, 3\}$ and width $\in \{32, 64\}$, the width being the latent dimension d of SLiSDE and of the Neural SDE and the hidden dimension h of SLiCE.
- Fixed across families: block-diagonal transition structure with block size 8 for SLiSDE and SLiCE, noise dimension $m = 8$, RMS normalisation.
- Fixed within SLiSDE: general noise, gate amplitude $\varepsilon = 0.1$, time-embedding dimension 8, sequential evaluation during training; time-dependent coefficients on TOY and DAX (static coefficients on SPX); the causal convolution of Appendix C.3 with kernel size 8 on DAX and SPX (none on TOY); gate weights initialised with standard deviation 0.25 on TOY and 0.05 on the option surfaces.
- Fixed within the Neural SDE: drift and diffusion networks of width 128 with $\tanh$ activations, full diffusion matrix, layer normalisation.
- Fixed within SLiCE: Euler transition; the Brownian path, rather than its increments, as input.

DAX ablations. Two one-factor sweeps around the two-layer DAX cell of SLiSDE ($d = 64$, block size 8, general noise): the noise type $\in \{$multiplicative, additive, general$\}$ at block size 8, and the block size $\in \{1, 4, 8, 16\}$ at general noise.

Girsanov controllers.

- state-space controller with the latent-path branch and the initial-state shift (Appendix C.6);
- controller objective $\in \{$cross-entropy, SNIS-variance$\}$; far-tail estimate $\in \{$combined, tilted batch only$\}$;
- KL wall $\kappa \in \{2, 6\}$ nats, $\lambda_{\mathrm{KL}} = 10^{-4}$;
- last-layer driver correlation $\rho \in \{0.0, 0.9\}$.

E ADDITIONAL EXPERIMENTAL RESULTS

This appendix collects the results referred to from Section 6: the SPX surface at two and three layers, the structural ablations (noise type and block size), the timing study, the distributional-recovery study on TOY, and the Girsanov tilt studies on DAX and TOY. Unless stated otherwise, each study varies a single factor of the tuned gated in-flow cell of the corresponding main-text table, with the calibration targets held fixed.

SPX OPTION-SURFACE DATASET

Table 4 reports the SPX option-surface benchmark for two- and three-layer configurations of every family under the protocol of Table 3 (held-out 25% strike mask; tuned configuration per model; seven seeds). SLiSDE is within one standard error of the held-out loss of SLiCE with two layers and within 6% of it with three, at 6–8$\times$ fewer parameters and 3.5–4$\times$ faster epochs; with three layers it attains the lowest far-put error. The Neural SDE is clearly worse on both metrics at both depths.

Model	Params	Loss (10^{-4})	Far put (10^{-6})	ms / epoch
Neural SDE – $L = 2$	79K	11.90 ± 0.11	135.4 ± 2.9	94.1
SLiCE – $L = 2$	142K	$\mathbf{6.43 \pm 0.12}$	$\mathbf{81.4 \pm 4.6}$	170.1
SLiSDE gated in-flow – $L = 2$	18K	6.44 ± 0.40	83.7 ± 4.8	**44.3**
Neural SDE – $L = 3$	157K	10.29 ± 0.13	136.0 ± 2.8	109.7
SLiCE – $L = 3$	213K	$\mathbf{6.53 \pm 0.26}$	82.9 ± 3.6	248.4
SLiSDE gated in-flow – $L = 3$	34K	6.29 ± 0.24	$\mathbf{70.1 \pm 6.6}$	**70.5**

Table 4: SPX option-surface results without Girsanov tilt, two and three layers. Same masks and protocol as Table 3; best cell of the search grid per family and depth, seven seeds, mean $\pm$ standard error. The narrower three-layer SLiCE cell diverged on every seed and is excluded.

NOISE TYPE AND BLOCK SIZE

Each layer can carry multiplicative noise only ($d_t^j = 0$), additive noise only ($C_t^j = 0$) or both (the noise-type remark of Section 3.1), and its structured coefficients are block-diagonal with block size b_s. Table 5 varies each factor separately from the two-layer DAX cell of Table 3. The three noise types are within one standard error of each other on the total loss; additive noise is cheaper, while the variants with multiplicative noise fit the far puts better. The total loss is flat in the block size; blocks of 4 or more improve the far puts over the diagonal model, at a cost in time per epoch that grows with b_s.

Variant	Params	Loss (10^{-4})	Far put (10^{-6})	ms / epoch
Noise type (block size 8)				
multiplicative	135K	2.540 ± 0.134	$\mathbf{13.0 \pm 2.5}$	73.5
additive	62K	$\mathbf{2.282 \pm 0.175}$	22.6 ± 6.9	**62.2**
general (both)	136K	2.413 ± 0.141	13.4 ± 2.1	79.4
Block size (general noise)				
$b_s = 1$ (diagonal)	63K	2.290 ± 0.116	19.4 ± 4.2	**47.7**
$b_s = 4$	94K	$\mathbf{2.276 \pm 0.153}$	$\mathbf{12.7 \pm 2.5}$	61.0
$b_s = 8$	136K	2.413 ± 0.141	13.4 ± 2.1	79.4
$b_s = 16$	219K	2.345 ± 0.198	$\mathbf{12.7 \pm 2.7}$	112.2

Table 5: Structural ablations on DAX, each varying one factor of the two-layer gated in-flow cell of Table 3 (general noise, block size 8; seven seeds, mean $\pm$ s.e.; held-out loss and far-put loss). Top: noise type of the structured layers. Bottom: block size of the block-diagonal structure.

TIMING STUDY

Setup. Tables 6 and 7 report wall-clock times after compilation on one NVIDIA RTX 6000 Ada (42.6 GiB usable), as the median of five repeats after two warm-ups, across batch sizes B and horizons T. The models are the two-layer DAX cell of SLiSDE (136K parameters), evaluated *sequentially* so that the comparison isolates the structured layers from the scan, the tuned Neural SDE (124K), a parameter-matched Neural SDE (137K, hidden width 138) and SLiCE at the same depth (142K).

Training step. SLiSDE is 5–7× faster than the Neural SDE at $B = 64$, 2.7× at $B = 256$ and 1.2–1.4× at $B = 1024$ (against the tuned and the parameter-matched Neural SDE respectively). At $B = 4096$, where batch parallelism saturates the GPU, the Neural SDE step is 1.6× faster, and SLiCE runs out of memory beyond $B = 1024$ at $T = 512$.

B	T	SLiSDE (seq.)	Neural SDE	SLiCE	speed-up
64	512	9.5	50.8	10.1	5.3×
64	2048	35.8	185.1	42.8	5.2×
64	8192	117.1	732.9	156.9	6.3×
256	2048	79.7	218.8	155.2	2.7×
1024	512	69.0	81.6	155.0	1.2×
1024	2048	258.9	322.7	OOM	1.2×
4096	512	242.1	153.6	OOM	0.6×

Table 6: Wall-clock time per training step (ms; forward plus backward, post-compilation, median of five repeats) as a function of batch size B and horizon T for SLiSDE (136K parameters, sequential evaluation), the tuned Neural SDE (124K) and SLiCE (142K). Speed-up is Neural SDE over SLiSDE; OOM: exceeds the 42.6 GiB of the GPU.

Forward pass and scan. Table 7 separates forward and backward passes and adds the chunked associative scan (best chunk size of 16, 64, 256). The scan shortens the forward pass by 1.4× at $B = 64$, $T = 512$ and is otherwise on par with or slower than the sequential evaluation: each sequential structured update is already cheap and the chunked scan roughly doubles the arithmetic, so its advantage lies in parallel depth rather than in wall-clock time at these sizes.

B	T	SLiSDE fwd (seq.)	SLiSDE fwd (scan)	NSDE fwd	SLiSDE fwd+bwd	NSDE fwd+bwd	NSDE (matched) fwd+bwd
64	512	5.7	4.1	12.7	9.5	50.8	54.0
64	2048	14.9	15.8	45.5	35.8	185.1	212.0
64	8192	48.4	50.4	168.6	117.1	732.9	838.9
256	2048	32.5	44.6	51.9	79.7	218.8	257.1
1024	512	27.6	39.9	21.6	69.0	81.6	89.4
1024	2048	89.9	OOM	74.2	258.9	322.7	359.5
4096	512	80.5	OOM	32.1	242.1	153.6	178.6

Table 7: Wall-clock time (ms, post-compilation, NVIDIA RTX 6000 Ada) of forward passes and full training steps as a function of batch size and horizon, including the chunked associative scan (best of chunk sizes 16, 64, 256) and the parameter-matched Neural SDE (137K parameters).

DISTRIBUTIONAL RECOVERY ON TOY

Because the TOY generator is known, the learned *distribution* can be assessed directly, well beyond the calibrated targets. Table 8 reports Kolmogorov–Smirnov and Wasserstein-1 distances to the ground truth, computed from 2×10^4 model paths for the best two-layer cell of each family (seven seeds), for the terminal value and for the running maximum. SLiSDE and the Neural SDE recover the terminal marginal comparably well (SLiSDE better in KS, the Neural SDE in W_1); on the law of the running maximum, the genuinely path-dependent statistic, SLiSDE is markedly better than both baselines.

	KS (terminal)	W_1 (terminal)	KS (running max)	W_1 (running max)
Neural SDE ($L = 2$)	0.019 ± 0.002	$\mathbf{0.020 \pm 0.002}$	0.059 ± 0.002	0.047 ± 0.002
SLiCE ($L = 2$)	0.022 ± 0.002	0.031 ± 0.003	0.081 ± 0.003	0.067 ± 0.003
SLiSDE ($L = 2$)	$\mathbf{0.015 \pm 0.001}$	0.023 ± 0.002	$\mathbf{0.036 \pm 0.002}$	$\mathbf{0.037 \pm 0.002}$

Table 8: Distances between the learned and the ground-truth laws of the terminal value and of the running maximum on TOY (2×10^4 model paths; best two-layer cell of each family; seven seeds, mean $\pm$ s.e.).

GIRSANOV TILT: CALIBRATION ON DAX AND RARE-EVENT ESTIMATION ON TOY

Calibration with the tilt on DAX: protocol. Table 9 compares five arms over seven seeds under one evaluation protocol, plain Monte Carlo under the reference law with 32 768 paths on the held-out strikes: the vanilla model; the untilted reference-law model with 1 024 and with 2 048 paths per step, the latter being the path budget of the tilted arms; and two tilted arms, which simulate 1 024 reference paths and two tilted batches of 512 paths, one controller per tail. The tilted arms differ in the far-tail estimate: one combines the tilted batch with the reference batch, the other uses the tilted batch alone. The far-tail targets are restricted to strikes backed by a quote (Appendix D).

Calibration with the tilt on DAX: results. Relative to the vanilla model, the combined variant halves the far-call error ($0.50\times$, six seeds of seven, paired $p = 0.03$) and lowers the total held-out loss by 15%; the tilted-batch-only variant halves the far-put error on average, with a four times smaller spread across seeds, and lowers the total loss by 24%. Against the budget-matched untilted arm the tilted arms are never worse and are better on average on the far tails ($0.66\times$ on far calls and $0.72\times$ on far puts for the respective variants), differences that are not significant at seven seeds. The tilt is a genuine change of measure here, $\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) \approx 3$ nats and $\mathrm{ESS}/n_q \approx 0.05$ on the put side, and costs $1.3\times$ the epoch time of the budget-matched untilted arm.

Arm	paths / step	Loss (10^{-4})	Far put (10^{-6})	Far call (10^{-6})	ESS $/n_q$ put / call	KL put / call
Vanilla	1024	1.15 ± 0.12	0.93 ± 0.46	3.15 ± 0.33	–	–
Untilted, $\rho = 0$	1024	1.18 ± 0.19	0.87 ± 0.42	3.62 ± 0.78	–	–
Untilted, $\rho = 0$	2048	0.95 ± 0.13	0.81 ± 0.30	2.59 ± 0.49	–	–
Tilt, combined far-tail estimate	$1024 + 2\times512$	0.98 ± 0.10	0.82 ± 0.27	$\mathbf{2.03 \pm 0.78}$	0.05 / 0.09	2.9 / 1.5
Tilt, tilted batch only	$1024 + 2\times512$	$\mathbf{0.88 \pm 0.10}$	$\mathbf{0.48 \pm 0.08}$	3.25 ± 1.04	0.05 / 0.12	3.0 / 1.4

Table 9: DAX with the Girsanov tilt (seven seeds, mean $\pm$ standard error; same masks and protocol as Table 3 otherwise). Far-tail losses are held-out MSEs on the quote-backed far strikes (Appendix D); every arm is evaluated by plain Monte Carlo under its reference law with 32 768 paths. The tilted arms use $\rho = 0$, the cross-entropy controller with a 6-nat KL wall and the initial-state shift; ESS/n_q and $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})$ (nats) are end-of-training values of the put / call controllers; time per epoch 130 / 130 / 229 / 303 / 295 ms.

The tilt as a learned importance sampler: protocol. To isolate the benefit of the tilt from the calibration task, we evaluate it on TOY in a controlled estimator study against ordinary Monte Carlo at the same compute budget. The reference law is the three-layer TOY cell of Table 2, simulated at $N = 512$ steps with $\rho = 0$ and calibrated seven times with different seeds. The targets are the exceedance probabilities $\mathbb{P}(Y_T > H)$ with H at the model's own quantiles of nominal probability 10^{-2}, 10^{-3} and 10^{-4}; reference values come from 8×10^5 plain Monte-Carlo paths. For each backbone one controller per threshold is trained post hoc, in three stages of 1 000 cross-entropy steps on batches of 1 024 tilted paths, the stage for the rarer event starting from the controller of the previous one, under the wall $\kappa = -\log p + 2$ nats. Every estimator receives the same per-estimate path budget n; in addition, plain Monte Carlo receives a cost-matched budget inflated by the measured per-path overhead of the tilted simulation (a factor 1.18). The relative root-mean-square error of every estimator is measured over 100 independent replications, and the tables report the mean and standard error over the seven backbones.

The tilt as a learned importance sampler: results. Table 10 reports the budget $n = 1024$ for plain Monte Carlo, for the tilt with the vanilla importance-sampling estimator and with the self-normalised one, together with the fraction of replications in which plain Monte Carlo returns exactly zero because no path reaches the threshold. The picture is the one predicted by the variance argument of Section 4:

the gain of the tilt grows with the rarity of the event, from a factor 2 in relative error at $p = 10^{-2}$ to a factor 4 at $p = 10^{-4}$, where plain Monte Carlo returns zero in 90% of the replications. The self-normalised estimator is markedly weaker, because the initial-state shift assigns large weights to tilted paths that miss the event and those weights enter only its normalisation.

Estimator ($n = 1024$), rel. RMSE	$p \approx 10^{-2}$	$p \approx 10^{-3}$	$p \approx 10^{-4}$
Plain MC	0.320 ± 0.005	1.051 ± 0.019	3.39 ± 0.39
Plain MC, cost-matched ($\times 1.18$)	0.278 ± 0.009	0.934 ± 0.047	3.15 ± 0.37
Girsanov tilt, IS	$\mathbf{0.156 \pm 0.018}$	$\mathbf{0.394 \pm 0.104}$	$\mathbf{0.785 \pm 0.184}$
Girsanov tilt, SNIS	0.274 ± 0.021	0.789 ± 0.106	1.88 ± 0.42
Share of plain-MC runs returning exactly 0	0%	34%	90%

Table 10: Rare-event estimation on TOY: relative RMSE of each estimator at a budget of $n = 1024$ paths per estimate (cost-matched plain Monte Carlo receives $1.18\,n$ paths), over 100 replications for each of seven independently calibrated models; mean $\pm$ standard error over the models.

Budgets and diagnostics. Table 11 extends Table 10 to the budgets $n \in \{256, 1024, 4096\}$ and adds the truncated vanilla estimator (weights capped at $\sqrt{n}$). At the end of training the controllers use 2.1, 3.5 and 5.1 nats of divergence, and the tilted paths hit the event at rates of 26%, 20% and 15% for $p = 10^{-2}$, 10^{-3} and 10^{-4} (plain Monte Carlo: 1%, 0.1%, 0.01%), at ESS $/n_q$ of 0.045, 0.024 and 0.011. Two features of the table deserve comment. The self-normalised estimator is uniformly worse than the vanilla one, for the reason given in Appendix C.7, and the truncated estimator coincides with the untruncated one except at $p = 10^{-2}$. The weights are heavy-tailed ($\mathbb{E}^{\mathbb{Q}}[w]$ estimated on the tilted batch is 0.75–0.97 instead of 1), so the standard errors across backbones are large and individual cells can lose to plain Monte Carlo; the gain of the tilt nevertheless grows with the rarity of the event at every budget, and at $p = 10^{-4}$ it is the only estimator that returns a non-zero value in most replications at $n \leq 1024$.

	rel. RMSE	$n = 256$	$n = 1024$	$n = 4096$
$p \approx 10^{-2}$	Plain MC	0.645 ± 0.017	0.320 ± 0.005	0.163 ± 0.005
	Plain MC, cost-matched	0.560 ± 0.012	0.278 ± 0.009	0.151 ± 0.003
	Tilt, IS	$\mathbf{0.411 \pm 0.084}$	$\mathbf{0.156 \pm 0.018}$	0.140 ± 0.033
	Tilt, IS truncated	0.391 ± 0.070	0.156 ± 0.018	$\mathbf{0.119 \pm 0.017}$
	Tilt, SNIS	0.645 ± 0.131	0.274 ± 0.021	0.197 ± 0.034
	plain MC returning 0	8%	0%	0%
$p \approx 10^{-3}$	Plain MC	1.953 ± 0.083	1.051 ± 0.019	0.501 ± 0.014
	Plain MC, cost-matched	1.780 ± 0.039	0.934 ± 0.047	0.460 ± 0.015
	Tilt, IS	$\mathbf{0.649 \pm 0.147}$	$\mathbf{0.394 \pm 0.104}$	$\mathbf{0.330 \pm 0.075}$
	Tilt, IS truncated	0.649 ± 0.147	0.394 ± 0.104	0.330 ± 0.075
	Tilt, SNIS	1.610 ± 0.287	0.789 ± 0.106	0.482 ± 0.066
	plain MC returning 0	75%	34%	1%
$p \approx 10^{-4}$	Plain MC	6.64 ± 0.95	3.39 ± 0.39	1.65 ± 0.10
	Plain MC, cost-matched	6.96 ± 0.40	3.15 ± 0.37	1.65 ± 0.13
	Tilt, IS	$\mathbf{1.77 \pm 0.59}$	$\mathbf{0.785 \pm 0.184}$	$\mathbf{0.816 \pm 0.235}$
	Tilt, IS truncated	1.77 ± 0.59	0.785 ± 0.184	0.816 ± 0.235
	Tilt, SNIS	4.06 ± 1.08	1.88 ± 0.42	1.81 ± 0.72
	plain MC returning 0	97%	90%	68%

Table 11: Rare-event estimation on TOY: relative RMSE for three budgets and three nominal probabilities, over 100 replications for each of seven independently calibrated models (mean $\pm$ standard error over the models); cost-matched plain Monte Carlo receives $1.18\,n$ paths; the last row of each block is the share of plain-Monte-Carlo replications with no path above the threshold.

EFFECT OF THE CONTROLLER OBJECTIVE, THE KL WALL AND THE FAR-TAIL ESTIMATE

The wall κ in Eq. 85 controls how far the Girsanov proposal may move from the reference measure, and the choice of controller objective decides whether it moves at all. Table 12 varies the wall, the objective and the way the far-tail terms are estimated on DAX. The variance-type objective, which minimises the self-normalised estimate of the far-tail loss directly, converges to the identity tilt (ESS $/n_q = 0.998$) and its arm is then an untilted model with $1\,024$ extra paths on the far-tail terms; the cross-entropy objective produces a proposal with 2–3 nats of divergence at either wall. Combining the tilted-batch estimate with the reference-batch estimate helps the far calls, whose controller is the weaker of the two, and hurts the far puts, where the tilted batch alone is the better estimator.

Controller	Loss (10^{-4})	Far put (10^{-6})	Far call (10^{-6})	ESS $/n_q$ put / call
CE, $\kappa = 6$, combined (7 seeds)	0.98 ± 0.10	0.82 ± 0.27	2.03 ± 0.78	0.05 / 0.09
CE, $\kappa = 6$, tilted batch only (7 seeds)	0.88 ± 0.10	0.48 ± 0.08	3.25 ± 1.04	0.05 / 0.12
CE, $\kappa = 2$, combined (3 seeds)	1.13 ± 0.20	0.59 ± 0.01	2.46 ± 1.03	0.06 / 0.08
SNIS-variance, combined (3 seeds)	1.18 ± 0.35	1.18 ± 0.33	2.71 ± 1.10	1.00 / 0.99

Table 12: Controller ablations on DAX far-tail calibration ($\rho = 0$, quote-backed targets, protocol of Table 9; mean $\pm$ standard error over seeds).