THE GAUGE FREEDOM OF ROTATIONAL DRIFT IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The forward process of a diffusion model is a design choice, yet most standard constructions use a purely dissipative drift $f = -\beta(t)x$ — the gradient component of the Helmholtz decomposition. We ask what the missing divergence-free half contributes within the exactly solvable family $dx = (-\gamma I + r(\tau)J_0)x\,d\tau + \sqrt{2}\,dW_\tau$, where $J_0^\top = -J_0$ is fixed. It leaves the generated distribution unchanged: an explicit time-dependent change of coordinates — rotation, rescaling, and time reparametrization — reduces the Fokker–Planck equation to the standard heat equation for every skew profile $J(\tau)$, so rotating and non-rotating processes differ only by the explicit map $R(\tau)$, which is the identity at the generation endpoint. Rotation is not inert, though — at constant rotation strength the process carries a stationary probability current with entropy production rate exactly $\sigma = \rho^2\|J_0\|_F^2 = 2\lfloor d/2\rfloor\rho^2$, breaking detailed balance underneath an endpoint law that is untouched. The rotational content resides in the path measure and is absent from the marginal score in the co-moving frame. The two conditions delimiting this gauge are not symmetric in kind. The reduction requires conformal dissipation, but the gauge then holds throughout the entire conformal class, anisotropic noise included, so the first condition is a design degree of freedom rather than a boundary. Realized through the co-moving frame of the rotational field, the gauge persists exactly for Killing fields; a differential rotation, though divergence-free and norm-preserving, already escapes it. Under model misspecification, rotational mismatch becomes visible through a single-peaked visibility window whose position is set by the trade-off between rotation and the dissipative decay of non-radial structure, weakly dependent on rotation strength and shifted later by multimode geometry, while its depth grows with rotation strength and its shape separates covariance-driven from multimode-driven visibility. We verify these claims analytically and through closed-form evaluations in 2D and at $d = 256$ on MNIST, with trained-model checks in the local heat-ball framework. Source code is available at https://github.com/Anonymous-510/anonymous-submission.

1 INTRODUCTION

A diffusion model is specified by two choices: a forward stochastic process that destroys structure, and a learned score used to construct the reverse time dynamics. The first is, in the canonical construction, a single fixed template — an Ornstein–Uhlenbeck process with isotropic gradient drift $f(x,t) = -\beta(t)x$ (Ho et al., 2020; Song et al., 2021; Karras et al., 2022) that contracts, does no transport, and satisfies detailed balance.

For a drift, the Helmholtz decomposition separates the gradient and divergence-free components, $f = f_{\mathrm{dis}} + f_{\mathrm{rot}}$ with $f_{\mathrm{dis}} = \nabla\phi$ and $\nabla \cdot f_{\mathrm{rot}} = 0$. Standard diffusion uses only the linear dissipative component. We ask what the missing divergence-free component contributes to the marginal flow, the path measure, and the score-estimation problem.

We study this question in the exactly solvable family $f(x,\tau) = -\gamma x + r(\tau)J_0x$, $J_0^\top = -J_0$ (the drift of equation 1 below). The symmetric part is a gradient field; the skew-symmetric part is divergence-free and norm-preserving — the canonical linear form of the Helmholtz split (Proposition 5), with $\rho := \omega/\gamma$ the rotation–dissipation ratio.

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.

What we find. Under isotropic dissipation, a linear skew-symmetric drift is a *gauge freedom*: an explicit time-dependent change of coordinates — rotation, rescaling, and time reparametrization — reduces the Fokker–Planck equation to the standard heat equation, so intermediate marginals differ from their non-rotating counterparts only by $R(\tau)$ and the endpoint generative law is unchanged (Proposition 6, Corollary 1; verified against direct simulation in Section 6). This is stronger than the trivial restoration of p_0 by an exact reverse process: since marginals differ only by $R(\tau)$, score fields, training objectives, and discretization errors coincide — even an imperfect learner cannot detect the linear rotation. The rotational content instead remains in the path measure: at stationarity, the process carries a nonzero probability current with entropy production rate $\sigma = \rho^2 \|J_0\|_F^2 = 2\lfloor d/2 \rfloor \rho^2$ (Proposition 1).

Under model misspecification, rotation becomes visible through a single-peaked score-bias window. Its position is set by the rotation–survival trade-off of the forward process — weakly decreasing in ρ, shifted later by multimode geometry — whereas its depth grows with ρ and is gated by which geometric feature (covariance anisotropy or multimode structure) survives longest. The conformal criterion $F + F^\top \propto I$ delimits where the gauge ceases to rule out rotational effects (Section 8).

Local versus global score information. Denoising score matching recovers $\nabla \log p_t$ as a conditional expectation over the *entire* data set; the Fokker–Planck equation is *local*. Bai et al. (2026) develop the local alternative — mean-value averages over backward heat balls, with exact representations of density, score and entropy — for a *general* drift f; the price is an inhomogeneous source term for nonlinear f.

Scope: where J acts. J acts on the model's state space, not on physical space, and the forward process is a modeling choice, not an observation; the Helmholtz decomposition motivates the question, which we answer as one of forward-process design.

Contributions.

C1. A canonical Helmholtz reading of the linear forward drift (Proposition 5), an exact caloric reduction for every skew profile $J(\tau)$ (Proposition 6), and, for isotropic diffusion, the conformal criterion $F + F^\top \propto I$ (Proposition 7).

C2. A gauge-invariance statement (Corollary 1) with its exact path-level complement $\sigma = \rho^2 \|J_0\|_F^2 = 2\lfloor d/2 \rfloor \rho^2$ (Proposition 1), verified by trajectory estimation at $d = 256$.

C3. A misspecification analysis: the blind-bias *visibility window*, radial invisibility (Lemma 1), the window law (Proposition 2), the exact π-zero (Proposition 3), block reduction (Proposition 4), and the window-to-speciation comparison.

C4. Experimental verification in 2D across five geometries and three skew-symmetric rotation profiles, and at $d = 256$ (Section 7).

C5. A derivative-free first-moment score estimator for the heat-ball representation with a complete error budget ($O(r^2)$ bias, $\sqrt{d/h}$ Monte-Carlo factor; Section 2.1, Appendix E).

2 SETUP

Consider the Itô process on $\mathbb{R}^d$

$$dx = \big(-\gamma I + J(\tau)\big)x\, d\tau + \sqrt{2}\, dW_\tau, \qquad J(\tau) = r(\tau) J_0, \tag{1}$$

with $J_0 = \mathrm{blockdiag}(J_2, \ldots, J_2)$ (for odd d append 0), $J_2 = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. $\|J_0\|_F^2 = 2\lfloor d/2 \rfloor$ (d for even d, $d-1$ for odd d). We work in cumulative-variance time τ and normalize $\gamma = 1$, so that $\mathcal{N}(0, I)$ is stationary. Writing ω for the scale of r, the rotation–dissipation ratio is $\rho := \omega/\gamma$ and the const profile is $r(\tau) \equiv \rho$. The density $u(x, \tau)$ solves $\partial_\tau u = \Delta u - \nabla \cdot \big[(-I + r(\tau) J_0) x u\big]$.

2.1 THE LOCAL HEAT-BALL FRAMEWORK

This subsection reviews the framework of Bai et al. (2026), except the first-moment estimator equation 5, derived in Appendix E; the κ-sampler is used as an off-the-shelf baseline throughout.

Reduction and caloric structure. A cumulative-variance time change reduces a variable-coefficient Fokker–Planck equation to a standard (in general inhomogeneous) heat equation (Bai et al., 2026,

Sec. 2); for the linear drift equation 1 the source term vanishes identically (Proposition 6), and the safe window of the reduction shrinks like d^{-1} (Bai et al., 2026, Thm. 1).

Heat balls and the mean-value representation. Let $\Gamma(z,s) = (4\pi s)^{-d/2} e^{-|z|^2/4s}$ be the heat kernel and, for $r > 0$, let

$$E_r(z_0, s_0) = \Big\{(z,s) \;:\; s < s_0,\;\; \Gamma(z_0 - z,\, s_0 - s) > r^{-d}\Big\} \tag{2}$$

be the backward heat ball. Watson's mean-value theorem Watson (1973) states that a caloric w satisfies $w(z_0,s_0) = \iint_{E_r} w\, d\kappa$, where the κ-*measure* has density

$$d\kappa \;=\; r^{-d} \left|\nabla_z \log \Gamma(z_0 - z, s_0 - s)\right|^2 dz\, ds \;=\; \tfrac{1}{4r^d}\, \tfrac{|z_0 - z|^2}{(s_0 - s)^2}\, dz\, ds, \tag{3}$$

a probability measure on E_r ($w \equiv 1$ is caloric). Bai et al. (2026) extend this to exact local mean-value representations for the density u, the log-density $v = \log u$, the score ∇v, and the entropy density $u \log u$ (Bai et al., 2026, Thm. 5, Cor. 6). Writing $\psi = \log w$ and differentiating under the integral sign gives the *self-normalized* score representation

$$\nabla \log w(z_0, s_0) = \frac{\iint_{E_r} \nabla w\, d\kappa}{\iint_{E_r} w\, d\kappa} \;\approx\; \sum_i \omega_i\, \nabla\psi(z_i, s_i), \qquad \omega = \mathrm{softmax}\big(\psi(z_i, s_i)\big), \tag{4}$$

with $(z_i, s_i) \sim \kappa$ — a local PDE average, not a global conditional expectation (Bai et al., 2026, Thm. 8).

Derivative-free first-moment estimator. For settings in which $\nabla\psi$ is unavailable or expensive, we derive a first-moment estimator built only from function values,

$$S = \tfrac{1}{m^2 r} \cdot \frac{\sum_i z_i\,(R_i^+ - R_i^-)}{\sum_i (R_i^+ + R_i^-)}, \qquad R_i^{\pm} = \exp\big(\psi(y \pm r z_i,\, s_i) - \psi(y, s)\big), \tag{5}$$

where m^2 is the per-coordinate second moment of the κ-sampled directions and h is the number of directions. The estimator requires no derivatives of ψ; it is consistent, with discretization bias $O(r^2)$ and an $O(\sqrt{d/h})$ Monte-Carlo factor (Appendix E).

Training losses and frame independence. The network ψ_θ is trained with the constraint losses of Bai et al. (2026): an anchor term, a caloric heat-ball residual, and a score heat-ball residual enforcing equation 4. All rotation in this paper enters through coordinate changes; the sampler is never modified.

3 Theory

Under isotropic dissipation the rotational drift is a pure gauge. With $R(\tau) = \exp(\theta(\tau) J_0)$, $\theta(\tau) = \int_0^\tau r(s)\, ds$, the time-dependent change of coordinates $z = e^\tau R(\tau)^\top x$ (a rotation composed with a rescaling), $s = (e^{2\tau} - 1)/2$, and $w(z,s) = e^{-d\tau} u(x,\tau)$ reduces the Fokker–Planck equation of equation 1 to the standard heat equation $\partial_s w = \Delta_z w$. The reduction needs no commutativity in time: it holds for every skew profile $J(\tau)$, with R replaced by the time-ordered exponential $O' = J(\tau)O$, $O(0) = I$ (Proposition 6 in Appendix A). More generally, a linear drift $F(\tau)$ with diffusion $Q(\tau)$ admits such a reduction if and only if $Q' - (FQ + QF^\top) = c(\tau)Q$ (Proposition 7). When $Q(\tau) \propto I$, this reduces to $F + F^\top \propto I$, and only in this isotropic case does the antisymmetric part drop out of the criterion.

Corollary 1 (Gauge invariance)**.** *For radial data the marginal $u(\cdot,\tau)$ is independent of J; in general $u^J(\cdot,\tau) = R(\tau)_{\#} u^0(\cdot,\tau)$. The rotational content lives in the pullback $x = e^{-\tau} R(\tau) z$ and the score transport $\nabla_x \log u = e^\tau R(\tau) \nabla_z \log w$: the reverse-time SDE is the $\rho = 0$ law composed with an orthogonal map that equals the identity at $\tau = 0$; since the reverse process starts from the invariant base law, for which $R(T)_{\#} p_T = p_T$, the endpoint generative law coincides with the $\rho = 0$ law. Sampling starts in practice at a finite horizon T from the invariant base law; since p_∞ is radial and $R(T)$ orthogonal, this standard truncation is unaffected by the rotation.*

Remark 1 (The rotational gauge is the Killing-field case)**.** *The reduction requires the transport generator $b \cdot \nabla$ of the divergence-free component to commute with Δ, which holds iff b is a Killing field, $b(x) = Jx + c$ with $J^\top = -J$. A differential rotation $b(x) = \omega(|x|^2) J_0 x$ with non-constant ω is divergence-free and norm-preserving, yet has $[b \cdot \nabla, \Delta] \neq 0$ and breaks the gauge (Appendix A.1). Linearity is therefore not a tractability assumption but exactly the class in which the phenomenon exists.*

Proposition 1 (Where the rotational content goes). *Consider equation 1 with $\gamma = 1$ and $J = \rho J_0$. Statements (i)–(ii) refer to the stationary initialization $p_\infty = \mathcal{N}(0, I)$, whereas statement (iii) holds for an arbitrary initial density p_0.*

(i) the stationary probability current is $j_\infty(x) = \rho\,(J_0 x)\,p_\infty(x)$, with current velocity $v(x) = j_\infty/p_\infty = \rho J_0 x$, which is nonzero whenever $\rho \neq 0$;

(ii) the entropy production rate is exactly

$$\sigma(\rho) \;=\; \mathbb{E}_{p_\infty}\big[\|v\|^2\big] \;=\; \rho^2\,\mathrm{tr}\big(J_0^\top J_0\big) \;=\; \|J\|_F^2 \;=\; 2\lfloor d/2\rfloor\rho^2, \tag{6}$$

(reducing to $d\rho^2$ for even d) and it coincides with the relative-entropy rate between the stationary path measure on $[0,T]$ and its time reversal, $D_{\mathrm{KL}}\big(P_{[0,T]} \,\|\, \bar{P}_{[0,T]}\big) = T\,\sigma(\rho)$. In particular $\sigma(0) = 0$: the non-rotating process satisfies detailed balance;

(iii) for any initial density p_0, every time-marginal differs from its $\rho = 0$ counterpart only by the rotation $R(\tau)$, and coincides with it pointwise when p_0 is radial (Corollary 1).

Proof. See Appendix B. □

Remark 2 (The conformal criterion is stronger than commutation). *For isotropic diffusion, write $F = S + J$ with $S = S^\top$ and $J = -J^\top$, so that $F + F^\top = 2S$: the conformal criterion is exactly $S \propto I$. Two one-way implications organise the landscape. First, $S \propto I$ forces $[S, J] = 0$ for every skew-symmetric J: the gauge regime of this paper sits* inside *the commuting case, not equal to it (e.g. in $d = 4$, $S = \mathrm{diag}(-1,-1,-2,-2)$ commutes with a rotation J acting on the first two coordinates but is not conformal). Second, by contraposition, $[S, J] \neq 0$ forces $S \not\propto I$: at fixed isotropic noise, genuinely hypocoercive coupling—rotation acting on anisotropic dissipation (Hwang et al., 1993; Lelièvre et al., 2013)—lies entirely outside the gauge. With anisotropic noise the boundary moves: the general criterion with $Q \not\propto I$ can hold together with $[A, J] \neq 0$, and the reduction persists (Proposition 8). Commutation is thus necessary but not sufficient for the gauge at $D = I$; breaking conformality is the boundary beyond which rotational effects are no longer ruled out, and our experiments verify the strongest (vanishing-commutator) case at the level of trained generators (Table 1).*

Remark 3 (Reading Proposition 1). *The linear skew-symmetric drift thus acts on marginals only through $R(\tau)$ while contributing exactly $2\lfloor d/2\rfloor\rho^2$ of path-measure irreversibility per unit time: invisible in co-moving marginals, but detectable through the probe of Definition 1.*

Remark 4 (An exact reference exponent). *Equation equation 6 gives $\sigma \propto \rho^2$ exactly, while the blind bias grows sub-quadratically ($\approx \rho^{1.4\text{–}1.9}$; Section 6). The two are normalized differently (a path-level rate versus a squared score deficit at fixed τ), so only the trend is comparable: the marginal-level observable recovers only the fraction of the path-level signal that the data geometry exposes (Proposition 3).*

Definition 1 (Rotation-blind analyst). *A* blind *estimator uses the rotation-free forward kernel $p^{\mathrm{bl}}_{\tau|0} = \mathcal{N}(e^{-\tau}x_0,\,(1 - e^{-2\tau})I)$ although the data were generated by equation 1. Its bias is $B(\tau;\rho) = \mathbb{E}_{x\sim u}\big\|\nabla \log u(x,\tau) - \nabla \log p^{\mathrm{bl}}_\tau(x)\big\|^2$, where $\nabla \log p^{\mathrm{bl}}_\tau$ is the score of the* marginal *of the blind forward process at time τ. For Gaussian mixtures B is available in closed form (both scores are mixtures of linear fields; the explicit expressions are equation 15 and equation 16 of Appendix C.1); we report B normalized by the true score magnitude.*

Remark 5 (Why a blind analyst is not a straw man). *B is not a model of error but a* probe*: by Lemma 1 it vanishes on radial data, so its window measures when non-radial structure is present. Its π-zero tests central symmetry rather than modality itself: a centrally symmetric multimodal distribution can retain the zero.*

Lemma 1 (Radial invisibility). *If p_0 is radial, $B(\tau;\rho) \equiv 0$ for all ρ, τ.*

Proof. Radial u is invariant under $R(\tau)$: $R(\tau)_\# u = u$, and $R_\#$ commutes with the rotation-free semigroup; hence $u = p^{\mathrm{bl}}$. □

Proposition 2 (Single-scale window approximation). *For a single centered anisotropic scale $\Sigma_0 \neq I$, the amplitude of the score-field discrepancy scales as $\sin(\omega\tau)\,e^{-2\tau}$ (structure rotated while it survives the OU contraction), so the bias of Definition 1, being its square, obeys*

$B(\tau) \propto \sin^2(\omega\tau)\, e^{-4\tau}$*; maximizing gives the window position*

$$\tan(\omega\tau^*) = \tfrac{\omega}{2} \quad \Longleftrightarrow \quad \tau^*(\rho) = \tfrac{\arctan(\rho/2)}{\rho}, \tag{7}$$

with the large-ρ limit $\tau^* \sim \pi/(2\rho)$. The estimate is weakly *decreasing in ρ (a factor* 1.26 *over $\rho \in [0.3, 2]$), and multimode geometries shift the window later; both features are confirmed in Section 6. Derivation and assessment against the exact bias: Appendix C.2.*

Proposition 3 (π-zero under central symmetry)**.** *Suppose d is even. At $\theta = \pi$ ($\tau = \pi/\omega$), one has $R(\pi) = -I$, and therefore $B = 0$ for every initial law invariant under $x \mapsto -x$. This includes centered Gaussian data and centrally symmetric multimodal mixtures. Generic asymmetric multimodal data need not retain the zero and may leave a residual (proof: Appendix C.3).*

Proposition 4 (Block reduction; the window of a mixture)**.** *Let $J = \mathrm{blockdiag}(\omega_1 J_0, \ldots, \omega_m J_0)$ on invariant planes and let p_0 factorize across them. Then the blind bias decomposes exactly, $B(\tau) = \sum_k B_k(\tau)$, and the normalized bias is the $\|\nabla \log u_k\|^2$-weighted average of the per-block biases. Consequently: (i) if all blocks carry identical structure and rates, the window position and depth equal those of the planar problem, for every d — the high-dimensional problem is a weighted mixture of planar problems and nothing more; (ii) blocks at a common speed contribute windows at a common position (the mixture deepens but does not shift); blocks at different speeds contribute displaced windows, and in all tested cases the mixture peak lies between the component maxima and is dominated by the deepest block. Both clauses are verified numerically in R5 (proof: Appendix D).*

Remark 6 (Exact bias on real data)**.** *For any data set, taking the empirical measure as p_0 makes both forward marginals Gaussian mixtures with closed-form scores, so the blind bias is estimable without training and without time discretization at any d; Section 7 uses this on MNIST.*

4 Related work

Non-reversible drift and accelerated convergence. Adding an antisymmetric perturbation to a reversible diffusion accelerates convergence without changing the invariant measure Hwang et al. (1993; 2005); for Gaussian targets, Lelièvre et al. (2013) characterize the optimal non-reversible linear drift and give a constructive procedure attaining the maximal spectral gap $\mathrm{tr}(U)/d$, with related variance-reduction results in Duncan et al. (2016); Rey-Bellet & Spiliopoulos (2015) and a complete parametrization of invariant-measure-preserving dynamics in Ma et al. (2015). Closest to the sampling side, Futami et al. (2020; 2021) accelerate diffusion samplers with skew-symmetric drift perturbations — asking *how much* a rotation accelerates; we ask the prior question of *what it can change* in a generative forward process, where the marginal flow, not only the invariant law, is the object of interest. Our isotropic setting is the degenerate case of this theory: $[\gamma I, J] = 0$ leaves no anisotropic structure for J to act on, which is why the spectral gap — and, by Corollary 1, the endpoint generative law — is independent of ρ.

Non-reversible and phase-space forward processes in generative models. Closest to our setting is Lu & Tang (2026), who decompose the forward OU drift as $A = (I + Q)U$ with $U \succ 0$ symmetric and Q antisymmetric, construct an exponentially optimal Q following Lelièvre et al. (2013), and analyze the effect on the speciation and collapse transitions of Biroli et al. (2024); they report that suitable non-reversible perturbations advance speciation while collapse is invariant under antisymmetric perturbation. Proposition 6 shows that under isotropic dissipation the marginals change only by the pushforward $R(\tau)$, so within the linear families covered by our criterion the acceleration of Lu & Tang (2026) originates in the anisotropy of U rather than in Q per se; Proposition 1 then explains the insensitivity of trace-controlled quantities to Q. Section 6 compares our visibility window against their speciation criterion. Rotation also appears, not as a design parameter, in phase-space diffusions Dockhorn et al. (2022); Sterling et al. (2025); Chen et al. (2024); Rojas et al. (2025); Pandey & Mandt (2023); Peluchetti (2023) bypasses time reversal altogether, and Li (2024) replaces the Gaussian driving noise itself — a direction orthogonal to the drift decomposition studied here.

Anisotropic and degenerate forward processes. That the forward process is a genuine design degree of freedom was established by Bansal et al. (2023), who invert arbitrary deterministic degradations. Forward processes with a diagonal but non-scalar drift ($J = 0$) are well studied: blurring diffusion Hoogeboom & Salimans (2023), inverse heat dissipation Rissanen et al. (2023), and score matching under general linear corruptions Daras et al. (2023). Hoogeboom & Salimans (2023) in particular show an equivalence with standard diffusion under a change of variables — the anisotropic

analogue of the gauge proved here; Section 8 delimits the reach of such equivalences through the conformal criterion. A related invariance appears in the training objective: Kingma & Gao (2023) show that standard diffusion objectives equal a weighted ELBO invariant under a family of forward reparametrizations; our gauge is the corresponding invariance of the forward drift's rotational component.

Local mean-value representations. The heat-ball mean-value method goes back to Pini (1951); Watson (1973) and extends to variable-coefficient parabolic Fabes & Garofalo (1987); Garofalo & Lanconelli (1989) and degenerate Kolmogorov operators Kupcov (1980); Garofalo & Lanconelli (1990); Lanconelli & Polidoro (1994); see Anceschi & Polidoro (2020). The construction is always the same: super-level sets of the fundamental solution weighted by $\langle A\nabla\Gamma, \nabla\Gamma\rangle/\Gamma^2$.

Statistical physics of generative diffusion. Biroli et al. (2024); Biroli & Mézard (2023) identify the speciation and collapse transitions, refined as spontaneous symmetry breaking by Raya & Ambrogioni (2023); our visibility window is a different observable but, as Section 6 shows, is related to the marginal symmetry-breaking time scale in a well-defined regime.

Entropy production and probability currents. A stationary probability circulation underneath a time-independent marginal is the classical *irreversible circulation of fluctuation* of Tomita & Tomita (1974); the current-velocity terminology $v = j/p$ follows Nelson (1966), and the time-reversal construction behind Lemma 2 goes back to Anderson (1982); Föllmer (1985); Haussmann & Pardoux (1986); see Cattiaux et al. (2023) for a modern treatment. Estimating entropy production from high-dimensional data is established Seifert (2012); Kim et al. (2020); Otsubo et al. (2020); Boffi & Vanden-Eijnden (2024) use score estimation for this purpose in active matter. Proposition 1 places our forward process in this language and supplies a closed-form ground truth $\sigma = \|J\|_F^2$ for it.

Generative models for fluids. Rotational structure also arises in generative models for fluids and turbulence Li et al. (2024; 2026); these applications are outside our scope.

5 Experimental protocol

Datasets. We consider five 2D densities: iso_gauss (radial control), aniso_gauss (single mode, covariance ratio 5:1; pure covariance gating), asym_gmm (mixed), sparse_dense_gmm (pure multimode geometry), and ring (tangential structure). scaled_gmm (mode distance halved) probes the structure scale $\Lambda = \log(|\mu|/\sigma)$. A single-mode determinant-one family ($\Sigma_0 = \mathrm{diag}(\kappa, 1/\kappa)$, $\kappa \in \{2, 5, 20\}$) varies the anisotropy over two orders of magnitude at fixed volume (the controlled scan of R6). For d=256 we use MNIST16 (MNIST downsampled to 16×16, pixels standardized, floor 0.05), whose between-class scatter has $\lambda_{\max}(\Sigma_B) = 10.7$.

Training. A log-density network $\psi_\theta(z, s)$ is trained with the heat-ball constraint losses of Bai et al. (2026) for 6000–10^4 steps in 2D (4000 steps at d=256) with fixed seeds; the κ-sampler is used unchanged. Profiles: const ($r \equiv \omega = \rho$), osc ($r = \omega \sin k\tau$), ramp ($r = \omega \tanh k\tau$), $k = 3$; comparisons across ρ use a matched step budget. To test Proposition 6 independently of the implementation, we retrain on data from direct Euler–Maruyama simulation of equation 1 (R1). Hyperparameters, seeds and compute: Appendix G.

6 Results in 2D

R1. Rotation is a gauge freedom (trained-pipeline check). In the reduced pipeline ω does not enter the training loop (Appendix G.2), so agreement across $\rho \in \{0.3, 1, 2\}$ and all three profiles holds *by construction* and confirms only correct implementation of the reduction. The substantive, non-circular test is the direct-simulation control — data from Euler–Maruyama integration of equation 1 in x — which trains the *same* pipeline on data that physically carry the rotation and reproduces the closed-form pipeline on all distribution-level metrics (Table 1). Its elevated held-out score MSE at 2D is a memorization artifact of the control's fixed simulated table: enlarging the table to 10^4 trajectories brings the same seed from 0.306 to 0.136, inside the closed-form band $[0.105, 0.145]$ (Appendix G.3).

R2. The visibility window is set by the data, only weakly by ρ. The blind bias is single-peaked in τ for every ρ (Figure 1a): measured peak positions are $\tau^* = 0.66/0.56/0.46$ for

Table 1: Gauge invariance of the trained pipeline. 2D: sparse_dense_gmm, 6000 steps; MNIST16: best sliced-W over training. In 2D the direct-simulation row is better on MMD and sliced-W and worse on score MSE, with no systematic direction, as Corollary 1 requires; at d=256 it is mildly worse throughout, consistent with the memorization artifact of Appendix G.3.

setting	2D			MNIST16 (d=256)		
	MSE@τ=0.3	MMD	slic. W	slic. W	gen std	nn ratio
$\omega = 0$, closed form	0.101	0.00254	0.152	0.1887	0.634	1.063
$\omega = 2$, closed form	0.105	0.00254	0.152	0.1897	0.628	1.060
$\omega = 2$, direct sim	0.124	0.00200	0.137	0.1965	0.646	1.098

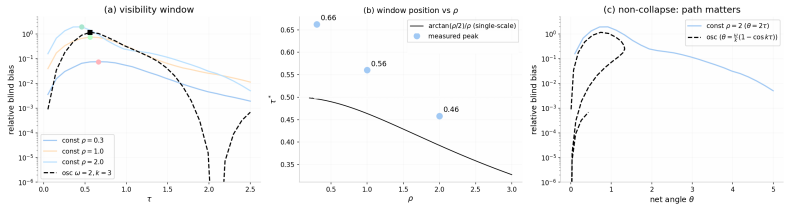

Figure 1: The visibility window in 2D. (a) Blind bias against τ for $\rho \in \{0.3, 1, 2\}$ (const) and osc (ω=2, k=3). (b) Peak position against ρ, with equation 7 as a trend line. (c) Bias against net angle θ: the const and osc curves do not collapse.

$\rho = 0.3/1/2$ — a 6.7× change in ρ moves the window by only 1.4× — while the peak depth grows $7.5\% \to 75\% \to 194\%$ (the three headline points alone give exponent 1.72); a ten-point scan gives depth $\propto \rho^{1.61}$ (multimode) and $\rho^{1.84}$ (covariance-gated), sub-quadratic against the exact path-level exponent 2 (Remark 4; scan details: Appendix G.4). The single-scale estimate equation 7 predicts $0.50/0.46/0.39$ — the right order and weak trend, but systematically early by 17–33%, consistent with the multimode shift of Proposition 2; we claim no ρ-law for the position. The operative statement: *the window position is set by the rotation–survival trade-off of Proposition 2, hence weakly decreasing in ρ and modulated by the data geometry by at most a factor of two, while the depth grows with ρ* (made quantitative in R6).

R3. Path effects: the window does not reopen. Aligning the osc run ($\omega = 2$) with const by net angle θ (Figure 1c), the curves do *not* collapse: at equal θ the oscillating profile carries less bias, having accumulated angle after more structure decayed. At $\theta = 0$ the bias is 5×10^{-12} (exact unwind); the second θ-hump produces only 10^{-4}–10^{-3} — **visibility is effectively one-way**: once the OU contraction erases the non-radial structure, rotation can no longer expose itself.

R4. Mechanism taxonomy and the π-zero. At $\rho = 2$, covariance-gated aniso_gauss peaks at $\tau^* = 0.36$ with 55% depth and exhibits the π-zero of Proposition 3: the value at the grid point nearest $\theta = \pi$ is 3.6×10^{-6}, the resolution floor of the $\Delta\tau \approx 0.05$ scan (the exact value is zero; Appendix C.3); the asymmetric multimode geometry sparse_dense keeps 2.4% there because sign reversal does not preserve its mixture law (Figure 5). Depth ordering: covariance 55% < mixed 163% < multimode 194%; halving Λ closes the window earlier ($\tau^* = 0.25$). Lemma 1: iso_gauss bias $\leq 5 \times 10^{-32}$ at $\rho \in \{1, 2\}$.

R5. Block heterogeneity: position is robust, depth mixes. Mixing two covariance-gated blocks (anisotropies 5:1 and 20:1) at the same speed ($\rho = 2$) leaves the window position unchanged ($\tau^* = 0.41$) while the depth becomes the $\|\nabla \log u_k\|^2$-weighted average of the components (4.39, from 1.34 and 7.57); rotating them at different speeds ($\rho_1 = 2$, $\rho_2 = 0.6$; single-block windows at 0.41 and 0.56) yields a single peak at the faster–deeper block's position, within the convex hull of the block windows. The high-dimensional window is thus a weighted mixture of planar windows, and τ^* is insensitive to the anisotropy ratio at fixed volume (used in R6).

R6. Window position and speciation time are independent scales that cross at $\lambda_{\max} \approx 1.3$. As a secondary connection we compare τ^* against the *speciation time* t_S of Biroli et al. (2024) in the form of Lu & Tang (2026), which here collapses to $t_S = \frac{1}{2}\ln(1 + \lambda_{\max}(\Sigma_B))$ (Lemma 3); rotation drops out of the spectrum, so t_S is ρ-independent.

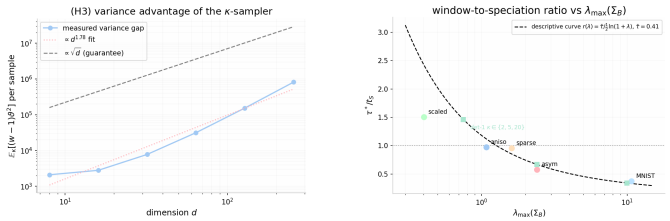

Figure 2: (a) Variance advantage of the κ-sampler over uniform importance sampling; measured slope 1.78 against the $\Omega(\sqrt{d})$ guarantee (vertical offset arbitrary). (b) Window-to-speciation ratio against $\lambda_{\max}(\Sigma_B)$ with the curve $\bar{\tau}/\frac{1}{2}\log(1+\lambda)$, $\bar{\tau} = 0.41$; squares: the determinant-one family.

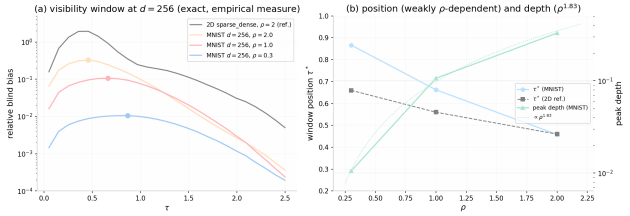

Figure 3: Visibility window at $d = 256$ (MNIST16), computed exactly from the empirical measure (Remark 6). (a) Blind bias for $\rho \in \{0.3, 1, 2\}$, peaks marked; grey: 2D multimode reference (ρ=2). (b) Peak position and depth against ρ; depth exponent 1.83.

Across the eight non-radial geometries (the four non-radial base densities, scaled_gmm, and the three determinant-one members) τ^* varies by less than a factor 2 (0.25–0.46) while t_S varies by a factor 7 (0.17–1.23), so their ratio follows the single-parameter curve $r(\lambda) = \bar{\tau}/\frac{1}{2}\ln(1+\lambda)$, $\bar{\tau} = 0.41$ fitted to the eight ratios (Figure 2). The curve is descriptive, not a law: its content is the near-constancy of τ^* itself; individual geometries depart from it (scaled_gmm by $\approx 38\%$, measured 1.50 against 2.43; the rest by at most 15%), with nothing fitted beyond $\bar{\tau}$. The determinant-one family localizes the mechanism: t_S grows $0.280 \to 1.198$ across $\kappa \in \{2, 5, 20\}$ while the window's grid argmax stays at the fine-grid point 0.407 for all three — the underlying continuous peaks $0.395/0.404/0.418$ move by less than half a grid step — three points on one curve traced by a single controlled parameter. At grid resolution these three ratios coincide with $\bar{\tau}/t_S$ by construction and add no independent evidence for the curve; evaluated at the continuous peaks they depart from it by at most 4%. The equality $\tau^* = t_S$ is therefore a *crossing* at $\lambda_{\max} \approx 1.3$, not a band. MNIST16 sits at $\lambda_{\max} = 10.7$, $\tau^* = 0.458 = 0.37\, t_S$, on the same curve (Table 2, Appendix F).

7 Scaling to $d = 256$

Three dimension-dependent claims of Sections 3–6 — the window law, the $\Omega(\sqrt{d})$ sampler variance advantage, and $\sigma = 2\lfloor d/2 \rfloor \rho^2$ — are verified on MNIST16 ($d = 256$).

H1. The gauge holds at scale. At d=256 the $\omega = 0$ and $\omega = 2$ reduced-pipeline runs are bit-identical by construction, and final metrics agree within the ∼1–2% MPS nondeterminism (Table 1); the direct-simulation control is the non-circular, end-to-end check (Appendix I).

H2. The visibility window at $d = 256$. Computed exactly from the empirical measure (Remark 6), the window is single-peaked for every ρ: its position moves only 1.9× under a 6.7× change of ρ ($0.87 \to 0.46$), while depth grows as $\rho^{1.83}$ (Remark 4); Figure 3, Table 3.

H3. Sampler variance advantage. The previously established $\Omega(\sqrt{d})$ variance advantage of the κ-sampler persists at d=256 (Figure 2a).

H4. Entropy production at scale. Simulating equation 1 at stationarity (M=5000 trajectories, Heun scheme, dt=2×10^{-3}) and estimating $\widehat{\sigma} = \frac{1}{T}\int_0^T \rho J_0 x_t \circ dx_t$ recovers $\sigma = 2\lfloor d/2 \rfloor \rho^2$ up to a sign-consistent +0.15–0.31% offset: $\widehat{\sigma} = 64.20 \pm 0.26$, 256.39 ± 0.58, 1025.6 ± 1.5 against 64, 256, 1024 at $\rho = 0.5, 1, 2$, compatible with the Heun discretization bias at this dt (consistency:

Remark 10). The gauge parameter is thus an observable growing linearly in d, measurable from raw trajectories alone.

8 DISCUSSION

Outlook: beyond the conformal class. For isotropic diffusion the obstruction to the caloric reduction is the conformal condition $F + F^\top \propto I$ (Proposition 7), to which rotation never contributes: no choice of $J(\tau)$ can change the endpoint generative law. With anisotropic noise the condition becomes

$$FD + DF^\top = -cD, \qquad F = -A + J, \tag{8}$$

and here more is true: equation 8 is equivalent to $D^{-1/2} F D^{1/2} = -\mu I + K$ with $\mu = c/2$, $K^\top = -K$, so $y = D^{-1/2}x$ obeys equation 1 and K is gauged away in y — the whole conformal class remains a gauge, with gauge map $D^{1/2}O(\tau)D^{-1/2}$ and closed-form propagator and covariance (Proposition 8). Co-designing D through equation 8 thus stays *inside* the gauge; rotational effects on the generative law require violating equation 8, where no closed-form kernel exists and the heat-ball source term (Section 2.1) becomes essential. In the eigenbasis of $D \neq I$, equation 8 pins J from A and $\mathrm{spec}(D)$ and equalizes the eigenvalue decay rates of F at $\mu = \mathrm{tr}(A)/d$, matching the optimal rate of Lelièvre et al. (2013). There is a sharper reading of this match: chaining the conjugation above with Proposition 6, the entire conformal co-design family is, up to the fixed preconditioning $x \mapsto D^{-1/2}x$ and an orthogonal frame, exactly the isotropic Ornstein–Uhlenbeck process at rate μ — the optimal decay rate is thus attained *inside* the gauge, and against preconditioned isotropic OU the family buys no new generative capability, only a change of coordinates.

What the answer to the Helmholtz question is. Under isotropic dissipation, linear skew-symmetric drift is a marginal gauge but produces path-space irreversibility at rate $2\lfloor d/2 \rfloor \rho^2$. Rotational effects on the generative law therefore require leaving the conformal linear regime.

What the practitioner gets. (i) *Detection and diagnosis*: the blind-bias curve is an instrument — its depth scales with ρ, its window locates the time scale of surviving non-radial structure, and its π-zero tests central symmetry. (ii) *A calibrated ground truth*: equation 6 gives the entropy production rate in closed form under stationary initialization, directly usable for validating local entropy-production estimators (Boffi & Vanden-Eijnden, 2024). (iii) *A negative design guideline*: a *linear* rotation added to any forward process in the conformal class equation 8 — isotropic noise included — cannot change the endpoint generative law, because the whole class is one isotropic Ornstein–Uhlenbeck process at rate μ seen through a fixed preconditioning and an orthogonal frame (Proposition 8); reported acceleration from non-reversible drift must be attributed to a symmetric part outside the conformal class. It does not extend to nonlinear divergence-free drifts (Remark 1).

Limitations. Two scope conditions are essential and independent. Isotropic dissipation: anisotropic A blocks the caloric reduction unless the diffusion matrix is co-designed as in equation 8. Linearity of the rotational field: the present caloric gauge holds exactly for Killing fields $b(x) = Jx + c$ and fails for nonlinear divergence-free drifts, including norm-preserving ones (Remark 1). The window law equation 7 is a single-scale estimate, systematically early for multimode data and reported as a trend; no law is fitted to the window position, and $\tau^* = t_S$ is claimed only at the crossing $\lambda_{\max} \approx 1.3$. We include no DSM/DDPM comparison and claim no improved sample quality or training efficiency: the experiments test the structural invariance and path-level predictions, using the baseline Bai et al. (2026) unchanged. Two gaps remain: the trained-model checks (R1, H1) include no trained-pipeline *positive control* (the closed-form version, a deliberately non-conformal drift the instrument must detect, is verified in Appendix H), and the Killing boundary of Remark 1 is verified analytically but not numerically. Window positions are grid argmax on two nested grids (headline scans at $\Delta\tau \approx 0.1$; R5 and the determinant-one family on the fine grid $\Delta\tau \approx 0.05$, so the quoted 0.41 is the grid point 0.407) and depth exponents are unweighted ten-point log-log fits (Appendix G.4); both conclusions are qualitative.

AI USE STATEMENT

AI-based assistants were used for code prototyping, numerical debugging, LaTeX typesetting, and language editing; all mathematical claims, proofs, experiments, and their interpretation were developed and verified by the authors, who take full responsibility.

REFERENCES

Francesca Anceschi and Sergio Polidoro. A survey on the classical theory for Kolmogorov equation. *Le Matematiche*, 75(1):221–258, 2020. doi: 10.4418/2020.75.1.11.

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. doi: 10.1016/0304-4149(82)90051-5.

Jiayao Bai, Lang Deng, Yi Du, and Yifei Jia. Local Fokker–Planck geometry for score estimation: Heat-ball mean-value representations and exact high-dimensional sampling. *arXiv preprint arXiv:2606.27954*, 2026.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 35, 2023.

Giulio Biroli and Marc Mézard. Generative diffusion in very large dimensions. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(9):093402, 2023. doi: 10.1088/1742-5468/acf8ba.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024. doi: 10.1038/s41467-024-54281-3.

Nicholas M. Boffi and Eric Vanden-Eijnden. Deep learning probability flows and entropy production rates in active matter. *Proceedings of the National Academy of Sciences*, 121(25):e2318106121, 2024. doi: 10.1073/pnas.2318106121.

Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 59(4):1844–1881, 2023. doi: 10.1214/22-AIHP1320.

Tianrong Chen, Jiatao Gu, Laurent Dinh, Evangelos Theodorou, Joshua M. Susskind, and Shuangfei Zhai. Generative modeling with phase stochastic bridge. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tUtGjQEDd4.

Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G. Dimakis, and Peyman Milanfar. Soft diffusion: Score matching for general corruptions. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=W98rebBxlQ.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped Langevin diffusion. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=CzceR82CYc.

Andrew B. Duncan, Tony Lelièvre, and Grigorios A. Pavliotis. Variance reduction using nonreversible Langevin samplers. *Journal of Statistical Physics*, 163(3):457–491, 2016. doi: 10.1007/s10955-016-1491-2.

Eugene B. Fabes and Nicola Garofalo. Mean value properties of solutions to parabolic equations with variable coefficients. *Journal of Mathematical Analysis and Applications*, 121(2):305–316, 1987. doi: 10.1016/0022-247X(87)90249-6.

Hans Föllmer. An entropy approach to the time reversal of diffusion processes. In *Stochastic Differential Systems (Lecture Notes in Control and Information Sciences, vol. 69)*, pp. 156–163. Springer, 1985. doi: 10.1007/BFb0005070.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Accelerating the diffusion-based ensemble sampling by non-reversible dynamics. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3337–3347, 2020. URL https://proceedings.mlr.press/v119/futami20a.html.

Futoshi Futami, Tomoharu Iwata, Naonori Ueda, and Issei Sato. Accelerated diffusion-based sampling by the non-reversible dynamics with skew-symmetric matrices. *Entropy*, 23(8):993, 2021. doi: 10.3390/e23080993.

Nicola Garofalo and Ermanno Lanconelli. Asymptotic behavior of fundamental solutions and potential theory of parabolic operators with variable coefficients. *Mathematische Annalen*, 283(2): 211–239, 1989. doi: 10.1007/BF01446432.

Nicola Garofalo and Ermanno Lanconelli. Level sets of the fundamental solution and Harnack inequality for degenerate equations of Kolmogorov type. *Transactions of the American Mathematical Society*, 321(2):775–792, 1990. doi: 10.2307/2001585.

Ulrich G. Haussmann and Étienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986. doi: 10.1214/aop/1176992362.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OjDkC57x5sz.

Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating Gaussian diffusions. *The Annals of Applied Probability*, 3(3):897–913, 1993. doi: 10.1214/aoap/1177005371.

Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating diffusions. *The Annals of Applied Probability*, 15(2):1433–1444, 2005. doi: 10.1214/105051605000000025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Dong-Kyun Kim, Youngkyoung Bae, Sangmin Lee, and Hawoong Jeong. Learning entropy production via neural networks. *Physical Review Letters*, 125(14):140604, 2020. doi: 10.1103/PhysRevLett.125.140604.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Leonid P. Kupcov. On parabolic means. *Doklady Akademii Nauk SSSR*, 252(2):296–301, 1980.

Ermanno Lanconelli and Sergio Polidoro. On a class of hypoelliptic evolution operators. *Rendiconti del Seminario Matematico Università e Politecnico di Torino*, 52(1):29–63, 1994.

Tony Lelièvre, Francis Nier, and Grigorios A. Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2):237–274, 2013. doi: 10.1007/s10955-013-0769-x.

Henry Li. Non-normal diffusion models. 2024. URL https://arxiv.org/abs/2412.07935.

Tianyi Li, Luca Biferale, Fabio Bonaccorso, Martino Andrea Scarpolini, and Michele Buzzicotti. Synthetic Lagrangian turbulence by generative diffusion models. *Nature Machine Intelligence*, 6 (4):393–403, 2024. doi: 10.1038/s42256-024-00810-0.

Tianyi Li, Michele Buzzicotti, Fabio Bonaccorso, and Luca Biferale. Physics-constrained diffusion model for synthesis of 3D turbulent data. *Nature Communications*, 2026.

Haiqi Lu and Ying Tang. Steering dynamical regimes of diffusion models by breaking detailed balance. *arXiv preprint arXiv:2602.15914*, 2026.

Yi-An Ma, Tianqi Chen, and Emily B. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems 28*, pp. 2917–2925, 2015.

Edward Nelson. Derivation of the Schrödinger equation from Newtonian mechanics. *Physical Review*, 150(4):1079–1085, 1966. doi: 10.1103/PhysRev.150.1079.

Shun Otsubo, Sosuke Ito, Andreas Dechant, and Takahiro Sagawa. Estimating entropy production by machine learning of short-time fluctuating currents. *Physical Review E*, 101(6):062106, 2020. doi: 10.1103/PhysRevE.101.062106.

Kushagra Pandey and Stephan Mandt. A complete recipe for diffusion generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4261–4272, 2023. doi: 10.1109/ICCV51070.2023.00393.

Stefano Peluchetti. Non-denoising forward-time diffusions. 2023. URL https://arxiv.org/abs/2312.14589.

Bruno Pini. Sulle equazioni a derivate parziali, lineari del secondo ordine in due variabili, di tipo parabolico. *Annali di Matematica Pura ed Applicata. Serie Quarta*, 32(1):179–204, 1951. doi: 10.1007/BF02417958.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 66377–66389, 2023. URL https://openreview.net/forum?id=lxGFGMMSVl.

Luc Rey-Bellet and Konstantinos Spiliopoulos. Irreversible Langevin samplers and variance reduction: a large deviations approach. *Nonlinearity*, 28(7):2081–2103, 2015. doi: 10.1088/0951-7715/28/7/2081.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *The Eleventh International Conference on Learning Representations*, 2023.

Kevin Rojas, Yixin Tan, Molei Tao, Yuriy Nevmyvaka, and Wei Deng. Variational Schrödinger momentum diffusion. *arXiv preprint arXiv:2501.16675*, 2025.

Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, 2012. doi: 10.1088/0034-4885/75/12/126001.

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

Benjamin Sterling, Chad Gueli, and Mónica F. Bugallo. Critically-damped higher-order Langevin dynamics for generative modeling. *arXiv preprint arXiv:2506.21741*, 2025.

Kazuhisa Tomita and Hiroyuki Tomita. Irreversible circulation of fluctuation. *Progress of Theoretical Physics*, 51(6):1731–1749, 1974. doi: 10.1143/PTP.51.1731.

Neil A. Watson. A theory of subtemperatures in several variables. *Proceedings of the London Mathematical Society. Third Series*, s3-26(3):385–417, 1973. doi: 10.1112/plms/s3-26.3.385.

APPENDIX

The appendix provides proofs for the principal theoretical statements and records additional experimental details. Appendix A proves the exact caloric reduction of Proposition 6 and the general conformal criterion of Proposition 7; Appendix B proves the path-measure identity of Proposition 1(ii); Appendix C gives the closed form of the blind bias and derives the window law (Proposition 2) and the π-zero (Proposition 3); Appendix D proves the block reduction (Proposition 4); Appendix E derives the derivative-free first-moment estimator equation 5 and establishes its consistency, discretization bias and variance; Appendix F derives the speciation time used in R6; Appendix G records experimental details; Appendix H verifies the closed-form positive control; Appendix I collects additional experimental figures.

A CALORIC REDUCTION

We first establish the linear Helmholtz decomposition, then prove the exact caloric reduction for the rotated Ornstein–Uhlenbeck family, and finally derive the general conformal criterion.

Proposition 5 (Helmholtz decomposition of a linear drift)**.** *Let $f(x) = Mx$ on $\mathbb{R}^d$ and write $M = S + J$ with $S = \frac{1}{2}(M + M^\top)$, $J = \frac{1}{2}(M - M^\top)$. Then*

(i) $Sx = \nabla\Phi(x)$ with $\Phi(x) = \frac{1}{2}x^\top Sx$: the symmetric part is a gradient field, hence curl-free;

(ii) $\nabla \cdot (Jx) = \operatorname{tr} J = 0$ and $x^\top Jx = 0$ for all x: the antisymmetric part is divergence-free and norm-preserving, and generates the one-parameter rotation group $\{e^{\theta J}\} \subset SO(d)$;

(iii) the splitting is the unique *decomposition $M = S' + J'$ in which $S'x$ is a gradient field and $J'x$ is both divergence-free and norm-preserving.*

Proposition 6 (Exact caloric reduction)**.** *Let $J(\tau)$ be any measurable skew-symmetric matrix field, $J(\tau)^\top = -J(\tau)$, and let $O(\tau)$ solve $O' = J(\tau)O$, $O(0) = I$. Then $O(\tau)$ is orthogonal for every τ, and with $z = e^\tau O(\tau)^\top x$, $s = (e^{2\tau} - 1)/2$, and $w(z, s) = e^{-d\tau}u(x, \tau)$, the Fokker–Planck equation associated with equation 1 — now with the general drift $-x + J(\tau)x$ — reduces to*

$$\partial_s w = \Delta_z w \tag{9}$$

for every skew profile $J(\tau)$, time-varying in direction included. For the scalar family $J(\tau) = r(\tau)J_0$ this is $O(\tau) = R(\tau) = \exp(\theta(\tau)J_0)$, $\theta(\tau) = \int_0^\tau r(s)\,ds$.

Proof. Since $(OO^\top)' = J(OO^\top) - (OO^\top)J$ vanishes on $OO^\top = I$, $O(\tau) \in O(d)$ for all τ; no commutativity of $J(\tau)$ at different times is needed, as O is the time-ordered exponential. Write $\varphi(\tau, z) = e^{-\tau}O(\tau)z$ for the inverse of $z = e^\tau O(\tau)^\top x$ and $w(z, s) = e^{-d\tau}u(\varphi(\tau, z), \tau)$ with $\tau = \tau(s)$, $s = \frac{1}{2}(e^{2\tau} - 1)$, so that $\mathrm{d}s/\mathrm{d}\tau = e^{2\tau}$. From $O' = J(\tau)O$, $\partial_\tau\varphi = -\varphi + J(\tau)\varphi$. Therefore

$$\partial_s w = e^{-d\tau - 2\tau}\Big[\partial_\tau u - \varphi \cdot \nabla u + \big(J(\tau)\varphi\big) \cdot \nabla u - d\,u\Big],$$

while orthogonality of the Jacobian $e^{-\tau}O$ gives $\Delta_z w = e^{-d\tau - 2\tau}\,\Delta_x u$. The Fokker–Planck equation $\partial_\tau u = \Delta u - \nabla \cdot [(-I + J)\varphi\, u]$ expands, using $\nabla \cdot \varphi = d$ and $\nabla \cdot [(J\varphi)u] = (J\varphi) \cdot \nabla u$ (since $\operatorname{tr} J = 0$), to $\partial_\tau u = \Delta u + d\,u + \varphi \cdot \nabla u - (J\varphi) \cdot \nabla u$, which makes the two expressions coincide: $\partial_s w = \Delta_z w$. □

Proposition 7 (Conformal criterion, full form)**.** *Consider $dx = F(\tau)x\,d\tau + \sqrt{2Q(\tau)}\,dW_\tau$ with $Q(\tau) \succ 0$, and let Φ be the fundamental matrix, $\Phi' = F\Phi$, $\Phi(0) = I$. Set $y = \Phi^{-1}x$. Then y has zero drift and diffusion matrix*

$$P(\tau) = \Phi^{-1}Q\,\Phi^{-\top}, \tag{10}$$

and the density of y can be reduced to a solution of the standard heat equation by a time reparametrization together with one fixed linear map — within this reduction class, if and only if $P(\tau) = \varphi(\tau)P_0$ for a scalar φ and a constant matrix P_0. This holds if and only if

$$Q' - \big(FQ + QF^\top\big) = c(\tau)\,Q, \qquad c = \varphi'/\varphi. \tag{11}$$

For constant Q this reads $FQ + QF^\top = -cQ$; for $Q \propto I$ it reads $F + F^\top \propto I$, in which the antisymmetric part of F does not appear.

Proof. Since $(\Phi^{-1})' = -\Phi^{-1}\Phi'\Phi^{-1} = -\Phi^{-1}F$,

$$dy = (\Phi^{-1})'x\,d\tau + \Phi^{-1}dx = -\Phi^{-1}Fx\,d\tau + \Phi^{-1}\big(Fx\,d\tau + \sqrt{2Q}\,dW\big) = \Phi^{-1}\sqrt{2Q}\,dW,$$

so the drift vanishes identically and the quadratic variation is $2P(\tau)\,d\tau$ with P as in equation 10. Because P does not depend on y, the density q of y satisfies $\partial_\tau q = \nabla \cdot (P\nabla q)$.

Only two operations are available to normalise this equation: a time reparametrization $ds = \varphi\,d\tau$, which can remove one *scalar* factor, and a fixed linear change of variables $\zeta = P_0^{-1/2}y$, which can normalise one *constant* matrix. (A time-*dependent* linear map $\zeta = O(\tau)y$ is not available: it reintroduces a drift $\dot{O}(\tau)O(\tau)^\top\zeta$, so keeping the drift identically zero forces O to be constant.) Their composition normalises exactly the family $P(\tau) = \varphi(\tau)P_0$; conversely if the direction of P

varies with τ then no single fixed linear map can render every $P(\tau)$ isotropic, and the reduction fails.

It remains to convert $P = \varphi P_0$ into a condition on (F, Q). Differentiating equation 10 and using $(\Phi^{-1})' = -\Phi^{-1}F$ and $(\Phi^{-\top})' = -F^\top \Phi^{-\top}$,

$$P' = \Phi^{-1}\Big[Q' - \big(FQ + QF^\top\big)\Big]\Phi^{-\top}.$$

The requirement $P' = (\varphi'/\varphi)P = c\,\Phi^{-1}Q\Phi^{-\top}$ becomes, after left multiplication by Φ and right multiplication by $\Phi^\top$ (both invertible), exactly equation 11. Specialising to constant Q gives $FQ + QF^\top = -cQ$. Specialising further to $Q = qI$ and writing $F = S + J$ with $S = S^\top$, $J = -J^\top$ gives $q(F + F^\top) = 2qS = -cqI$, i.e. $S \propto I$: the antisymmetric part cancels identically and never enters. □

Remark 7. *Proposition 7 is the precise sense in which rotation "never obstructs" the reduction: J drops out of equation 11 the moment $Q \propto I$, so the existence of the caloric frame is a property of the symmetric part alone. Equation equation 11 with $Q \neq I$ is equation 8 of Section 8; there J re-enters the condition through $[J, Q]$, but satisfying the condition keeps the reduction intact: the whole conformal class remains a gauge (Proposition 8).*

Proposition 8 (The whole conformal class is a gauge)**.** *Let $D \succ 0$ be constant and let F satisfy equation 8 with constant c: $FD + DF^\top = -cD$. Write $\mu = c/2$. Then $G := D^{-1/2}FD^{1/2}$ satisfies $G + G^\top = -cI$, hence $G = -\mu I + K$ with $K^\top = -K$, and $y = D^{-1/2}x$ obeys*

$$dy = \big(-\mu I + K\big)y\,d\tau + \sqrt{2}\,dW_\tau, \tag{12}$$

an instance of equation 1. Consequently Proposition 6 applies in y: the rotation K is a gauge throughout the conformal class, and in x the gauge map is $D^{1/2}O(\tau)D^{-1/2}$ with $O' = KO$, $O(0) = I$. Moreover, with no commutativity requirement on $[A, J]$,

$$\Phi(t) = e^{-\mu t}D^{1/2}e^{tK}D^{-1/2}, \qquad \Sigma(t) = \frac{1 - e^{-2\mu t}}{\mu}\,D, \qquad \Sigma_\infty = \frac{D}{\mu}. \tag{13}$$

A time-varying $c(\tau)$ is absorbed by the cumulative-variance time change of Section 2 before applying the same argument.

Proof. Conjugating equation 8 by $D^{-1/2}$ gives $G + G^\top = -cI$, so the symmetric part of G is $-\mu I$ and $G = -\mu I + K$ with K skew. Since $D^{-1/2}\sqrt{2D} = \sqrt{2}\,I$, the substitution $y = D^{-1/2}x$ gives equation 12, which is equation 1 with $\gamma = \mu$ and $J = K$; the caloric reduction of Proposition 6 therefore gauges K away in y, and $x = D^{1/2}y$ is a fixed K-independent linear map. For the closed forms, $F = D^{1/2}GD^{-1/2}$ gives $\Phi(t) = e^{tF} = D^{1/2}e^{tG}D^{-1/2} = e^{-\mu t}D^{1/2}e^{tK}D^{-1/2}$ because $[-\mu I, K] = 0$; then $e^{sF}D\,e^{sF^\top} = e^{-2\mu s}D^{1/2}e^{sK}e^{-sK}D^{1/2} = e^{-2\mu s}D$, so $\Sigma(t) = \int_0^t 2e^{sF}De^{sF^\top}ds = \frac{1-e^{-2\mu t}}{\mu}D$, with limit $\Sigma_\infty = D/\mu$ (which indeed solves $F\Sigma_\infty + \Sigma_\infty F^\top = -2D$ by equation 8). □

A.1 Scope: why the rotational field must be linear

Proposition 7 characterizes the *dissipative* part. The following characterizes the *rotational* part, and shows that the linear family of equation 1 is not a simplification but the exact boundary of the phenomenon.

Proposition 9 (Killing criterion for the rotational field)**.** *Let b be a smooth vector field on $\mathbb{R}^d$ and consider $dx = \big(-\gamma x + b(x)\big)d\tau + \sqrt{2}\,dW_\tau$. A caloric reduction in the sense of Proposition 6 — a time-dependent change of variables carrying the Fokker–Planck equation to $\partial_s w = \Delta_z w$ — when realized through the co-moving frame of b, requires the transport generator $b \cdot \nabla$ to commute with the diffusion generator Δ. This holds if and only if b is a Killing field of Euclidean space,*

$$b(x) = Jx + c, \qquad J^\top = -J,\ c \in \mathbb{R}^d. \tag{14}$$

In particular, being divergence-free and norm-preserving — the pointwise conditions of Proposition 5(ii) — is not *sufficient.*

Proof. If the flow φ_t of b acts by isometries then $\Delta(f \circ \varphi_t) = (\Delta f) \circ \varphi_t$ for all smooth f; differentiating at $t = 0$ gives $\Delta(b \cdot \nabla f) = b \cdot \nabla(\Delta f)$, i.e. $[b \cdot \nabla, \Delta] = 0$. Conversely, a direct expansion

gives $[b\cdot\nabla,\Delta]f = -(\Delta b)\cdot\nabla f - 2\sum_{ij}(\partial_i b_j)\,\partial_i\partial_j f$; the second-order part vanishes for all f iff the symmetrized Jacobian of b vanishes, which is the Killing equation, and the first-order part then vanishes because $\Delta b_j = \partial_i(\partial_i b_j) = -\partial_i(\partial_j b_i) = -\partial_j(\nabla\cdot b) = 0$. Equivalently, φ_t preserves the Laplace–Beltrami operator and hence the Euclidean metric, so b generates a one-parameter group of isometries and is a Killing field; the Killing fields of $\mathbb{R}^d$ with the Euclidean metric are exactly equation 14. We use the necessity direction only in the co-moving-frame sense stated in the proposition: a general reduction not built from the flow of b is constrained a posteriori by the conformal criterion of Proposition 7. □

Remark 8 (A divergence-free, norm-preserving counterexample)**.** *Take the differential rotation $b(x) = \omega(|x|^2)J_0x$ in $d = 2$. Then $\nabla\cdot b = \nabla\omega\cdot(J_0x) + \omega\,\mathrm{tr}\,J_0 = 0$ because $x^\top J_0 x = 0$, and $x\cdot b = \omega\, x^\top J_0 x = 0$: the field is divergence-free and norm-preserving for every profile ω. It is Killing only for constant ω. For $\omega(r^2) = r^2$ a direct computation gives*

$$[\,b\cdot\nabla,\Delta\,]f = 4(x_2^2 - x_1^2)\,\partial_1\partial_2 f + 4x_1x_2\big(\partial_1^2 - \partial_2^2\big)f + 8\big(x_2\partial_1 - x_1\partial_2\big)f \;\not\equiv\; 0,$$

and similarly for $\omega(r^2) = (1+r^2)^{-1}$. Both therefore fall outside the gauge: the rotation no longer commutes with the diffusion, the co-moving frame is no longer an isometry, and the marginal flow is genuinely altered.

Remark 9 (Two independent scope conditions)**.** *Propositions 7 and 9 are logically independent and together delimit the regime studied here: for the isotropic diffusion considered in the main text, the dissipative part must be conformal ($F + F^\top \propto I$) and the rotational part must be Killing. Failure of either is a route by which rotational effects are no longer ruled out by the present gauge argument; Section 8 pushes the first boundary outward — with anisotropic noise the entire conformal class equation 8 remains a gauge (Proposition 8) — and leaves the second open.*

B Entropy production and path-measure irreversibility

We give the full proof of Proposition 1, including the path-measure identity (ii), so that the paper is self-contained.

Lemma 2. *Let $dx = f(x)\,dt + \sqrt{2}\,dW$ admit the stationary density p, and let $v := f - \nabla\log p$ be the current velocity. The time-reversed stationary process solves $d\bar{x} = \bar{f}(\bar{x})\,dt + \sqrt{2}\,d\bar{W}$ with $\bar{f} = -f + 2\nabla\log p$. Then, for every $T > 0$,*

$$D_{\mathrm{KL}}\big(P_{[0,T]}\,\|\,\bar{P}_{[0,T]}\big) = T\,\mathbb{E}_p\big[\|v\|^2\big].$$

Proof. For two diffusions with common diffusion matrix $a = 2I$ and drifts b_1, b_2, the Girsanov relative-entropy rate on path space is $\frac{1}{2}(b_1-b_2)^\top a^{-1}(b_1-b_2) = \frac{1}{4}\|b_1-b_2\|^2$. Here $b_1 - b_2 = f - \bar{f} = 2\big(f - \nabla\log p\big) = 2v$, whence the rate is $\frac{1}{4}\|2v\|^2 = \|v\|^2$. Averaging over the stationary law and integrating over $[0,T]$ gives the claim. □

Proof of Proposition 1. (i) With $D = I$ the current is $j = fp - \nabla p$. For $p = p_\infty \propto e^{-|x|^2/2}$ we have $\nabla p_\infty = -x\,p_\infty$, hence $j_\infty = (-x + \rho J_0 x)p_\infty + x\,p_\infty = \rho(J_0x)p_\infty$. It is divergence-free: $\nabla\cdot j_\infty = \rho\,\mathrm{tr}(J_0)p_\infty + \rho(J_0x)\cdot\nabla p_\infty = -\rho\,(x^\top J_0 x)\,p_\infty = 0$, consistent with stationarity. (ii) By part (i), $v(x) = \rho J_0 x$ and $p_\infty = \mathcal{N}(0, I)$, so

$$\sigma = \int \tfrac{|j_\infty|^2}{p_\infty} = \rho^2\,\mathbb{E}\big[x^\top J_0^\top J_0 x\big] = \rho^2\,\mathrm{tr}\,\big(J_0^\top J_0\,\Sigma_\infty\big) = \rho^2\|J_0\|_F^2 = 2\lfloor d/2\rfloor\rho^2,$$

using $\Sigma_\infty = I$ and $\|J_0\|_F^2 = 2\lfloor d/2\rfloor$, equal to d for even d and $d-1$ for odd d (one zero block). The identification of σ with the time-reversal relative-entropy rate is the standard stochastic-thermodynamic identity for overdamped diffusions Seifert (2012); a self-contained Girsanov proof for the present process is Lemma 2. For $\rho = 0$ the current velocity vanishes, $v \equiv 0$, so $\bar{f} = f$ and the process is reversible. (iii), including its assertion for arbitrary p_0, is Corollary 1. □

Remark 10 (Consistency of the trajectory estimator of H4)**.** *The estimator $\widehat{\sigma} = \frac{1}{T}\int_0^T \rho J_0 x_t \circ dx_t$ used in H4 is consistent. Writing the Stratonovich integral in Itô form, $(\rho J_0x)\circ dx = (\rho J_0x)\cdot dx + \frac{1}{2}\,d\langle\rho J_0x, x\rangle$, and $d\langle\rho J_0x, x\rangle = 2\rho\,\mathrm{tr}(J_0)\,dt = 0$ because $\mathrm{tr}\,J_0 = 0$. Taking expectations at stationarity, $\mathbb{E}[(\rho J_0x)\cdot(-x+\rho J_0x)] = \rho\,\mathbb{E}[x^\top J_0^\top(-x)] + \rho^2\mathbb{E}\|J_0x\|^2 = 2\lfloor d/2\rfloor\rho^2$, the first term vanishing by antisymmetry ($x^\top J_0 x = 0$), and the martingale part having zero mean. Hence $\mathbb{E}[\widehat{\sigma}] = \sigma$.*

C THE BLIND BIAS: CLOSED FORM, WINDOW LAW, AND THE π-ZERO

C.1 CLOSED FORM FOR GAUSSIAN MIXTURES

Let $p_0 = \sum_m \pi_m \mathcal{N}(\mu_m, \Sigma_m)$ (the empirical measure of a data set is the special case $\Sigma_m = 0$, $\pi_m = 1/M$, used in Section 7 via Remark 6). Under equation 1 the forward marginal is again a Gaussian mixture,

$$u(\cdot, \tau) = \sum_m \pi_m \mathcal{N}\big(\mu_m(\tau),\, \Sigma_m(\tau)\big), \tag{15}$$

where

$$\mu_m(\tau) = e^{-\tau} R(\tau)\mu_m, \quad \Sigma_m(\tau) = e^{-2\tau} R\Sigma_m R^\top + (1 - e^{-2\tau})I,$$

and the rotation-blind kernel of Definition 1 gives the same expression with $R(\tau)$ replaced by I. Both scores are softmax-weighted linear fields,

$$\nabla \log u(x, \tau) = -\sum_m \omega_m(x)\, \Sigma_m(\tau)^{-1}\big(x - \mu_m(\tau)\big), \tag{16}$$

where

$$\omega_m(x) = \mathrm{softmax}_m\Big(\log \pi_m + \log \mathcal{N}\big(x; \mu_m(\tau), \Sigma_m(\tau)\big)\Big),$$

so that $B(\tau; \rho) = \mathbb{E}_{x\sim u}\|\nabla \log u - \nabla \log p^{\mathrm{bl}}\|^2$ is computable to Monte-Carlo accuracy by exact sampling from equation 15 — no training and no time discretization, at any d. This is the computation reported in H2.

C.2 THE WINDOW LAW

Proposition 2 is derived for a single centered Gaussian, where everything is explicit.

Proposition 10 (Window law, derived form)**.** *Let $p_0 = \mathcal{N}(0, \Sigma_0)$ in $d = 2$ with $\Sigma_0 \neq I$, and let $r(\tau) \equiv \omega$. Write Σ_τ and $\Sigma_\tau^{\mathrm{bl}}$ for the rotated and blind forward covariances of equation 15. Then*

$$B(\tau; \rho) = \mathrm{tr}\left[\big(\Sigma_\tau^{-1} - (\Sigma_\tau^{\mathrm{bl}})^{-1}\big)\, \Sigma_\tau \,\big(\Sigma_\tau^{-1} - (\Sigma_\tau^{\mathrm{bl}})^{-1}\big)\right], \tag{17}$$

and, decomposing $\Sigma_0 = I + A_0$ with A_0 symmetric,

$$\Delta_\tau := \Sigma_\tau - \Sigma_\tau^{\mathrm{bl}} = e^{-2\tau}\big(RA_0R^\top - A_0\big), \qquad \|\Delta_\tau\|_F = 2\,\|A_0^{\mathrm{tf}}\|_F\, e^{-2\tau}\,\big|\sin(\omega\tau)\big|, \tag{18}$$

where A_0^{tf} is the trace-free part of A_0. Consequently, in the regime $\|\Delta_\tau\| \ll 1$ where $\Sigma_\tau \approx I$,

$$B(\tau) \;\simeq\; \|\Delta_\tau\|_F^2 \;\propto\; \sin^2(\omega\tau)\, e^{-4\tau}, \tag{19}$$

which is maximized at

$$\frac{d}{d\tau}\Big[\sin^2(\omega\tau)e^{-4\tau}\Big] = 0 \iff 2\omega\sin(\omega\tau)\cos(\omega\tau) = 4\sin^2(\omega\tau) \iff \tan(\omega\tau^*) = \frac{\omega}{2},$$

i.e. $\tau^*(\rho) = \arctan(\rho/2)/\rho$, which is equation 7.

Proof. Equation equation 17 is immediate from equation 16 with a single centered component ($\omega_m \equiv 1$, $\mu_m = 0$), since $\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_\tau)}\|Mx\|^2 = \mathrm{tr}(M\Sigma_\tau M^\top)$. For equation 18, the isotropic parts of Σ_τ and $\Sigma_\tau^{\mathrm{bl}}$ coincide and $R\,I\,R^\top = I$, so only A_0 contributes; the trace-free symmetric 2×2 matrices form a two-dimensional space on which conjugation by $R(\theta)$ acts as a planar rotation by 2θ, whence $\|RA_0^{\mathrm{tf}}R^\top - A_0^{\mathrm{tf}}\|_F = \|A_0^{\mathrm{tf}}\|_F\,|e^{2i\theta} - 1| = 2\|A_0^{\mathrm{tf}}\|_F|\sin\theta|$ with $\theta = \omega\tau$. Finally, writing $(\Sigma^{\mathrm{bl}})^{-1} = (\Sigma_\tau - \Delta)^{-1} = \Sigma_\tau^{-1} + \Sigma_\tau^{-1}\Delta\Sigma_\tau^{-1} + O(\|\Delta\|^2)$ and substituting into equation 17 gives $B = \mathrm{tr}\left[\Sigma_\tau^{-1}\Delta\Sigma_\tau^{-1}\Delta\Sigma_\tau^{-1}\right] + O(\|\Delta\|^3)$, which reduces to $\|\Delta\|_F^2$ when $\Sigma_\tau \approx I$. □

Remark 11 (Accuracy of the estimate)**.** *Evaluating equation 17 numerically without any approximation, for $\Sigma_0 = \mathrm{diag}(\sqrt{5}, 1/\sqrt{5})$ (anisotropy 5:1 at unit determinant), the exact maximizer is $\tau^* = 0.4965/0.4585/0.3803$ for $\rho = 0.3/1/2$, against $\arctan(\rho/2)/\rho = 0.4963/0.4636/0.3927$: the estimate is accurate to 0.04%, 1.1% and 3.3% respectively, the error growing with ρ because $\Sigma_\tau \approx I$ degrades as the window moves earlier. These are the maximizers of the raw bias equation 17; on the score-normalized curve $B(\tau)/\mathbb{E}_{u_\tau}\|\nabla \log u_\tau\|^2$ from which the quoted window positions are read (Appendix G.4), the same geometry peaks at $0.5113/0.4741/0.3958$, and the window law remains accurate to 3.0%, 2.3% and 0.8% there. Proposition 10 is therefore a controlled*

approximation for the single-scale geometry; the systematic 17–33% *offset reported in R2 for multimode data is a separate, geometry-induced shift and not an artifact of this expansion.*

C.3 THE π-ZERO

Proof of Proposition 3. The zero is exact, not perturbative. For even d, at $\theta(\tau) = \pi$ we have $R = \exp(\pi J_0) = -I$. If p_0 is centrally symmetric, then $R_{\#}p_0 = p_0$; rotational equivariance of the isotropic OU semigroup therefore gives $u^J(\cdot, \tau) = u^0(\cdot, \tau)$, so the two scores coincide pointwise and $B = 0$. A Gaussian mixture may satisfy this symmetry through a permutation of its components, for example by pairing means μ and $-\mu$ with equal weights and matching covariances.

If p_0 is not invariant under sign reversal, the reflected and unreflected mixtures need not coincide and a nonzero residual may remain. For odd d, $R(\pi) = \mathrm{diag}(-I_{d-1}, 1)$ under the convention of equation 1, so the same conclusion requires invariance under this partial reflection rather than under $x \mapsto -x$. □

Remark 12. *Numerically, for the centered anisotropic Gaussian above, the closed-form evaluation of $B(\tau = \pi/\omega)$ gives 5×10^{-37} ($\rho = 1$) and 3×10^{-34} ($\rho = 2$), i.e. zero up to floating-point path dependence; the exact value is zero because the two kernels coincide at $R = -I$. The 3.6×10^{-6} residual reported in R4 is the resolution floor of the $\Delta\tau \approx 0.05$ scan — the value at the grid point nearest $\theta = \pi$ — not a property of the target; at the gauge angles the two kernels coincide samplewise, so the Monte-Carlo error there is exactly zero (Appendix G.4). R4 is a closed-form evaluation in the sense of Appendix G.1 and involves no trained network.*

D BLOCK REDUCTION

Proof of Proposition 4. Let $\mathbb{R}^d = \bigoplus_{k=1}^m V_k$ be the decomposition into the invariant planes of $J = \mathrm{blockdiag}(\omega_1 J_0, \ldots, \omega_m J_0)$, and suppose $p_0 = \bigotimes_k p_0^k$ with p_0^k supported on V_k. The drift $-\gamma I + J$ and the diffusion $\sqrt{2}\, I$ are both block diagonal with respect to this decomposition, so the forward process decouples: the components x_τ^k are independent and $u(\cdot, \tau) = \bigotimes_k u_k(\cdot, \tau)$. The same holds for the blind kernel, which is block diagonal a fortiori. Therefore

$$\nabla \log u = \bigoplus_k \nabla \log u_k, \qquad \nabla \log p^{\mathrm{bl}} = \bigoplus_k \nabla \log p_k^{\mathrm{bl}},$$

and the discrepancy field is the direct sum $\delta = \bigoplus_k \delta_k$ with $\delta_k = \nabla \log u_k - \nabla \log p_k^{\mathrm{bl}}$ depending only on x^k. Since the blocks are independent under u,

$$B(\tau) = \mathbb{E}_{x\sim u}\|\delta\|^2 = \sum_k \mathbb{E}_{x^k \sim u_k}\|\delta_k\|^2 = \sum_k B_k(\tau),$$

which is the claimed exact decomposition. Writing $G_k(\tau) := \mathbb{E}\|\nabla \log u_k\|^2$, the normalized bias is

$$\frac{B(\tau)}{\sum_k G_k(\tau)} = \sum_k \frac{G_k(\tau)}{\sum_j G_j(\tau)} \cdot \frac{B_k(\tau)}{G_k(\tau)}, \tag{20}$$

a G_k-weighted average of the per-block normalized biases, as claimed.

Clause (i) follows at once: if all blocks carry the same p_0^k and the same ω_k, every summand in equation 20 is the planar normalized bias, so the weighted average equals it identically — in particular its maximizer and maximum are the planar ones, for every m and hence for every $d = 2m$.

Clause (ii) follows from equation 20 together with unimodality of the planar curves. Blocks at a common speed but different structure contribute curves with a common maximizer (the maximizer depends on ω through equation 7 and on the structure only through the second-order shift of Remark 11), so the convex combination attains its maximum at the same place while its height is the weighted average of the heights. Blocks at different speeds contribute curves with displaced maximizers; for the present block family the exact formula equation 20 allows the mixture maximum to be computed directly, and in all tested cases it lies between the component maxima and is dominated by the deepest block (we do not claim a general unimodality-preservation statement for arbitrary unimodal summands). The quantitative statements — including the absence of an intermediate peak — are reported in R5. □

E THE DERIVATIVE-FREE FIRST-MOMENT ESTIMATOR

This appendix derives equation 5 and quantifies its error. Throughout, $\psi = \log w$ is the log-density in the caloric frame, y is the evaluation point, $r > 0$ the finite-difference radius, and $z_1, \dots, z_h$ are i.i.d. draws from a law ν on $\mathbb{R}^d$ of *radial–angular product form*,

$$z = \varsigma\, U, \qquad U \sim \mathrm{Unif}(S^{d-1}) \text{ independent of the magnitude } \varsigma \geq 0, \qquad \mathbb{E}[\varsigma^4] < \infty. \tag{21}$$

This is stronger than isotropy of the covariance alone and is what makes the fourth-moment computation of part (iv) below exact; it is also the structure the κ-sampler actually has. It implies $\mathbb{E}_\nu[z] = 0$ and $\mathbb{E}_\nu[zz^\top] = m^2 I$ with $m^2 = \mathbb{E}[\varsigma^2]/d$. In the heat-ball application the z_i are produced by the κ-sampler of Bai et al. (2026, Prop. 11, Alg. 1), which factorizes into independent radial, temporal and spherical components with the angular factor uniform on S^{d-1}: exactly the form equation 21. Hence

$$m^2 = \frac{1}{d}\,\mathbb{E}_\kappa\|z\|^2 = \frac{1}{d}\,\mathbb{E}[\varsigma^2], \tag{22}$$

the *per-coordinate* second moment. This convention fixes the constant in equation 5 and must be matched exactly by any implementation: at $d = 2$ it evaluates to $1/(16\pi) = 0.0198944$, and the sampler reproduces it to 0.019886, the residual being grid discretization.

E.1 DERIVATION

The score representation equation 4 requires $\nabla\psi$ at each sampled ball point. The following identity replaces it by two function evaluations per direction.

Proposition 11 (Derivative-free first moment)**.** *Let ν satisfy equation 21 with bounded support, and let $\psi \in C^4$ on a neighbourhood of $y \pm r\,\mathrm{supp}(\nu)$; set*

$$R_i^{\pm} = \exp\big(\psi(y \pm r z_i) - \psi(y)\big), \qquad S_h = \frac{1}{m^2 r} \cdot \frac{\sum_{i=1}^h z_i\,(R_i^+ - R_i^-)}{\sum_{i=1}^h (R_i^+ + R_i^-)}.$$

Then

(i) (odd moment) $\mathbb{E}\big[z\,(R^+ - R^-)\big] = 2\,r\,m^2\,\nabla\psi(y) + O(r^3)$;

(ii) (even moment) $\mathbb{E}\big[R^+ + R^-\big] = 2 + O(r^2)$; *moreover $R^\pm > 0$, so for r small enough the denominator satisfies $\mathbb{E}[R^+ + R^-] \geq c > 0$ and $\mathbb{E}\big[(R^+ + R^-)^2\big] < \infty$;*

(iii) (consistency and error budget) *for fixed r, $S_h \to \nabla\psi(y) + O(r^2)$ almost surely as $h \to \infty$, and*

$$S_h = \nabla\psi(y) + \underbrace{O(r^2)}_{\textit{discretization}} + \underbrace{O(h^{-1})}_{\textit{ratio}} + \underbrace{O_P\big(h^{-1/2}\big)}_{\textit{Monte Carlo}};$$

(iv) (variance) *for $g = \nabla\psi(y) \neq 0$, writing $\beta := \mathbb{E}[\varsigma^4]/\big(\mathbb{E}[\varsigma^2]\big)^2$ (≥ 1 by Cauchy–Schwarz), the relative standard deviation of S_h is*

$$\mathrm{RSD}(S_h) = \frac{1}{\sqrt{h}}\,\big(\beta d - 1\big)^{1/2}\,\big(1 + O(r^2)\big). \tag{23}$$

Corollary 2 (The constant for the κ-sampler)**.** *For the κ-measure the magnitude factorizes as $\varsigma = \varrho\,R(\sigma)$, where $\varrho \in [0,1]$ is the radius normalized by the ball radius R at the sampled time offset σ, and ϱ, σ and U are mutually independent (Bai et al., 2026, Prop. 11). Since β is scale-invariant it factorizes accordingly, $\beta = \beta_\varrho\,\beta_R$, and both factors are explicit.*

Radial factor. *Bai et al. (2026, Prop. 11) place the* $\mathrm{Beta}(d/2+1, 1)$ *law on $t = \varrho^2$, not on ϱ; hence $\varrho \sim \mathrm{Beta}(d+2, 1)$, $\mathbb{E}[\varrho^k] = (d+2)/(d+2+k)$, and*

$$\beta_\varrho = \frac{(d+4)^2}{(d+2)(d+6)} = 1 + \frac{4}{(d+2)(d+6)} = 1 + \frac{4}{d^2} + O(d^{-3}). \tag{24}$$

Temporal factor. *Setting $r = 1$ in equation 2, the ball radius at time offset σ solves $(4\pi\sigma)^{-d/2} e^{-R^2/4\sigma} = 1$, i.e. $R^2 = -2d\sigma\ln(4\pi\sigma)$. Writing $L := -\ln(4\pi\sigma)$ this is*

$$R(\sigma)^2 = \frac{d}{2\pi}\,L\,e^{-L}, \qquad L \sim \mathrm{Gamma}\Big(\tfrac{d+4}{2},\ \text{rate } \tfrac{d}{2}\Big). \tag{25}$$

The Gamma moment generating function gives $\mathbb{E}[L^k e^{-tL}] = \lambda^\alpha \Gamma(\alpha+k)/\big(\Gamma(\alpha)(\lambda+t)^{\alpha+k}\big)$ *with* $\alpha = \frac{d+4}{2}$, $\lambda = \frac{d}{2}$, *whence* $\beta_R = \mathbb{E}[L^2 e^{-2L}]/\big(\mathbb{E}[Le^{-L}]\big)^2$ *evaluates in closed form to*

$$\beta_R = \frac{d+6}{d+4} \cdot \frac{(d+2)^{\,d+6}}{d^{\,(d+4)/2}\,(d+4)^{\,(d+8)/2}} = 1 + \frac{2}{d^2} + O(d^{-3}). \tag{26}$$

Consequence. *Therefore*

$$\beta = \beta_\varrho\, \beta_R = 1 + \tfrac{6}{d^2} + O(d^{-3}), \quad \mathrm{RSD}(S_h) = \sqrt{\tfrac{\beta d - 1}{h}} = \sqrt{\tfrac{d-1+O(d^{-1})}{h}} \asymp \sqrt{\tfrac{d}{h}}, \tag{27}$$

which is the $O(\sqrt{d/h})$ *Monte-Carlo factor quoted in the main text. For comparison, a deterministic magnitude gives* $\beta = 1$ *and a Gaussian* z *gives* $\beta = 1 + 2/d$, *i.e.* $\mathrm{RSD} = \sqrt{(d+1)/h}$: *the* κ*-measure is closer to the deterministic case than to the Gaussian one.*

Remark 13 (Why the correction is $O(d^{-2})$ and not $O(d^{-1/2})$)**.** *The temporal coordinate concentrates only at rate* $d^{-1/2}$ *(Bai et al., 2026, Cor. 13) — indeed* $\mathbb{E}[L] = 1 + 4/d$ *and* $\mathrm{Var}(L) = 2(d+4)/d^2$, *so* $L = 1 + O_P(d^{-1/2})$ *— yet* $\beta_R - 1$ *is* $O(d^{-2})$. *The reason is that* β_R *measures the fluctuation of* $R^2 \propto Le^{-L}$, *and* $L \mapsto Le^{-L}$ *is* stationary *exactly at* $L = 1$, *which is the point* L *concentrates on. The first-order fluctuation therefore cancels: with* $f(L) = Le^{-L}$, $f'(1) = 0$ *and* $f''(1) = -e^{-1}$, *so* $R^2 \propto 1 - \frac{1}{2}(L-1)^2 + \dots$ *and*

$$\beta_R - 1 = \frac{\mathrm{Var}(R^2)}{\big(\mathbb{E}R^2\big)^2} \simeq \mathrm{Var}\Big(\tfrac{1}{2}(L-1)^2\Big) \simeq \frac{2}{d^2},$$

matching equation 26. Concentration of the sampler is thus better than the first-order rate suggests for this particular functional.

Remark 14 (Numerical confirmation)**.** *Evaluating equation 24–equation 27 at* 50*-digit precision and comparing with a direct measurement of* $\mathbb{E}\|z\|^4/\big(\mathbb{E}\|z\|^2\big)^2$ *from* 4×10^6 κ*-samples:*

d	2	32	256	1024
β_ϱ *equation 24*	1.125000	1.003096	1.000059	1.000004
β_R *equation 26*	1.404664	1.002144	1.000031	1.000002
β *(closed form)*	1.580247	1.005247	1.000090	1.000006
β *(sampled)*	1.5802	1.00525	1.000090	—

Agreement is within the 5×10^{-4} *Monte-Carlo error. The scaled deviation* $(\beta - 1)d^2$ *equals* 5.373, 5.829, 5.913, 5.978 *at* $d = 32, 128, 256, 1024$, *converging to the predicted* 6. *At* $d = 256$ *the total departure of* β *from unity is* 9×10^{-5}, *so the constant in equation 27 is* $\sqrt{d-1}$ *to five digits.*

Proof. Fix a direction z and put $\phi(t) = w(y+tz)/w(y) = \exp(\psi(y+tz) - \psi(y))$, so that $R^\pm = \phi(\pm r)$ and $\phi(0) = 1$, $\phi'(0) = z \cdot \nabla\psi(y)$.

(i) Since $\psi \in C^4$, so is ϕ, and Taylor expansion with Lagrange remainder cancels the even orders in the symmetric difference:

$$R^+ - R^- = \phi(r) - \phi(-r) = 2r\,\phi'(0) + \tfrac{r^3}{3}\phi'''(0) + \tfrac{r^4}{24}\big[\phi''''(\xi_+) - \phi''''(\xi_-)\big] = 2r\,\big(z \cdot \nabla\psi(y)\big) + O(r^3),$$

the bracket being $o(1)$ by continuity of ϕ''''. We deliberately do not push the expansion to $O(r^5)$, which would require $\psi \in C^5$; $O(r^3)$ is all that the statement needs. Multiplying by z and taking expectations, the first term gives $2r\,\mathbb{E}[zz^\top]\nabla\psi = 2rm^2\nabla\psi$ by equation 21; the remainder is $O(r^3)$ because $|\phi'''(0)| \lesssim \|z\|^3$ and $\mathbb{E}[\varsigma^4] < \infty$. Note that the leading error is $O(r^3)$, not $O(r^2)$: the symmetric difference is what removes the second-order term, and this is why equation 5 uses $R^+ - R^-$ rather than a one-sided difference.

(ii) Similarly $R^+ + R^- = 2\phi(0) + r^2\phi''(0) + O(r^4) = 2 + O(r^2)$, and $\phi''(0) = z^\top\big(\nabla^2 w/w\big)z$ has finite expectation.

(iii) By the strong law applied separately to numerator and denominator,

$$\frac{h^{-1}\sum_i z_i(R_i^+ - R_i^-)}{h^{-1}\sum_i (R_i^+ + R_i^-)} \xrightarrow[h\to\infty]{\text{a.s.}} \frac{2rm^2\nabla\psi + O(r^3)}{2 + O(r^2)} = r\,m^2\,\nabla\psi\,\big(1 + O(r^2)\big),$$

the denominator being bounded away from zero for small r by part (ii). Dividing by $m^2 r$ gives $\nabla\psi(y)\,(1 + O(r^2))$. The estimator is therefore *consistent in the joint limit* $h \to \infty$, $r \to 0$, and *second-order accurate in the finite-difference radius*, with a discretization bias of order r^2. We stress that it is *not* an unbiased estimator at finite h: being a ratio of two sample averages it carries the usual ratio bias of order h^{-1}, which requires exactly the two conditions of part (ii) — a denominator with finite second moment and a population mean bounded away from zero — and which is dominated by the $h^{-1/2}$ Monte-Carlo fluctuation. The $O_P(\cdot)$ term is the central limit theorem applied to the numerator, with the constant given by part (iv).

(iv) To leading order in r the numerator summand is $X := z\,(R^+ - R^-) \simeq 2r\,z\,(z \cdot g)$ with $g = \nabla\psi(y)$. Using the product form equation 21, $\|z\|^2 = \varsigma^2$ and $z \cdot g = \varsigma\,(U \cdot g)$ with $\mathbb{E}[(U \cdot g)^2] = \|g\|^2/d$, so

$$\mathbb{E}\|X\|^2 = 4r^2\,\mathbb{E}[\varsigma^4]\,\frac{\|g\|^2}{d}, \qquad \|\mathbb{E}X\|^2 = 4r^2 m^4 \|g\|^2 = 4r^2\,\frac{\big(\mathbb{E}[\varsigma^2]\big)^2}{d^2}\|g\|^2,$$

and therefore

$$\operatorname{tr}\operatorname{Cov}(X) = \mathbb{E}\|X\|^2 - \|\mathbb{E}X\|^2 = 4r^2 m^4 \|g\|^2\big(\beta d - 1\big), \qquad \beta = \frac{\mathbb{E}[\varsigma^4]}{\big(\mathbb{E}[\varsigma^2]\big)^2}.$$

Averaging h i.i.d. copies and dividing by the denominator (which concentrates at 2) gives equation 23. It is essential here to subtract $\|\mathbb{E}X\|^2$: the raw second moment $\mathbb{E}\|X\|^2$ alone would give $\sqrt{\beta d/h}$ and overstate the fluctuation — for Gaussian z, for instance, the correct factor is $\sqrt{d+1}$ and not $\sqrt{d+2}$. □

Remark 15 (Numerical confirmation of the constant)**.** *We verify equation 5 against the exact score of a caloric target, using the κ-sampler itself rather than a generic isotropic law, so that the convention equation 22 is exercised as implemented. In $d = 6$, with 4×10^6 sampled directions, anchor $y = (0.3, -0.5, 0.8, 0.1, -0.2, 0.6)$ and $s = 0.7$, the target is $\nabla\psi = -y/(1 + 2s)$. The relative error of S_h is*

$$1.85 \times 10^{-3}\ (r = 0.20), \qquad 1.74 \times 10^{-3}\ (r = 0.10), \qquad 1.72 \times 10^{-3}\ (r = 0.05).$$

The error is flat in r and equals the Monte-Carlo floor $\approx 2 \times 10^{-3}$ implied by $h = 4 \times 10^6$ and Proposition 11(iv). This confirms the constant $1/(m^2 r)$ — an error in it would appear as a fixed multiplicative offset, not as a floor — and shows that the discretization bias is already below the sampling floor at $r = 0.20$ for a target of this smoothness. It does not resolve the $O(r^2)$ rate of Proposition 11(iii), which lies below the sampling floor at this h; a direct rate test would require a target with larger third derivatives or substantially larger h.

E.2 When to prefer equation 5

Reverse-mode automatic differentiation obtains the full d-dimensional gradient at a cost of two to three forward passes, so equation 5, which needs $2h$ forward passes, is never cheaper when ψ is a differentiable network and backpropagation is available. Its purpose is different: it is the only route when $\nabla\psi$ is unavailable in principle — when ψ is a black-box density oracle, a non-differentiable simulator output, or an externally supplied energy — and it removes autodifferentiation from the generation-time loop, where the graph would otherwise have to be retained through every sampled ball point. Proposition 11(iii)–(iv) quantifies exactly what is paid for that: a bias of order r^2, which is the same order as the heat-ball radius already present in equation 4, and a variance inflation of $\sqrt{(\beta d - 1)/h} \asymp \sqrt{d/h}$ (Proposition 11(iv) and Corollary 2).

F The speciation time for the rotated family

R6 compares τ^* with the speciation time t_S of Biroli et al. (2024), defined by the eigenvalue-crossing criterion $\lambda_{\min}\big(\widetilde{M}(t_S)\big) = 0$ with $\widetilde{M}(t) = \Sigma_{\mathrm{sto}}(t) - e^{-At}\Sigma_{\mathrm{B}}\big(e^{-At}\big)^\top$ in the form given by Lu & Tang (2026). For the isotropic-dissipation family the criterion collapses to a closed form, and the collapse itself shows that t_S cannot depend on ρ.

Lemma 3 (Speciation time, closed form)**.** *For equation 1 with $\gamma = 1$ and $J = \rho J_0$ one has $A = I - \rho J_0$, $e^{-At} = e^{-t} R(\rho t)$ with R orthogonal, and $\Sigma_{\mathrm{sto}}(t) = (1 - e^{-2t})I$. Hence*

$$\widetilde{M}(t) = (1 - e^{-2t})\,I - e^{-2t}\,R\,\Sigma_{\mathrm{B}} R^\top, \qquad \lambda_{\min}\big(\widetilde{M}(t)\big) = (1 - e^{-2t}) - e^{-2t}\lambda_{\max}(\Sigma_{\mathrm{B}}),$$

Table 2: Visibility-window position τ^* (ρ=2) against the speciation time $t_S = \frac{1}{2}\ln(1+\lambda_{\max}(\Sigma_B))$, with Σ_B the traceless part of the data covariance. The ratio τ^*/t_S follows the one-parameter curve $r(\lambda) = \bar{\tau}/\frac{1}{2}\ln(1+\lambda)$ of Figure 2; the identity $\tau^* = t_S$ is a crossing at $\lambda_{\max} \approx 1.3$, not a band.

dataset	$\lambda_{\max}(\Sigma_B)$	τ^* (ρ=2)	t_S	ratio
aniso_gauss	1.08	0.36	0.366	0.97
asym_gmm	2.41	0.36	0.613	0.58
sparse_dense_gmm	1.61	0.46	0.480	0.96
scaled_gmm	0.40	0.25	0.169	1.50

Table 3: Visibility window at $d = 256$ (MNIST16). t_S is ρ-independent by construction; computed from the between-class scatter of the ten digit classes (traceless-covariance variant: $t_S = 1.46$, same conclusion).

ρ	τ^*	peak depth	τ^*_{2D} (ref.)	t_S
0.3	0.87	1.1%	0.66	1.23
1	0.66	10.7%	0.56	1.23
2	0.46	32.9%	0.46	1.23

and the crossing condition gives

$$t_S = \tfrac{1}{2}\ln\big(1 + \lambda_{\max}(\Sigma_{\mathrm{B}})\big), \tag{28}$$

which is independent of ρ.

Proof. Because γI commutes with J, the fundamental matrix factorizes as $e^{-At} = e^{-t}R(\rho t)$ with $R(\theta) = \exp(\theta J_0) \in SO(d)$ (Proposition 6). Orthogonal conjugation preserves the spectrum, so $R\Sigma_{\mathrm{B}}R^\top$ has the same eigenvalues as Σ_{B}; the rotation therefore drops out of every spectral functional of $\widetilde{M}$, which is the structural reason for the ρ-independence. Since $\widetilde{M}$ is $(1-e^{-2t})I$ minus a positive semidefinite matrix, its smallest eigenvalue is $(1 - e^{-2t}) - e^{-2t}\lambda_{\max}(\Sigma_{\mathrm{B}})$, and setting it to zero gives $e^{2t_S} = 1 + \lambda_{\max}(\Sigma_{\mathrm{B}})$. □

Remark 16. *Solving $\lambda_{\min}(\widetilde{M}(t)) = 0$ numerically reproduces equation 28 to all printed digits, giving $t_S = 0.366184,\ 0.613356,\ 1.229794$ for aniso_gauss, asym_gmm and MNIST16 respectively, and the value is unchanged between $\rho = 0.5$ and $\rho = 2$, as Lemma 3 requires. Note that Table 2 lists $\lambda_{\max}(\Sigma_{\mathrm{B}})$ rounded to $1.08,\ 2.41,\ 10.7$; recomputing equation 28 from the rounded values gives $0.365944,\ 0.612916,\ 1.229495$, which differ from the quoted t_S in the fourth decimal only. Table entries are rounded ($0.366,\ 1.23$) while the full-precision values are used throughout the computation. These are the t_S entries of Tables 2 and 3.*

Remark 17 (The ratio curve)**.** *Combining Lemma 3 with the near-constancy of τ^* established in R2 and R5 — the window position is set by the rotation–survival trade-off of Proposition 2 and is insensitive to the anisotropy ratio — the ratio plotted in Figure 2 is*

$$r(\lambda) = \frac{\tau^*}{t_S} \;\simeq\; \frac{\bar{\tau}}{\frac{1}{2}\ln(1+\lambda)}, \qquad \bar{\tau} \approx 0.41,$$

a one-parameter curve with no fitted quantity beyond the single scale $\bar{\tau}$. The identity $\tau^* = t_S$ is the solution of $\frac{1}{2}\ln(1+\lambda) = \bar{\tau}$, i.e. the isolated crossing $\lambda_{\max}(\Sigma_{\mathrm{B}}) = e^{2\bar{\tau}} - 1 \approx 1.3$. We stress that the content of the curve is the near-constancy of τ^* together with equation 28; it is a statement about how two independently defined time scales are related, not an additional fitted law.

G EXPERIMENTAL DETAILS

G.1 WHAT REQUIRES TRAINING AND WHAT DOES NOT

It is useful to separate the two kinds of experiment in this paper, because they have very different dependence on implementation choices.

Closed-form evaluations. R2–R6 and H2 evaluate the blind bias $B(\tau;\rho)$ of Definition 1 directly from the expressions equation 15–equation 16 of Appendix C.1. Both the rotated and the blind forward marginals are Gaussian mixtures with softmax-weighted linear scores, so the bias is obtained by exact sampling from the rotated marginal: *no network is trained, no time discretization is introduced, and no hyperparameter other than the Monte-Carlo sample size enters.* This covers the window law, the mechanism taxonomy, the π-zero, the block reduction and the entire $d = 256$ window analysis. The π-zero of Proposition 3 under its symmetry condition and the block decomposition of Proposition 4 are, in addition, exact algebraic identities that the numerics only confirm.

Trained-model runs. Only R1 and H1 involve a trained ψ_θ. Their content is that the trained metrics coincide across ρ, which is a consistency check on Corollary 1 at finite precision rather than a performance claim; in particular the conclusion is insensitive to the architecture, since the same network is used for every rotation setting and the comparison is within-architecture throughout. Two caveats apply. In the reduced pipeline ω does not enter the training loop, so the closed-form rows of Table 1 agree by construction and test only the implementation; the non-circular evidence is the direct-simulation row, which at d=256 is mildly worse on all three metrics — a pattern we attribute to the fixed-table design quantified in 2D (Appendix G.3) rather than to ρ-dependence, but which a seed-resolved d=256 control would settle more cleanly. Neither run is a trained-pipeline positive control; the closed-form version of that control — a deliberately non-conformal drift the instrument must detect as a violation — is verified in Appendix H, and the trained-pipeline version is left to future work (Section 8).

G.2 TRAINED-MODEL RUNS (R1, H1)

The log-density network $\psi_\theta(z, s)$ is trained with the constraint losses of Bai et al. (2026)—an anchor term, the caloric heat-ball residual ℓ_{hb}, and the score heat-ball residual ℓ_{shb}—for 6000–10000 steps in 2D and 4000 steps at $d = 256$, using a fixed seed (CPU in 2D and the MPS backend at $d = 256$); full hyperparameters are reported in Table 4. The observed MPS variation is reported in Table 1. The network uses explicit `kaiming_normal_` initialization: the PyTorch default `kaiming_uniform_`$(a = \sqrt{5})$ corresponds to a gain of 0.577 rather than the $\sqrt{2}$ appropriate for the activation used, which attenuates input gradients by roughly two orders of magnitude at depth 4 and suppresses the correction term that the heat-ball losses are supposed to shape. The κ-sampler of Bai et al. (2026, Alg. 1) is used as an off-the-shelf baseline, unmodified, with the same settings for every rotation setting: this is what makes the comparison across ρ a controlled one, and it is the reason the sampler never appears as a confound in R1 or H1.

Metrics. MMD is the squared maximum mean discrepancy with an RBF kernel whose bandwidth is the median squared pairwise distance on the evaluation grid (median heuristic), with the bias terms excluded; sliced-W averages the one-dimensional Wasserstein distance over 50 seeded random projections; gen std is the per-pixel standard deviation of generated samples (data std $= 1$ after standardization); and the nn ratio is the median distance from a generated sample to the training set divided by the median internal training distance, with the test-to-training value as an absolute reference.

Rotation enters only through the profile $r(\tau)$: const ($r \equiv \omega$), osc ($r = \omega \sin k\tau$) and ramp ($r = \omega \tanh k\tau$) with $k = 3$. All comparisons across ρ use a matched step budget. In the reduced pipeline ω does not enter the training loop at all, so the ω=0 value at 6000 steps is read off the (bit-identical) training trajectory rather than rerun.

Seeds and the two reported spreads. Table 1 reports two spreads that must not be conflated. *At fixed seed*, sweeping $\rho \in \{0.3, 1, 2\}$ changes the score MSE at τ=0.3 only within 0.0986–0.1048; training is bit-identical in the reduced pipeline, so this spread is evaluation-set rotation noise, not training noise. *At fixed* $\rho = 2$, sweeping four data-realization seeds gives 0.105–0.145 for the

Table 4: Hyperparameters. Left: the 2D pipeline (sparse_dense_gmm and the mechanism datasets). Right: MNIST16 (d=256). Wall-clock: ≈20 min per 6000-step 2D run (CPU, Apple M-series) and ≈30 s sampling; ≈13 min per 4000-step MNIST16 run (MPS). $\tau_{\max} = 5.025$ and $S_{\max} \approx 1.16 \times 10^4$ throughout.

	2D	MNIST16 (d=256)
network	MLP, width 256, depth 4, SiLU	conv (4× down.), widths 32–128, FiLM
time embedding	$u = \tau/\tau_{\max} + 8$ sinusoids	FiLM on $u + 8$ sinusoids
optimizer	Adam, 10^{-3}, cosine to 0.1×	Adam, 10^{-3}, cosine, wd 10^{-5}
batch / steps	512 / 6000–10^4	128 / 4000
gradient clip	10	5
heat-ball radius / κ's	$r^2 = 0.125\, s$ / 64	$r^2 = 0.02\, s$ / 16
loss weights	$w_{\rm hb}$=2, $w_{\rm shb}$=0.3, $w_{\rm mse}$=0.05, $w_{\rm dir}$=2	$w_{\rm hb}$=10, $w_{\rm bnd}$=0.01; exact target for $s \geq 0.3$ (4096 refs)
sampling	500 steps, 64 κ's, $s_{\min}$=10^{-5}	500 steps, 512 κ's, $s_{\min}$=10^{-5}
data / eval sizes	10^4 samples; eval 2×10^3	10^4 train / 2×10^3 test
seeds	0 (headline), 0–3 (bands)	0
direct-sim table	4096 trajectories × 80 grid ($dt = \tau_{\max}/79$)	—

closed-form pipeline and 0.124–0.306 for the direct-simulation control. The elevation of the latter is analysed in R1 and is a property of the control's design, not of the gauge.

G.3 The direct-simulation control

To test Proposition 6 independently of the closed-form reduction, training data are generated by explicit Euler–Maruyama integration of equation 1 in the x coordinate on an 80-point time grid (step size $\tau_{\max}/79$), then pulled back to z. The base configuration uses 4096 trajectories, held fixed and reused throughout training; the enlarged-table check of R1 uses 10^4 trajectories and brings the same seed's score MSE from 0.306 to 0.136, inside the closed-form band. This is the memorization artifact discussed in R1: a fixed simulated table adds a memorization term to held-out MSE that fresh closed-form sampling does not incur, which is why the distribution-level metrics — immune to it — overlap across pipelines while the pointwise MSE does not.

Note that the direct-simulation control and the entropy-production estimator of Appendix G.5 use different integrators (explicit Euler–Maruyama on a fixed grid versus a Heun scheme at fixed dt) for different purposes, and their parameters are not interchangeable.

G.4 Closed-form evaluations (R2–R6, H2)

The blind bias is evaluated on a uniform grid in τ by exact sampling from the rotated marginal equation 15; window positions τ^* are located as the grid argmax of the score-normalized bias $\mathrm{rel}(\tau) = B(\tau)/\mathbb{E}_{u_\tau}\|\nabla \log u_\tau\|^2$ — the quantity plotted in the bias-curve figures — and are quoted at the resolution of that grid. All planar scans evaluate the outer expectation of Appendix C.1 with $n = 4000$ independent samples; the d=256 scans use $n_{\rm mc} = 30000$ against mixtures of at most 8192 components (Remark 6). The Monte-Carlo standard error of the reported normalized bias is then $\lesssim 6 \times 10^{-2}$ (about 3% relative) on the multimode curves. At the exact gauge angles — $\theta = 0$, and $\theta = \pi$ for centrally symmetric single-mode data — the two kernels coincide and the bias vanishes identically for every sample, so its Monte-Carlo error is zero there. The small residuals quoted in R3–R4 are therefore floors of the *evaluations* rather than properties of the target: closed-form double-precision evaluations at the exact angle return path-dependent floating-point residuals ($\lesssim 10^{-34}$; cf. Appendix C.3), while the 3.6×10^{-6} of R4 is the value at the grid point nearest $\theta = \pi$ ($\Delta\theta \approx 2 \times 10^{-2}$ on the $\Delta\tau \approx 5 \times 10^{-2}$ grid), i.e. the resolution floor of the scan.

All scans, planar and $d = 256$, use $\tau \in [0.05, 2.5]$ with one of two *nested* uniform grids: $\Delta\tau = 2.45/48 \approx 0.051$ (49 points) for the depth-exponent scan, the block-heterogeneity experiment of R5 and the determinant-one family; and $\Delta\tau = 2.45/24 \approx 0.102$ (25 points, every second point of the former) for the headline window positions, the mechanism taxonomy of R4,

the speciation comparison of R6 and the $d = 256$ empirical-measure scan. Although the coarse grid is a subgrid of the fine one, the reported grid argmax can depend on the grid resolution; each experiment therefore reports τ^* at its specified grid resolution. Writing $\tau_k = 0.05 + k\,\Delta\tau$, every position quoted in the paper is a grid point: with the coarse spacing, $k = 2, 3, 4, 5, 6, 8$ give 0.254, 0.356, 0.458, 0.560, 0.663, 0.867, reported as 0.25, 0.36, 0.46, 0.56, 0.66, 0.87; the value $\tau^* = 0.41$ of R5 and of the determinant-one family is $k = 7$ of the fine grid (0.407), which the coarse grid does not contain. This is also why the MNIST16 window is quoted as $\tau^* = 0.458$ in R6 and appears rounded to 0.46 in Table 3: they are the same grid point. For the $d = 256$ computation the empirical measure of MNIST16 ($16 \times 16 = 256$ dimensions) is used as p_0 in the sense of Remark 6, so the M-component mixture is the data set itself and the evaluation is exact up to Monte-Carlo error in the outer expectation.

The depth exponents quoted in R2 ($\rho^{1.61}$ multimode, $\rho^{1.84}$ covariance-gated) and in H2 ($\rho^{1.83}$ at $d = 256$) are log-log least-squares fits over a common ten-point scan, $\rho \in \{0.1, 0.2, 0.3, 0.6, 1, 1.5, 2, 3, 4, 6\}$, so that the three values of ρ reported in the tables are interior points of the scan rather than its support. The three values of Table 3 alone give 1.80, consistent with the ten-point fit.

The speciation times t_S of Tables 2 and 3 are evaluated from the closed form equation 28 of Lemma 3; the entries 0.366, 0.613 and 1.23 are the rounded values of 0.366184, 0.613356 and 1.229794 obtained there.

G.5 Entropy-production estimation (H4)

$M = 5000$ trajectories of equation 1 are simulated at stationarity with a Heun scheme at $dt = 2 \times 10^{-3}$, and σ is estimated by the Stratonovich mid-point discretization of $\widehat{\sigma} = \frac{1}{T}\int_0^T \rho J_0 x_t \circ dx_t$, whose consistency is established in Remark 10. The quoted uncertainties are standard errors over the M independent trajectories.

G.6 Reproducibility

Every result other than R1 and H1 is reproducible from the closed forms of Appendices C.1 and F together with the data-set definitions of Section 5, without access to any trained model. Code for the closed-form evaluations and for the trained-model pipeline will be released with the paper.

H Closed-form positive control

All quantities in this section are closed-form. For the drift $F = -A + J$ the forward kernel is Gaussian, $x_\tau \mid x_0 \sim \mathcal{N}(e^{F\tau}x_0, \Sigma(\tau))$ with $\dot{\Sigma} = F\Sigma + \Sigma F^\top + 2I$, $\Sigma(0) = 0$; the blind kernel replaces F by $-A$. Both scores are linear fields, so the bias of Definition 1 is the exact trace quadratic form $B = \mathrm{tr}(\Delta P\,\Sigma_u\,\Delta P^\top)$, $\Delta P = \Sigma_u^{-1} - \Sigma_{\mathrm{bl}}^{-1}$, evaluated against radial data $p_0 = \mathcal{N}(0, I)$ for which Lemma 1 guarantees $B \equiv 0$ throughout the conformal class. Figure 4 shows that this invisibility is sharp: for the non-conformal drift the radial density acquires non-radial structure and the bias becomes nonzero, growing with ω; the same contrast appears in the marginal divergence $\mathrm{KL}(u_\omega \| u_0)$. This closes the logic of the trained-pipeline checks (R1, H1) at the instrument level: the estimator family responds to rotational drift exactly where, and only where, the gauge argument no longer forbids an effect; a trained-pipeline version of the control is left to future work.

I ADDITIONAL EXPERIMENTAL FIGURES

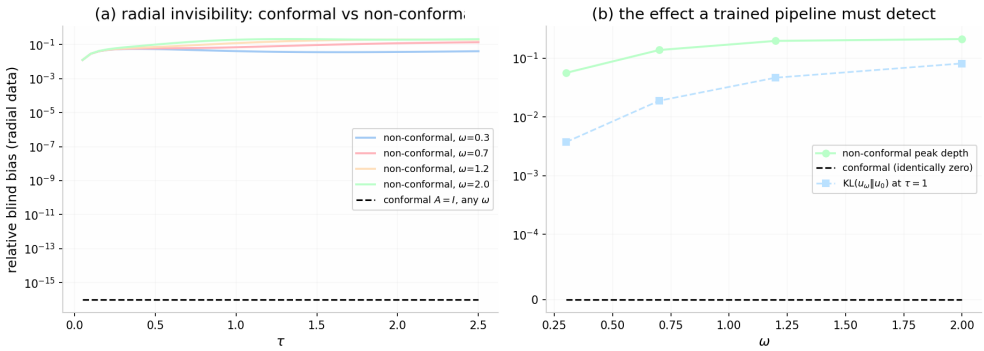

Figure 4: Closed-form positive control. With radial data ($p_0 = \mathcal{N}(0, I)$), the blind bias of Definition 1 is identically zero for every ω in the conformal family $A = \gamma I$ (dashed; at the float floor 2×10^{-29}), whereas a non-conformal drift $A = \mathrm{diag}(1, 0.25)$, $J = \omega J_0$ renders the radial density non-radial and the bias nonzero, with peak depth growing from 5.6% (ω=0.3) to 21% (ω=2). (b) Peak depth and $\mathrm{KL}(u_\omega \| u_0)$ at τ=1 against ω: the conformal line is exactly flat while the non-conformal effect grows monotonically — the signal a trained-pipeline positive control must detect.

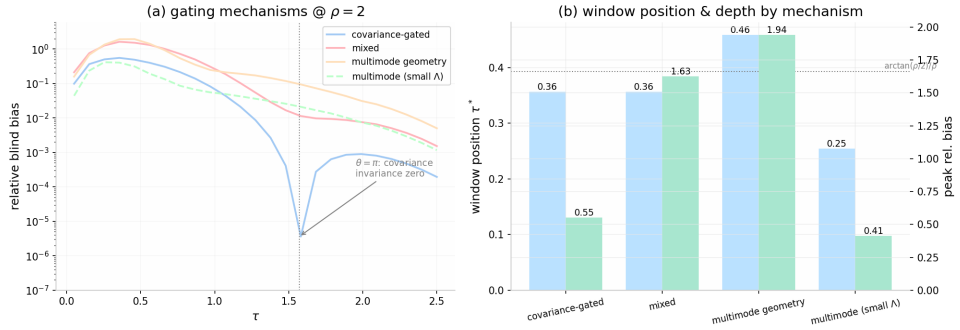

Figure 5: Gating mechanisms at $\rho = 2$. (a) The covariance-gated curve (blue) dives to its π-zero at $\theta = \pi$ ($\tau = \pi/2$); the asymmetric multimode curves do not vanish there. (b) Window position (left axis) and depth (right axis) across mechanisms.

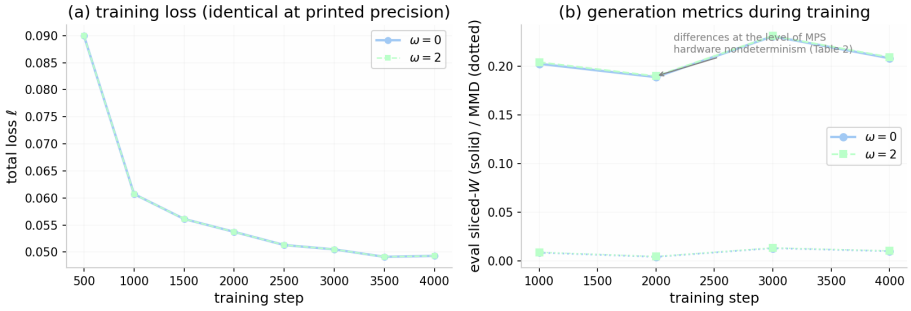

Figure 6: Training and evaluation curves of the two MNIST16 runs of Table 1. (a) The training losses are identical at printed precision: ω does not enter the reduced training loop (Corollary 1). (b) The generation metrics evaluated during training differ only at the level of hardware nondeterminism of the MPS backend, quantified in Table 1.

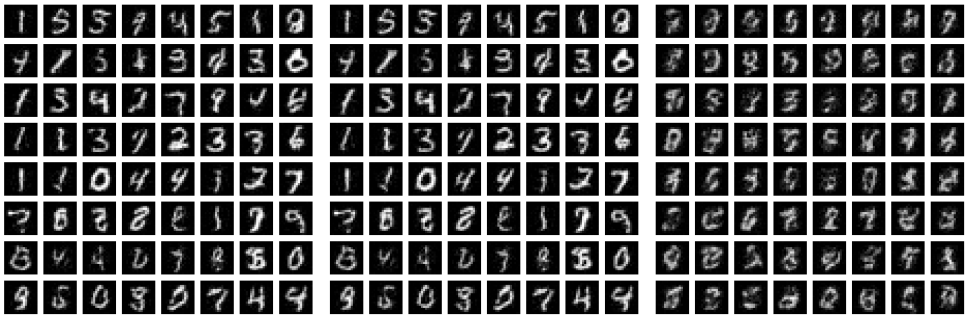

Figure 7: MNIST16 samples at $d{=}256$ (left to right): $\omega{=}0$; $\omega{=}2$ with the closed-form reduction; $\omega{=}2$ with the direct-simulation control. The first two panels are generated from the same prior draw and the same caloric-frame Brownian increments and agree up to the hardware nondeterminism of MPS training (Table 1) and the residual pullback rotation $\omega\tau(s_{\mathrm{floor}}) \approx 0.02$ rad, invisible at this resolution; the third uses a different data realization and agrees in distribution. The backbone is the publicly described heat-ball pipeline of Bai et al. (2026) and its sample quality is not a contribution of this paper; the figure verifies that the trained models are non-degenerate and that rotation leaves the samples visually unchanged.