Constraint-Preserving Denoising Diffusion Operators on Function Spaces

Liz Lemma   Future Detective

January 19, 2026

Abstract

Score-based diffusion models are typically defined on finite-dimensional grids and often violate structural constraints (e.g., incompressibility, boundary conditions), forcing ad hoc projections or penalties that break resolution invariance and uncertainty calibration. Building on the Denoising Diffusion Operators (DDO) framework for diffusion in Hilbert spaces, we develop a constraint-aware theory and algorithms in which both corruption and sampling live on a constrained function space H_c (e.g., divergence-free fields). We construct constrained Gaussian reference measures, define a constrained score operator as a logarithmic derivative with respect to the constrained reference, and derive a denoising score matching objective with finite loss. For sampling, we propose projected preconditioned Langevin dynamics driven by trace-class noise and prove that the constraint is preserved exactly in the continuum and under standard time discretizations. We further establish discretization-invariant convergence statements: Galerkin discretizations converge to the same constrained target measure with constants independent of spatial resolution. Experiments on 2D incompressible flow on the torus (where the Leray projector is explicit in Fourier space) validate exact incompressibility at all resolutions, high-fidelity generation of invariant measures for Navier–Stokes, and conditional sampling under sparse observations. The resulting models are valid-by-design diffusion generators for modern scientific machine learning.

Table of Contents

3. Problem setup: constrained spaces H_c, orthogonal projectors Π, and examples (divergence-free vector fields on $\mathbb{T}^2$, boundary/affine constraints). Formal statement of goals: exact constraint preservation + resolution invariance.

4. Constrained reference measures and corruption: constructing μ_0^c on H_c; equivalence conditions; optional smoothing operators A when μ lacks required regularity.

5. Constrained denoising score matching: definition of constrained score operator; tractable DSM objective; architecture constraints (projected neural operators) and range conditions.

6. Constraint-preserving sampling: projected preconditioned Langevin SPDE; invariance of ν; Euler–Maruyama and Crank–Nicolson discretizations preserving H_c.

7. Resolution-invariant convergence: Galerkin discretization framework; bounds for sampler error decomposition (score error + time discretization + annealing) with constants independent of $\dim(H_{c,h})$.

8. Lower bounds / necessity: why penalty-based methods require resolution-dependent scaling to enforce constraints; a formal separation between exact projection vs penalty enforcement under refinement.

9. Experiments: (i) unconditional generation of incompressible fields (spectral diagnostics, divergence error exactly zero), (ii) Navier–Stokes invariant measure learning and zero-shot super-resolution, (iii) conditional sampling with sparse observations (posterior mean/variance, calibration).

10. Discussion: extension to boundary constraints via lifting/affine projection; positivity via reparameterization; limitations (nonlinear manifolds), and future work (manifold constraints, learned projectors, non-Gaussian constrained references).

1 Introduction: constraints as first-class objects in function-space generative modeling; failures of penalty/projection heuristics under refinement; contributions and relation to DDO.

Constraints are not an incidental modeling preference in function-space generative modeling; they are part of the definition of the object we wish to generate. In many applications the data distribution lives on a closed subset of a Hilbert space that is most naturally described through a linear constraint. Typical examples include incompressible velocity fields (a divergence-free constraint), gauge constraints, and linear compatibility conditions induced by boundary value problems. In such settings the ambient Hilbert space H is merely a convenient container, while the meaningful state space is a closed linear subspace $H_c \subset H$. Our aim is therefore not only to learn a probability measure μ on H_c, but to construct a sampler that produces realizations $\tilde{u} \in H_c$ almost surely, and does so in a way that remains stable under changes of discretization.

A central difficulty is that most generative methodologies are designed in finite dimensions and only later transported to function spaces by discretization. If constraints are treated heuristically on each grid—for instance by penalizing constraint violation in a loss, or by post-processing outputs through a discrete projection—then the resulting procedure is typically not invariant under refinement. The reason is structural: as the discretization becomes finer, the number of directions in which a candidate sample can violate the constraint grows, and these directions generally populate higher-frequency modes. A penalty of the form $\lambda\|Lu\|^2$ (for a linear constraint operator L) can suppress some violations at a fixed resolution, but without a resolution-dependent scaling of λ it cannot uniformly control the violation over the proliferating high-frequency modes. Conversely, a projection step performed at the discrete level may enforce a discrete constraint, but the family of discrete constraints need not converge to the continuum constraint in a manner compatible with the learned drift or the injected noise; in particular, discretizations can introduce spurious nullspaces and aliasing effects that reappear as constraint leakage when one changes the grid, the truncation, or the numerical representation of the constraint. Thus, while penalty and projection heuristics can be effective engineering tools at a single resolution, they do not, by themselves, produce a principled continuum sampler on H_c with guarantees that persist as $h \to 0$.

We take the viewpoint that constraints must be first-class objects at the level of the continuum model. Concretely, we define learning and sampling directly on the constrained space H_c, and we require every component of the construction—reference noise, score/drift parameterization, and sampling dynamics—to have range in H_c. This "constraint-by-construction" requirement is not cosmetic: it is the only mechanism that can ensure exact feasibility $\tilde{u} \in H_c$ almost surely, uniformly over discretizations, without tuning resolution-dependent penalty parameters. It also aligns with the mathematical structure of diffusion-based generative modeling in Hilbert spaces, where the reference noise is typically Gaussian and the fundamental object learned is a logarithmic derivative (a score) defined with respect to a Gaussian reference measure. If the reference measure is chosen on the wrong space, or if the learned drift injects energy into directions orthogonal to H_c, then feasibility cannot be maintained except by ad hoc correction.

Our starting point is therefore to replace the ambient Gaussian reference $\mu_0 = \mathcal{N}(0, C)$ on H by a constrained Gaussian reference on H_c. This entails constructing a covariance operator compatible with the constraint, and working with the associated Cameron–Martin geometry intrinsic to H_c. We then consider a constrained corruption process (additive, or optionally preceded by a smoothing operator), which induces a perturbed measure ν on H_c. The key observation is that once the reference measure and the perturbation are both defined on H_c, one may formulate denoising score matching (DSM) and Langevin sampling entirely within the constrained space, thereby avoiding the spurious degrees of freedom that appear when one works on H and later attempts to project.

The contributions developed in this work can be summarized as follows. First, we provide conditions under which the perturbed measure ν is equivalent to the constrained Gaussian reference, so that a Radon–Nikodym density $\frac{d\nu}{d\mu_0^c}$ exists and admits a log-density potential Φ with a well-defined Fréchet derivative along the constrained Cameron–Martin directions. This is the measure-theoretic prerequisite for any score-based method in infinite dimensions: without equivalence and integrability of the logarithmic derivative, the score is not a meaningful object and the DSM objective may fail to be finite.

Second, we establish the constrained DSM identity that justifies learning the score (or, equivalently, a noise-prediction field) using pairs (u, η) with $u \sim \mu$ and η sampled from the constrained reference. The identity is the constrained analogue of the familiar Vincent-type relation: the minimizer of a score-matching objective under ν coincides with the minimizer of a denoising objective expressed through the conditional distribution of the corruption. In the constrained setting, the statement must be formulated in the correct dual space and then mapped back through the Riesz isomorphism associated with the constrained Cameron–Martin space; this is precisely where the function-space formulation matters, and where naive discretize-then-learn approaches obscure the geometry.

Third, we formulate a projected, preconditioned Langevin dynamics whose drift and noise both lie in H_c. Because H_c is a linear subspace, invariance of feasibility is immediate at the continuum level, and it persists under standard time discretizations (Euler–Maruyama, Crank–Nicolson) provided the increments are also constrained. This yields a sampler that preserves the constraint exactly at every iteration, rather than merely in expectation or up to numerical tolerance.

Fourth, we address resolution invariance. We adopt a continuum-first computational model: discretizations are introduced only to implement the constrained dynamics and the learned operator, via Galerkin subspaces $H_{c,h} \subset H_c$ with compatible projectors. Under standard assumptions on trace-class noise and stability of the drift, we obtain convergence of the induced invariant measures ν_h to ν in W_2, with constants that do not deteriorate with $\dim(H_{c,h})$. This provides a precise meaning to the claim that training and sampling can be performed on different discretizations without losing the constraint or changing the target measure in the limit.

Finally, we clarify why penalty-only enforcement is insufficient for uniform guarantees: to control constraint violation uniformly under refinement, the penalty parameter must scale with the largest eigenvalues associated with the constraint operator, which typically diverge as $h \to 0$. This observation motivates projection/range restriction as a necessity rather than a convenience in resolution-robust constrained generative modeling.

These considerations place our approach in direct relation to the DDO methodology in Hilbert spaces: we retain the measure-theoretic foundation (Gaussian reference measures, Cameron–Martin derivatives, DSM, and preconditioned Langevin sampling) but relocate the entire construction to the constrained state space. The subsequent background section recalls the required DDO ingredients in the unconstrained setting, after which we show how each piece is modified—and, in certain places, simplified—when constraints are treated as first-class objects.

2 Background: DDO in Hilbert spaces; Gaussian reference measures, Cameron–Martin spaces, denoising score matching, and preconditioned Langevin sampling.

In order to motivate the constrained construction, we recall the basic ingredients of drift–diffusion operator (DDO) learning in a real separable Hilbert space H with a centered Gaussian reference measure. The principal point is that, in infinite dimensions, the "score" is not an H-valued gradient in general; rather it is a logarithmic derivative defined with respect to a Gaussian measure and only along its Cameron–Martin directions. The DDO formalism makes this geometry explicit and thereby yields well-posed learning objectives and sampling dynamics.

Let $\mu_0 = \mathcal{N}(0, C)$ be a centered Gaussian measure on H with covariance

$C : H \to H$ self-adjoint, strictly positive, and trace-class. The associated Cameron–Martin space is

$$H_{\mu_0} := \mathrm{Im}(C^{1/2}) \subset H,$$

equipped with inner product $\langle h_1, h_2 \rangle_{\mu_0} := \langle C^{-1/2} h_1, C^{-1/2} h_2 \rangle$ and norm $\|h\|_{\mu_0} := \|C^{-1/2} h\|$. As is standard, H_{μ_0} is continuously embedded into H, but it is typically a strict subset; for μ_0-a.e. $x \in H$ one has $x \notin H_{\mu_0}$. This mismatch is the basic reason that gradients and divergences must be formulated carefully: differentiation is meaningful along H_{μ_0}, not in arbitrary directions of H.

Suppose ν is a probability measure on H such that $\nu \ll \mu_0$, and write

$$\frac{d\nu}{d\mu_0}(u) = \exp(\Phi(u)),$$

for a μ_0-a.s. defined potential Φ. The relevant notion of score is the Fréchet derivative of Φ along Cameron–Martin directions: for $h \in H_{\mu_0}$,

$$D_{H_{\mu_0}} \Phi(u)[h] := \lim_{\varepsilon \to 0} \frac{\Phi(u + \varepsilon h) - \Phi(u)}{\varepsilon},$$

when this limit exists and defines a bounded linear functional of h. Thus $D_{H_{\mu_0}} \Phi(u) \in H_{\mu_0}^*$ for μ_0-a.e. u. Identifying $H_{\mu_0}^* \cong H_{\mu_0}$ via the Riesz map $R : H_{\mu_0}^* \to H_{\mu_0}$, one may equivalently work with the preconditioned score $R D_{H_{\mu_0}} \Phi(u) \in H_{\mu_0} \subset H$. In many constructions it is precisely this Riesz-mapped object that appears as the non-linear component of a drift, hence the terminology "preconditioned."

The measure-theoretic role of $D_{H_{\mu_0}} \Phi$ is clarified by the Gaussian integration-by-parts identity: for sufficiently regular cylindrical test functions f and $h \in H_{\mu_0}$,

$$\int_H \langle C^{-1} u, h \rangle \, f(u) \, \mu_0(du) \;=\; \int_H Df(u)[h] \, \mu_0(du),$$

where $Df(u)[h]$ denotes the directional derivative. Under ν, the analogous formula involves the logarithmic derivative $D_{H_{\mu_0}} \Phi$, and one may interpret $D_{H_{\mu_0}} \Phi$ as the object that converts derivatives of test functions into expectations under ν. This is the infinite-dimensional analogue of the familiar finite-dimensional score $\nabla \log p$.

DDO learning is based on accessing ν through a corruption mechanism for which conditional scores are tractable. A canonical choice is additive Gaussian corruption. Let μ be a "data" measure and consider $u \sim \mu$, $\eta \sim \mu_0$ independent, and $v = u + \eta$. The induced law ν of v is the Gaussian convolution $\nu = \mu * \mu_0$. In infinite dimensions, absolute continuity $\nu \ll \mu_0$ is not automatic: for example, translating μ_0 by $u \notin H_{\mu_0}$ produces a singular measure. A standard sufficient condition is that μ is supported on H_{μ_0}, or, more generally, that one corrupts $Au + \eta$ for a bounded linear smoothing operator A mapping into H_{μ_0}. Under such hypotheses one obtains $\nu \sim \mu_0$, hence a well-defined Φ and logarithmic derivative.

The denoising score matching (DSM) identity provides an objective that avoids direct access to Φ. Formally, the "score-matching" problem is to approximate $D_{H_{\mu_0}}\Phi$ by a measurable map $G_\theta : H \to H^*_{\mu_0}$ in the sense of minimizing

$$\mathcal{J}(\theta) := \mathbb{E}_{v\sim\nu}\big\|D_{H_{\mu_0}}\Phi(v) - G_\theta(v)\big\|^2_{H^*_{\mu_0}},$$

whenever the right-hand side is finite. The DSM reformulation replaces the unknown score by a conditional expectation involving the corruption. In its most common (and practically convenient) form, one learns a preconditioned H-valued field F_θ such that $F_\theta(v) \approx RD_{H_{\mu_0}}\Phi(v)$, and one trains via noise prediction:

$$\min_\theta \; \mathbb{E}_{u\sim\mu,\, \eta\sim\mu_0}\big\|\eta + F_\theta(u+\eta)\big\|^2,$$

possibly with variance scaling when multiple noise levels are used. The equivalence between these objectives is a conditional-score calculation: the minimizer of the denoising objective is the conditional mean of $-\eta$ given v, which can be rewritten (after applying the appropriate covariance and Riesz identifications) as the score of the marginal ν. The key point for our purposes is that the objects being learned and the norms in which they are compared are dictated by μ_0 and its Cameron–Martin geometry, not by an arbitrary H-gradient.

Sampling is performed by a preconditioned Langevin dynamics designed to leave ν invariant. One convenient formulation is the Langevin SPDE

$$\dot{u}(t) = -u(t) + RD_{H_{\mu_0}}\Phi(u(t)) + \sqrt{2}\,\dot{W}(t),$$

where W is a C-Wiener process on H. Under standard Lipschitz and growth conditions on the drift (and trace-class assumptions ensuring that W is well-defined in H), this equation admits a unique strong solution and has ν as an invariant measure. In computation one typically discretizes time, for instance by Euler–Maruyama,

$$u_{n+1} = u_n + \Delta t\,\big(-u_n + F_\theta(u_n)\big) + \sqrt{2\Delta t}\,\xi_n, \qquad \xi_n \sim \mathcal{N}(0, C),$$

and uses annealing across noise scales to improve mixing and to match the time-dependent corruption used in training. The DDO perspective is that, provided one has learned $F_\theta \approx RD_{H_{\mu_0}}\Phi$ in an $L^2(\nu)$-sense uniformly across scales, the induced Markov chain approximates the intended invariant law in a way that can be analyzed without dimension-dependent constants, because the noise is trace-class and the drift is formulated in the correct preconditioned coordinates.

This unconstrained background will be used as a template: to obtain exact feasibility and resolution stability in the presence of linear constraints, we will replicate each step—reference Gaussian, Cameron–Martin derivative, DSM identity, and Langevin sampling—after relocating all objects from H to the constrained state space and enforcing range conditions at the level of the continuum formulation.

3 Problem setup: constrained spaces H_c, orthogonal projectors Π, and examples (divergence-free vector fields on $\mathbb{T}^2$, boundary/affine constraints). Formal statement of goals: exact constraint preservation + resolution invariance.

We now introduce the constrained state space on which both learning and sampling will be carried out. Let H be a real separable Hilbert space and let $H_c \subset H$ be a closed linear subspace encoding feasibility. We denote by $\Pi : H \to H_c$ the orthogonal projector. The basic structural facts we use are that Π is bounded, self-adjoint, and idempotent ($\Pi^2 = \Pi$), and that every $u \in H$ admits a unique orthogonal decomposition

$$u = \Pi u + (I - \Pi)u, \qquad \Pi u \in H_c, \quad (I - \Pi)u \in H_c^{\perp}.$$

In particular, constraint preservation is automatic for any evolution whose drift and noise take values in H_c: if $u_0 \in H_c$ and increments lie in H_c, then all iterates remain in H_c by closure and linearity. The role of Π in what follows is therefore not cosmetic; it is the device by which we ensure that every object we construct has range contained in the feasible subspace.

We keep the viewpoint that the constraint is linear and homogeneous, so that H_c is a linear subspace rather than a manifold. This covers the most common PDE constraints used in practice and is the setting in which orthogonal projection is well-defined without further choices. When constraints are expressed by a bounded linear operator $L : H \to Z$ into another Hilbert space Z, a canonical instance is

$$H_c = \ker(L),$$

with Π the orthogonal projection onto $\ker(L)$. We emphasize that the closedness of H_c is essential: it guarantees existence of Π and allows us to interpret feasibility as an exact property in H, not merely pointwise on a discretization.

A primary example, which motivates much of the discussion, is the divergence-free constraint for incompressible flow on the periodic torus. Let $H = L^2(\mathbb{T}^2; \mathbb{R}^2)$ with the usual L^2 inner product, and define

$$H_c := \{u \in L^2(\mathbb{T}^2; \mathbb{R}^2) : \nabla \cdot u = 0 \text{ in } \mathcal{D}'(\mathbb{T}^2)\}.$$

Then H_c is closed in H, and Π is the Leray–Helmholtz projection, characterized in Fourier variables by removing the component parallel to each wavevector $k \neq 0$. This example highlights why enforcing feasibility at the continuum level matters: divergence is a distributional constraint and is not faithfully represented by pointwise penalties on a grid unless one resolves all relevant modes and enforces compatibility conditions at the discrete level.

A second class of examples consists of boundary and affine constraints. Suppose H is a Sobolev space $H^1(\Omega)$ (or a product space for vector-valued fields) and let $\Gamma \subset \partial\Omega$ be a portion of the boundary. Homogeneous Dirichlet conditions can be encoded as the closed subspace

$$H_c := \{u \in H^1(\Omega) : \mathrm{Tr}_\Gamma\, u = 0\},$$

where Tr_Γ is the trace operator. More generally, affine constraints of the form $u \in u_\star + H_c$ can be reduced to the linear case by recentering: writing $u = u_\star + w$ with $w \in H_c$, we perform learning and sampling for w and then translate back. In practice this includes fixed mean, fixed flux, or prescribed boundary data, provided we can represent the corresponding projector (or at least a stable discrete projector) in the chosen numerical scheme.

Our objective is to design a generative procedure that outputs $\tilde{u} \in H_c$ almost surely while remaining stable under discretization refinement. We formulate this as two coupled requirements.

1. *Exact feasibility at every level.* The learned vector field (drift, score approximation, or noise prediction) must map into H_c, and the injected noise must live in H_c. In the continuum, this is achieved by construction; at the discrete level, it becomes an invariance property of the update map. Concretely, if $\widetilde{F}_\theta$ denotes an unconstrained neural operator on H, we will enforce feasibility by defining

$$F_\theta := \Pi \circ \widetilde{F}_\theta,$$

   so that $F_\theta(v) \in H_c$ for all admissible inputs v. This avoids the accumulation of constraint error over long sampling trajectories and prevents the appearance of spurious constraint-violating components that can grow with resolution.

2. *Resolution invariance.* We seek bounds on the discrepancy between the target law and the law induced by the discretized sampler whose constants do not depend on $\dim(H_{c,h})$, where $H_{c,h} \subset H_c$ is a Galerkin subspace (e.g. Fourier truncations, finite elements, or compatible mixed spaces) with projectors $\Pi_h \to I$ strongly on H_c. Informally, training on one discretization and sampling on another should be stable as $h \to 0$, provided the neural operator is discretization-consistent and the noise is sampled consistently with the constrained reference.

The second requirement is not a mere numerical preference. In constrained problems, the number of constraint-violating directions typically increases with the dimension of the discretization, and any procedure that suppresses these directions only through a fixed penalty is vulnerable to deterioration as resolution increases. For instance, if one replaces projection by a quadratic penalty $\lambda\|Lu\|^2$ in a drift or loss, then the effective stiffness required to control high-frequency violations generally scales with the largest eigenvalue of L^*L on the discretization. Thus, holding λ fixed while refining h does not yield uniform feasibility and can lead to $\mathbb{E}\|Lu\|$ bounded away from zero. By contrast, projection-based range restriction is scale-free: it removes the forbidden component exactly, independently of the number of degrees of freedom.

We therefore adopt a continuum-first methodology: all measures, derivatives, objectives, and dynamics are defined on H_c, and discretization is treated as an approximation of this constrained continuum problem. In particular, the corruption mechanism used for denoising must be supported on H_c, and the reference Gaussian used to define logarithmic derivatives must be constructed on H_c rather than on H. This forces us to revisit the measure-theoretic steps from the unconstrained setting, now with the Cameron–Martin geometry and absolute continuity formulated intrinsically on H_c. We turn next to the constrained reference measure μ_0^c, the induced corrupted law ν, and conditions guaranteeing $\nu \ll \mu_0^c$ (with optional smoothing operators A when the data measure does not have the requisite regularity).

Constrained reference Gaussian. Having fixed the feasible state space H_c, we next define a Gaussian reference measure intrinsically on H_c which will serve as the base measure for Radon–Nikodym derivatives and logarithmic derivatives. Let $\mu_0 = \mathcal{N}(0, C)$ on H, where $C : H \to H$ is self-adjoint, strictly positive, and trace-class. We define an operator on H_c by

$$C_c := \big(\Pi C \Pi\big)\big|_{H_c} : H_c \to H_c.$$

Since Π is bounded and C is trace-class, C_c is trace-class on H_c; it is also self-adjoint and nonnegative. We assume (and this holds in the standard PDE examples of interest) that C_c is strictly positive on H_c so that $\mathcal{N}(0, C_c)$ is a nondegenerate Gaussian measure on H_c. Equivalently, we may characterize μ_0^c as the law of Πw when $w \sim \mathcal{N}(0, C)$; then $\Pi w \in H_c$ almost surely and the induced covariance is precisely $\Pi C \Pi$ restricted to H_c. We denote this constrained reference by

$$\mu_0^c := \mathcal{N}(0, C_c) \quad \text{on } H_c.$$

Cameron–Martin geometry on H_c. Let

$$H_{\mu_0^c} := \mathrm{Im}\big(C_c^{1/2}\big) \subset H_c$$

be the Cameron–Martin space of μ_0^c, endowed with inner product

$$\langle h_1, h_2\rangle_{H_{\mu_0^c}} := \big\langle C_c^{-1/2} h_1,\, C_c^{-1/2} h_2\big\rangle_{H_c}, \qquad h_1, h_2 \in H_{\mu_0^c},$$

and associated norm $\|h\|^2_{H_{\mu_0^c}} = \langle h, h\rangle_{H_{\mu_0^c}}$. In particular, for each $h \in H_{\mu_0^c}$, the translate $\mu_0^c(\cdot - h)$ is equivalent to μ_0^c (Cameron–Martin theorem). Conversely, if $h \notin H_{\mu_0^c}$, then $\mu_0^c(\cdot - h)$ and μ_0^c are mutually singular. This dichotomy is the basic reason that we must place regularity assumptions on the shifts appearing in the corruption mechanism.

Constrained corruption and the induced perturbed law. Let μ be the data measure on H_c. We consider the constrained additive corruption

$$v = u + \eta, \qquad u \sim \mu, \quad \eta \sim \mu_0^c, \quad u \perp\!\!\!\perp \eta,$$

which induces the convolution measure $\nu := \mu * \mu_0^c$ on H_c. Conditionally on u, we have $v \sim \mathcal{N}(u, C_c)$ as a Gaussian measure on H_c. The objective in what follows is to define the logarithmic density $\Phi = \log \frac{d\nu}{d\mu_0^c}$ and to interpret its derivative along $H_{\mu_0^c}$; this requires at minimum $\nu \ll \mu_0^c$.

Absolute continuity and equivalence: the role of $H_{\mu_0^c}$. A sufficient condition for $\nu \ll \mu_0^c$ is that the random shifts u lie in the Cameron–Martin space:

$$\mu(H_{\mu_0^c}) = 1.$$

Indeed, for μ-almost every u, the conditional law $\mathcal{N}(u, C_c)$ is equivalent to μ_0^c, and hence any μ_0^c-null set is $\mathcal{N}(u, C_c)$-null. Integrating out u yields $\nu \ll \mu_0^c$. Moreover, under this assumption we can write an explicit, μ_0^c-a.s. strictly positive density. Let $h \in H_{\mu_0^c}$. The Cameron–Martin formula gives

$$\frac{d\mathcal{N}(h, C_c)}{d\,\mu_0^c}(v) = \exp\Big(\big\langle C_c^{-1} h,\, v\big\rangle_{H_c} - \frac{1}{2}\|h\|^2_{H_{\mu_0^c}}\Big), \qquad \mu_0^c\text{-a.s. } v \in H_c,$$

where $C_c^{-1}h$ is well-defined since $h \in \mathrm{Im}(C_c^{1/2})$. Consequently,

$$\frac{d\nu}{d\mu_0^c}(v) = \int_{H_c} \exp\Big(\big\langle C_c^{-1} u,\, v\big\rangle_{H_c} - \frac{1}{2}\|u\|^2_{H_{\mu_0^c}}\Big)\, \mu(du), \qquad \mu_0^c\text{-a.s.} \tag{1}$$

The integrand is strictly positive, hence the mixture density in (1) is strictly positive μ_0^c-almost surely, which implies not only $\nu \ll \mu_0^c$ but in fact $\nu \sim \mu_0^c$. We then define

$$\Phi(v) := \log \frac{d\nu}{d\mu_0^c}(v), \qquad \mu_0^c\text{-a.s. } v \in H_c.$$

When the data are rough: smoothing operators. In many infinite-dimensional inverse problems and PDE models, μ is supported on states rougher than $H_{\mu_0^c}$; in that case the condition $\mu(H_{\mu_0^c}) = 1$ fails, and the singularity part of the Cameron–Martin theorem precludes $\nu \ll \mu_0^c$ for the naive corruption $v = u + \eta$. A standard remedy is to modify the corruption by inserting a bounded linear smoothing operator

$$A : H_c \to H_{\mu_0^c}, \qquad \text{with } A(H_c) \subset H_{\mu_0^c},$$

and to corrupt via

$$v = Au + \eta.$$

Then $Au \in H_{\mu_0^c}$ almost surely regardless of whether $u \in H_{\mu_0^c}$, and the same argument as above applies with u replaced by Au. In particular, the induced law ν of v satisfies $\nu \sim \mu_0^c$, and its density admits the representation

$$\frac{d\nu}{d\mu_0^c}(v) = \int_{H_c} \exp\Big(\big\langle C_c^{-1} Au,\, v\big\rangle_{H_c} - \frac{1}{2}\|Au\|^2_{H_{\mu_0^c}}\Big)\, \mu(du), \qquad \mu_0^c\text{-a.s.}$$

Conceptually, A is a device ensuring that the random shift in the Gaussian mixture lies in the Cameron–Martin space of the reference noise. Typical choices are elliptic regularizers (e.g. fractional powers of $I - \Delta$ on periodic domains) or observation/measurement operators which are smoothing by construction. We stress that A is used here solely to place the corrupted law in the same measure class as μ_0^c; it is not a penalty and it does not relax feasibility, since $A(H_c) \subset H_c$ and $\eta \in H_c$ imply $v \in H_c$ almost surely.

Noise scales and compatibility with later annealing. For multi-scale denoising it is convenient to consider a family of constrained Gaussians $\mu_{0,t}^c := \mathcal{N}(0, tC_c)$ for $t \in I \subset (0, \infty)$, and corresponding corrupted laws $\nu_t := \mu * \mu_{0,t}^c$ (or with Au). The same Cameron–Martin reasoning applies uniformly in t after rescaling the Cameron–Martin norm by $t^{-1/2}$, yielding $\nu_t \sim \mu_{0,t}^c$ and hence a well-defined potential $\Phi_t = \log \frac{d\nu_t}{d\mu_{0,t}^c}$. Under the standing assumption that Φ_t is differentiable along $H_{\mu_0^c}$ with square-integrable logarithmic derivative, we may proceed to define the constrained score operator and derive a denoising score matching objective on H_c.

Constrained score operator and its preconditioned representative. Since $\nu \sim \mu_0^c$ we may regard $\Phi = \log \frac{d\nu}{d\mu_0^c}$ as a μ_0^c-a.s. defined real-valued functional on H_c. The relevant notion of differentiability is along the Cameron–Martin directions: for $v \in H_c$ and $h \in H_{\mu_0^c}$, we define the directional derivative

$$D_{H_{\mu_0^c}} \Phi(v)[h] := \lim_{\varepsilon \to 0} \frac{\Phi(v + \varepsilon h) - \Phi(v)}{\varepsilon},$$

whenever the limit exists. Under our standing assumption that Φ is Fréchet differentiable along $H_{\mu_0^c}$, the map $h \mapsto D_{H_{\mu_0^c}} \Phi(v)[h]$ is a bounded linear functional on $H_{\mu_0^c}$, hence an element of the dual space $H^*_{\mu_0^c}$. We denote this logarithmic derivative (the constrained score) by

$$D_{H_{\mu_0^c}}\Phi(v) \in H^*_{\mu_0^c}, \qquad D_{H_{\mu_0^c}}\Phi(v)[h] = \big\langle D_{H_{\mu_0^c}}\Phi(v),\, h\big\rangle_{H^*_{\mu_0^c}, H_{\mu_0^c}}.$$

To obtain a vector field taking values in a Hilbert space, we identify $H^*_{\mu_0^c} \cong H_{\mu_0^c}$ via the Riesz map $R : H^*_{\mu_0^c} \to H_{\mu_0^c}$, characterized by

$$\big\langle \ell,\, h\big\rangle_{H^*_{\mu_0^c}, H_{\mu_0^c}} = \big\langle R\ell,\, h\big\rangle_{H_{\mu_0^c}}, \qquad \ell \in H^*_{\mu_0^c},\ h \in H_{\mu_0^c}.$$

We will work with the preconditioned constrained score

$$v \longmapsto RD_{H_{\mu_0^c}}\Phi(v) \in H_{\mu_0^c} \subset H_c,$$

and with the associated drift component

$$F(v) := -v + RD_{H_{\mu_0^c}}\Phi(v) \in H_c, \tag{2}$$

which is the object entering the projected Langevin dynamics.

A tractable constrained DSM objective. Directly minimizing a score-matching objective in the dual norm $\|\cdot\|_{H^*_{\mu_0^c}}$ is conceptually natural but not computationally convenient. The denoising formulation yields an explicit target. Conditionally on u, the corrupted sample v has law $\gamma_u := \mathcal{N}(u, C_c)$ on H_c, and its Radon–Nikodym derivative with respect to μ_0^c is given by the Cameron–Martin formula. Writing

$$\Psi(v; u) := \log \frac{d\gamma_u}{d\mu_0^c}(v) = \big\langle C_c^{-1}u,\, v\big\rangle_{H_c} - \frac{1}{2}\|u\|^2_{H_{\mu_0^c}},$$

we obtain, for each fixed $u \in H_{\mu_0^c}$, the conditional score along $H_{\mu_0^c}$,

$$D_{H_{\mu_0^c}}\Psi(v; u) \in H^*_{\mu_0^c}, \qquad D_{H_{\mu_0^c}}\Psi(v; u)[h] = \big\langle C_c^{-1}u,\, h\big\rangle_{H_c}.$$

Applying the Riesz map yields the particularly simple identity

$$RD_{H_{\mu_0^c}}\Psi(v; u) = u \in H_{\mu_0^c}. \tag{3}$$

The constrained denoising score matching (DSM) objective is then to fit a measurable map $G_\theta : H_c \to H^*_{\mu_0^c}$ (or equivalently its Riesz representative) by minimizing

$$\mathcal{L}_{\mathrm{DSM}}(\theta) := \mathbb{E}_{u\sim\mu,\ v\sim\gamma_u}\Big\| D_{H_{\mu_0^c}}\Psi(v; u) - G_\theta(v)\Big\|^2_{H^*_{\mu_0^c}}, \tag{4}$$

which is finite under our assumption that $D_{H_{\mu_0^c}}\Phi \in L^2(\nu; H^*_{\mu_0^c})$ (and similarly for the conditional score). The standard conditional-expectation argument implies that any minimizer satisfies $G_\theta(v) = D_{H_{\mu_0^c}}\Phi(v)$ in $L^2(\nu; H^*_{\mu_0^c})$.

Noise-prediction form and the target drift. In practice we parameterize the drift (2) rather than the dual-valued score. Using (3), the conditional expectation of the preconditioned score is

$$RD_{H_{\mu_0^c}}\Phi(v) = \mathbb{E}\big[\, u \,\big|\, v \big], \qquad \nu\text{-a.s. } v \in H_c,$$

so that the drift component becomes

$$F(v) = -v + \mathbb{E}\big[\, u \,\big|\, v \big] = \mathbb{E}\big[\, u - v \,\big|\, v \big] = -\mathbb{E}\big[\, \eta \,\big|\, v \big]. \tag{5}$$

Consequently, if we parameterize $F_\theta : H_c \to H_c$ and minimize the noise-prediction loss

$$\mathcal{L}_{\mathrm{NP}}(\theta) := \mathbb{E}_{u\sim\mu,\ \eta\sim\mu_0^c} \big\| \eta + F_\theta(u+\eta) \big\|_{H_c}^2, \tag{6}$$

then the unique minimizer in $L^2(\nu; H_c)$ is $F_\theta = F$, i.e. $F_\theta(v) = -\mathbb{E}[\eta \mid v]$. This is the constrained analogue of the classical Vincent identity, with the key distinction that all objects are defined intrinsically on H_c and logarithmic derivatives are taken only along $H_{\mu_0^c}$.

Noise scales and a single family of range conditions. For a family of noise levels $t \in I \subset (0, \infty)$ we consider $\eta_t \sim \mathcal{N}(0, tC_c)$, $v = u + \eta_t$, and write $\Phi_t = \log \frac{d\nu_t}{d\mu_{0,t}^c}$ for $\nu_t = \mu * \mu_{0,t}^c$. The same derivation yields

$$F(v,t) := -v + RD_{H_{\mu_0^c}}\Phi_t(v) = -\mathbb{E}[\eta_t \mid v],$$

and we train by sampling t and minimizing $\mathbb{E}\|\eta_t + F_\theta(v,t)\|_{H_c}^2$ with the appropriate scaling absorbed into the sampling of η_t (or, equivalently, by rescaling the loss).

Architecture constraints via projection and range restriction. To guarantee exact feasibility of both training targets and sampled trajectories, we enforce the range condition $F_\theta(\cdot, t) \in H_c$ by construction. Concretely, we let $\widetilde{F}_\theta : H_c \times I \to H$ be an unconstrained neural operator and define

$$F_\theta(\cdot, t) := \Pi \circ \widetilde{F}_\theta(\cdot, t) : H_c \to H_c.$$

This removes any dependence of feasibility on discretization or resolution, and it prevents the learned drift from injecting energy into constraint-violating directions which are absent at coarse resolution but proliferate under refinement. In settings where we additionally require $F_\theta(\cdot, t) \in H_{\mu_0^c}$ (so that the drift coincides with the Riesz representative of a dual score), we may append a fixed bounded smoothing layer mapping $H_c \to H_{\mu_0^c}$ before projection; however, for the projected Langevin dynamics it suffices that F_θ takes values in H_c and is square-integrable (and, for well-posedness, Lipschitz or suitably dissipative). With these choices, the learned field F_θ is compatible with the constraint-preserving sampling dynamics developed next.

Constraint-preserving sampling via projected preconditioned Langevin. Having identified the target drift field F on H_c, we sample from ν by evolving a Langevin dynamics intrinsically on the constraint subspace. Let W be a C-Wiener process on H and set $W^c := \Pi W$, so that W^c is a C_c-Wiener process on H_c. We consider the stochastic evolution equation on H_c

$$dU_s = F(U_s)\,ds + \sqrt{2}\,dW_s^c = \big(-U_s + \Pi R D_{H_{\mu_0^c}}\Phi(U_s)\big)\,ds + \sqrt{2}\,dW_s^c. \quad (7)$$

Under standard dissipativity and local Lipschitz assumptions on $u \mapsto \Pi R D_{H_{\mu_0^c}}\Phi(u)$ (or, more generally, monotonicity plus polynomial growth), (7) admits a unique strong solution with continuous paths in H_c. Moreover, feasibility is automatic: since H_c is a closed linear subspace, the linear semigroup generated by $-I$ leaves H_c invariant, and both the nonlinear drift term and the noise term take values in H_c. In particular, writing the mild form

$$U_s = e^{-s}U_0 + \int_0^s e^{-(s-r)}\Pi R D_{H_{\mu_0^c}}\Phi(U_r)\,dr + \sqrt{2}\int_0^s e^{-(s-r)}\,dW_r^c,$$

we see term-by-term that $U_s \in H_c$ for all $s \geq 0$ whenever $U_0 \in H_c$.

Invariant measure. The dynamics (7) is the constrained analogue of the classical (preconditioned) Langevin diffusion with Gaussian reference. Formally, its infinitesimal generator acting on smooth cylindrical test functions $\varphi : H_c \to \mathbb{R}$ is

$$(\mathcal{L}\varphi)(u) = \big\langle F(u), D\varphi(u)\big\rangle_{H_c} + \mathrm{Tr}\big(C_c D^2\varphi(u)\big),$$

where $D\varphi$ and $D^2\varphi$ denote the first and second Fréchet derivatives on H_c. Since ν has density $\frac{d\nu}{d\mu_0^c}(u) = \exp(\Phi(u))$, invariance of ν can be verified by an integration-by-parts identity on the Gaussian space (H_c, μ_0^c). In essence, one rewrites the drift as the sum of the Ornstein–Uhlenbeck drift $-u$ (for which μ_0^c is invariant) and a perturbation given by the logarithmic derivative of ν relative to μ_0^c. Under our standing assumptions that $D_{H_{\mu_0^c}}\Phi \in L^2(\nu; H^*_{\mu_0^c})$ and is sufficiently regular to justify the manipulations, one obtains

$$\int_{H_c} (\mathcal{L}\varphi)(u)\,\nu(du) = 0 \qquad \text{for all smooth cylindrical } \varphi, \quad (8)$$

which identifies ν as an invariant measure for (7). When additionally the drift is of (preconditioned) gradient type, (8) strengthens to reversibility of ν with respect to the Markov semigroup generated by $\mathcal{L}$; we do not require this stronger property for constraint preservation, but it is useful when analyzing bias due to discretization or due to learned approximations $F_\theta \approx F$.

Constraint preservation at the level of time discretization. For sampling we employ time discretizations that inherit feasibility from the continuum equation. Fix a step size $\Delta t > 0$. The Euler–Maruyama scheme on H_c is

$$u_{n+1} = u_n + \Delta t\, F(u_n) + \sqrt{2\Delta t}\,\xi_n, \qquad \xi_n \sim \mathcal{N}(0, C_c) \text{ i.i.d.} \tag{9}$$

By construction, $F(u_n) \in H_c$ and $\xi_n \in H_c$ almost surely; hence, if $u_n \in H_c$ then $u_{n+1} \in H_c$. This argument is resolution-independent: it uses only the range condition on the drift and the support of the noise, not any coordinate representation. In particular, replacing F by a learned field $F_\theta = \Pi \circ \widetilde{F}_\theta$ preserves feasibility identically.

A semi-implicit Crank–Nicolson discretization is often preferable for stability, treating the linear contraction $-u$ implicitly while keeping the nonlinear term explicit. Writing $B(u) := \Pi R D_{H_{\mu_0^c}} \Phi(u)$, we obtain

$$\Big(I + \frac{\Delta t}{2} I\Big) u_{n+1} = \Big(I - \frac{\Delta t}{2} I\Big) u_n + \Delta t\, B(u_n) + \sqrt{2\Delta t}\,\xi_n, \qquad \xi_n \sim \mathcal{N}(0, C_c). \tag{10}$$

Since $\big(I + \frac{\Delta t}{2} I\big)^{-1}$ is a bounded operator on H_c and all right-hand-side terms lie in H_c, the update is well-defined in H_c and again preserves the constraint exactly. The same conclusion holds for other splitting or preconditioned schemes provided that (i) the drift evaluation is projected into H_c and (ii) the Gaussian increment has covariance C_c.

Annealed schedules and learned drifts. For multi-noise-scale sampling we index the drift by a noise level $t \in I$ and run a sequence of dynamics with decreasing t, using either (9) or (10) with $F(\cdot, t)$ (or $F_\theta(\cdot, t)$). At each level t we draw $\xi_n^{(t)} \sim \mathcal{N}(0, C_{c,t})$ with $C_{c,t} = tC_c$, so that every incremental noise remains supported on H_c. Thus, independently of approximation error in F_θ, the iterates remain feasible almost surely:

$$u_0 \in H_c \implies u_n \in H_c \ \text{ for all } n \text{ and all annealing levels.}$$

The role of learning is therefore separated cleanly from the role of enforcing the constraint: learning controls how closely the sampler approximates ν, while projection and constrained noise enforce exact membership in H_c at every step.

Transition to discretization and resolution invariance. The preceding construction is continuum-first: the SPDE and its discretizations are formulated on H_c, and feasibility is a consequence of subspace invariance rather than a grid-dependent penalty. To quantify how close a practical implementation is to the ideal constrained sampler, we next pass to Galerkin subspaces $H_{c,h} \subset H_c$ and compare the induced invariant measures and sampling distributions across resolutions. This yields a resolution-invariant error decomposition separating score approximation, time-discretization bias, annealing bias, and Galerkin truncation error, with constants that do not depend on $\dim(H_{c,h})$.

Resolution-invariant Galerkin convergence. We now formalize the passage from the continuum sampler on H_c to implementable samplers on finite-dimensional Galerkin spaces in a manner that is stable under refinement. Let $\{H_{c,h}\}_{h>0}$ be a family of finite-dimensional subspaces of H_c with orthogonal projectors $\Pi_h : H_c \to H_{c,h}$ such that $\Pi_h \to I$ strongly on H_c as $h \to 0$. We define the restricted covariance $C_{c,h} := \Pi_h C_c \Pi_h$ (viewed as an operator on $H_{c,h}$) and let $\mu^c_{0,h} := \mathcal{N}(0, C_{c,h})$ be the corresponding Gaussian measure on $H_{c,h}$. We shall compare the continuum perturbed measure ν on H_c with the finite-dimensional invariant measures induced by Galerkin-projected dynamics.

For each h, consider the Galerkin approximation of (7) on $H_{c,h}$,

$$dU^{(h)}_s = \Pi_h F\big(U^{(h)}_s\big)\, ds + \sqrt{2}\, dW^{c,(h)}_s, \qquad W^{c,(h)} := \Pi_h W^c, \tag{11}$$

where $W^{c,(h)}$ is a $C_{c,h}$-Wiener process on $H_{c,h}$. Under the same dissipativity and Lipschitz hypotheses ensuring well-posedness of (7), the finite-dimensional SDE (11) admits a unique strong solution for any initial condition in $H_{c,h}$. We write $\mathcal{P}_s$ and $\mathcal{P}^{(h)}_s$ for the Markov semigroups of (7) and (11), respectively.

The first point is that (11) approximates the continuum dynamics in a strong sense, uniformly on bounded time intervals, with constants that can be chosen independently of $\dim(H_{c,h})$. Indeed, coupling U and $U^{(h)}$ on a common probability space using the same driving noise W^c (and setting $W^{c,(h)} = \Pi_h W^c$) yields a pathwise error equation for $E^{(h)}_s := U_s - U^{(h)}_s$ of the form

$$dE^{(h)}_s = \Big(F(U_s) - \Pi_h F(U^{(h)}_s)\Big)\, ds + \sqrt{2}\,(I - \Pi_h)\, dW^c_s.$$

The linear contraction $-u$ contained in $F(u) = -u + \Pi R D_{H_{\mu^c_0}} \Phi(u)$ yields a uniform dissipativity estimate, while the noise mismatch term is controlled by the trace-class property of C_c and the strong convergence $\Pi_h \to I$. Consequently, for each fixed $T > 0$ we obtain an estimate of the schematic form

$$\sup_{0\le s\le T} \mathbb{E}\|U_s - U^{(h)}_s\|^2 \;\le\; C_T\Big(\|(I - \Pi_h)U_0\|^2 + \mathrm{Tr}\big((I - \Pi_h)C_c\big)\Big), \tag{12}$$

where C_T depends on T and on Lipschitz/dissipativity bounds for F, but not on $\dim(H_{c,h})$. The crucial observation is that the stochastic forcing enters only through C_c, and $\mathrm{Tr}((I-\Pi_h)C_c) \to 0$ by trace-class compactness, independently of how $\dim(H_{c,h})$ grows.

We next pass from finite-time approximation to invariant measures. Let ν_h denote an invariant measure for (11). Existence follows from standard Lyapunov arguments based on dissipativity; uniqueness follows under a mild irreducibility/strong Feller condition, which holds in the present additive-noise setting provided $C_{c,h}$ is non-degenerate on $H_{c,h}$. Combining exponential mixing (in Wasserstein distance) for both the continuum and Galerkin semigroups with the coupling estimate (12), we obtain

$$W_2(\nu_h, \nu) \;\longrightarrow\; 0 \qquad \text{as } h \to 0, \tag{13}$$

and the implied constants in the mixing and stability estimates may be chosen uniformly in h. At the level of proof, the independence from $\dim(H_{c,h})$ enters through: (i) dimension-free dissipativity of the linear part $-u$; (ii) operator-norm (rather than coordinate-wise) Lipschitz bounds on $u \mapsto \Pi R D_{H_{\mu_0^c}} \Phi(u)$ restricted to H_c; and (iii) trace-class control of the noise, which prevents growth of stochastic fluctuations with the number of retained modes.

We now incorporate learned drifts and time discretization. Fix h and let $F_{\theta,h} : H_{c,h} \times I \to H_{c,h}$ be the learned drift used in practice. We assume only that it is constraint-preserving (i.e. it maps into $H_{c,h}$) and that it approximates the projected target drift in the sense that, for each noise level $t \in I$,

$$\mathrm{Err}_{\mathrm{score}}(h,\theta) := \sup_{t\in I} \Big(\mathbb{E}_{u\sim\nu_t} \big\| \Pi_h F(u,t) - F_{\theta,h}(\Pi_h u, t) \big\|^2 \Big)^{1/2} \tag{14}$$

is small. This formulation makes explicit the discretization-consistency requirement: the same operator network should represent the drift across resolutions via projection of inputs/outputs. In particular, (14) is compatible with training at one h and sampling at another, provided the architecture respects the projection structure.

Let $\widehat{\nu}_{h,\theta,\Delta t}$ denote the stationary distribution (or long-time sampling distribution) of the time-Δt Euler–Maruyama chain on $H_{c,h}$ driven by $F_{\theta,h}$ (and similarly for a Crank–Nicolson-type scheme). A standard perturbation decomposition, combined with the contractivity of the linear part and the stability of invariant measures under drift perturbations, yields an error bound of the form

$$W_2\big(\widehat{\nu}_{h,\theta,\Delta t}, \nu\big) \;\le\; C\Big(\mathrm{Err}_{\mathrm{score}}(h,\theta) + \mathrm{Err}_{\Delta t}(h) + \mathrm{Err}_{\mathrm{anneal}}(h) + \mathrm{Err}_{\mathrm{Galerkin}}(h) \Big), \tag{15}$$

where $\mathrm{Err}_{\Delta t}(h)$ denotes the bias introduced by time discretization (typically $\mathrm{Err}_{\Delta t}(h) = O(\Delta t^{1/2})$ in strong metrics and $O(\Delta t)$ in weak metrics under sufficient regularity), $\mathrm{Err}_{\mathrm{anneal}}(h)$ quantifies the discrepancy due to running a finite annealing schedule rather than exact sampling at the terminal noise level, and $\mathrm{Err}_{\mathrm{Galerkin}}(h) := W_2(\nu_h, \nu)$ is the truncation error controlled by (13). The constant C depends on continuum quantities (dissipativity, Lipschitz bounds, and $\mathrm{Tr}(C_c)$), but is independent of $\dim(H_{c,h})$. We emphasize that (15) separates the sources of error in a way that is compatible with a continuum-first methodology: Galerkin truncation affects only $\mathrm{Err}_{\mathrm{Galerkin}}$, time stepping affects only $\mathrm{Err}_{\Delta t}$, annealing affects only $\mathrm{Err}_{\mathrm{anneal}}$, and learning affects only $\mathrm{Err}_{\mathrm{score}}$.

Finally, since each term in (15) is controlled by operator-level quantities rather than coordinate dimension, the convergence is resolution-invariant in the following sense: if we refine h while maintaining comparable score accuracy in (14) (as expected for discretization-consistent neural operators) and choose Δt according to stability rather than dimension, then $W_2(\widehat{\nu}_{h,\theta,\Delta t}, \nu) \to 0$ without any constants degenerating with $\dim(H_{c,h})$. This is precisely the regime in which training and sampling can be decoupled from the particular grid used to represent the fields.

Penalty enforcement lower bounds and the necessity of exact projection. A natural alternative to constraint preservation by construction is to enforce a linear constraint via a quadratic penalty. Concretely, let $L : \mathcal{D}(L) \subset H \to G$ be a (typically unbounded) closed linear operator into a Hilbert space G, and consider the constraint $Lu = 0$ (e.g. $L = \nabla\cdot$ on vector fields). A common strategy is to work on an unconstrained discretization $H_h \subset H$ and to modify either the learning objective or the sampling drift by adding a term proportional to $\lambda \|L_h u\|_G^2$, where $L_h : H_h \to G_h$ is a discrete approximation of L and $\lambda > 0$ is a penalty parameter. We now record the basic obstruction: unless λ is increased with resolution, penalty methods cannot yield constraint violation tending to zero under refinement, even in the most favorable Gaussian/quadratic setting. This is the precise sense in which projection (or, equivalently, range-restriction to H_c) is not merely convenient but structurally necessary for resolution-invariant feasibility.

To isolate the effect, we ignore learning error and consider the idealized stationary distribution induced by a purely quadratic penalization. Let H_h be finite-dimensional, let $L_h : H_h \to G_h$ be linear, and let $\Sigma_h : H_h \to H_h$ be a symmetric positive definite "base covariance" (for simplicity one may take $\Sigma_h = I$). Consider the penalized Gaussian measure $\rho_{h,\lambda}$ on H_h with density

$$\rho_{h,\lambda}(du) \;\propto\; \exp\Big(-\tfrac{1}{2}\langle \Sigma_h^{-1} u, u\rangle - \tfrac{\lambda}{2}\|L_h u\|_{G_h}^2\Big)\, du,$$

which is the invariant measure of the Ornstein–Uhlenbeck dynamics with linear drift $-\Sigma_h^{-1} u - \lambda L_h^* L_h u$ and additive white noise. This measure is the best-case target for any penalty-based Langevin sampler: if even $\rho_{h,\lambda}$ fails to concentrate on $\ker(L_h)$ uniformly in h, then no practical scheme with finite stepsizes and additional nonlinearities can remedy the obstruction without scaling λ.

Let $A_h := \Sigma_h^{1/2} L_h^* L_h \Sigma_h^{1/2}$ on H_h. Diagonalizing A_h yields eigenpairs $\{(\alpha_{j,h}, e_{j,h})\}_{j=1}^{\dim(H_h)}$ with $\alpha_{j,h} \geq 0$. Under $\rho_{h,\lambda}$, the transformed coordinates $z = \Sigma_h^{-1/2} u$ are Gaussian with covariance $(I + \lambda A_h)^{-1}$, and hence a direct calculation gives

$$\mathbb{E}_{\rho_{h,\lambda}} \|L_h u\|_{G_h}^2 = \mathrm{Tr}\Big(L_h \Sigma_h (I + \lambda A_h)^{-1} L_h^*\Big) = \sum_{j=1}^{\dim(H_h)} \frac{\alpha_{j,h}}{1 + \lambda \alpha_{j,h}}. \tag{16}$$

The summands satisfy $\frac{\alpha}{1+\lambda\alpha} \geq \frac{1}{2\lambda}$ whenever $\alpha \geq 1/\lambda$. Therefore, if A_h has many eigenvalues above $1/\lambda$ (which is exactly what occurs under refinement for differential constraints), the expected constraint violation cannot be made small without increasing λ.

Proposition 3.1 (Resolution-dependent penalty scaling)**.** *Assume that for a sequence $h \to 0$ the operator A_h has at least m_h eigenvalues satisfying $\alpha_{j,h} \geq \kappa\, \alpha_{\max,h}$ for some fixed $\kappa \in (0, 1]$, where $\alpha_{\max,h} := \|A_h\|_{\mathrm{op}}$. Then for all $\lambda > 0$,*

$$\mathbb{E}_{\rho_{h,\lambda}} \|L_h u\|_{G_h}^2 \;\geq\; m_h \frac{\kappa\, \alpha_{\max,h}}{1 + \lambda\kappa\, \alpha_{\max,h}} \;\geq\; \frac{m_h}{2\lambda}\, \mathbf{1}_{\{\lambda\kappa\, \alpha_{\max,h} \geq 1\}}.$$

In particular, achieving $\mathbb{E}_{\rho_{h,\lambda}} \|L_h u\|_{G_h}^2 \leq \varepsilon$ uniformly in h requires $\lambda \gtrsim m_h/\varepsilon$ once $\alpha_{\max,h} \to \infty$ and $m_h \to \infty$.

The hypothesis is mild: for standard Fourier or finite element discretizations of differential operators, $\alpha_{\max,h}$ grows polynomially in h^{-1} and a nontrivial fraction of modes accumulate near the high end of the spectrum. For example, in a Fourier truncation on $\mathbb{T}^d$ with maximal wavenumber $K \sim h^{-1}$ and $L = \nabla\cdot$, one has $L_h^* L_h$ acting diagonally on modes with eigenvalues $|k|^2$. Thus $\alpha_{\max,h} \sim K^2$, and there are $m_h \asymp K^d$ modes with $|k| \gtrsim K/2$. Inserting this into Proposition 3.1 yields a lower bound scaling like K^d/λ for the unnormalized ℓ^2 constraint energy. When the discrete L^2 norm is normalized by the grid volume, the same calculation yields an $\Omega(1/\lambda)$ lower bound independent of K, so that any fixed λ produces a nonvanishing divergence level as $h \to 0$.

Two conclusions follow. First, penalty methods are intrinsically *not* resolution-invariant: to maintain a fixed feasibility tolerance ε one must increase λ with the growth of $\|L_h^* L_h\|_{\mathrm{op}}$ and, in typical cases, also with the number of constraint-violating directions m_h. Second, increasing λ introduces stiffness: the penalized drift contains the factor $\lambda L_h^* L_h$, whose largest eigenvalue is $\lambda\alpha_{\max,h}$; explicit time stepping then requires $\Delta t \ll (\lambda\alpha_{\max,h})^{-1}$ for stability, which deteriorates rapidly under refinement. Thus penalty-based training and sampling jointly face a twofold resolution dependence: the coefficient must scale up to control feasibility, and the stepsize must scale down to control stiffness.

By contrast, exact projection separates feasibility from resolution. If we evolve directly on the constrained subspace H_c (or on $H_{c,h}$) and enforce $F_\theta = \Pi \circ \widetilde{F}_\theta$, then constraint satisfaction holds identically at every step and for every h, without tuning coefficients against $\alpha_{\max,h}$ and without introducing stiff directions. This is the formal separation we rely on: projection eliminates the growing family of constraint-violating modes rather than merely suppressing them, and therefore is the only mechanism among these two that can support feasibility with constants that do not degenerate as $\dim(H_{c,h}) \to \infty$.

Numerical experiments. We summarize three representative experiments designed to stress (a) exact feasibility at every iteration, (b) resolution invariance under Galerkin refinement, and (c) compatibility with conditioning. In all cases we train the constrained noise-prediction objective

$$\min_\theta \ \mathbb{E}_{u\sim\mu,\ \eta\sim\mu_0^c}\big\|\eta + F_\theta(u+\eta,t)\big\|_H^2, \qquad F_\theta := \Pi \circ \widetilde{F}_\theta,$$

with t either fixed (single-scale) or drawn from a prescribed annealing schedule. The constrained Gaussian $\mu_0^c = \mathcal{N}(0, C_c)$ is sampled on the chosen Galerkin space $H_{c,h}$ by drawing $z \sim \mathcal{N}(0, I)$ and setting $\eta = C_{c,h}^{1/2} z$; on periodic domains we implement $C_{c,h}^{1/2}$ and Π_h spectrally (FFT), while on non-periodic benchmarks we implement Π_h by an $H(\mathrm{div})$-compatible discrete Helmholtz projection. Sampling uses the annealed projected Euler–Maruyama scheme in the algorithmic context, so that $u_n \in H_{c,h}$ for all n by construction.

(i) Unconditional generation of incompressible fields. We first consider unconditional generation of divergence-free velocity fields on $\mathbb{T}^d$, where the constraint is $H_c = \{u \in L^2(\mathbb{T}^d;\mathbb{R}^d) : \nabla \cdot u = 0,\ \int u = 0\}$. Training data are i.i.d. snapshots $\{u_i\}$ from a prescribed target distribution on H_c (either synthetic turbulence-like spectra or snapshots from a reference solver). We evaluate samples $\tilde{u}$ using spectral diagnostics that are standard for incompressible flow statistics: the energy spectrum $E(k)$, the enstrophy spectrum, and low-order structure functions $S_p(\ell) = \mathbb{E}|\delta_\ell u|^p$ with $\delta_\ell u(x) = u(x+\ell) - u(x)$. In addition, we monitor distributional metrics on low-dimensional summaries (e.g. histograms of vorticity and energy) and two-point correlation functions to test whether long-range coherence is reproduced.

The feasibility check is explicit: since F_θ takes values in H_c and the injected noise $\xi_n \sim \mathcal{N}(0, C_{c,h})$ is also supported on H_c, each Langevin iterate satisfies $\nabla \cdot u_n = 0$ in the discrete sense implied by Π_h. Consequently, the divergence diagnostic (measured as $\|\nabla \cdot \tilde{u}\|_{L^2}$ on the evaluation grid) is at machine precision across all sampling steps and across all evaluation resolutions. This confirms the main qualitative claim: in contrast to penalty methods, there is no resolution-dependent tuning required to keep feasibility fixed under refinement; feasibility is an invariant of the dynamics rather than an emergent property.

(ii) Learning an invariant measure for Navier–Stokes and zero-shot super-resolution. We next target the invariant measure of the 2D incompressible Navier–Stokes equations on $\mathbb{T}^2$ in vorticity form, under standard stochastic or deterministic forcing and viscosity parameters chosen so that the dynamics are mixing at the resolved scales. We treat the stationary snapshot distribution as μ on H_c and train $F_\theta(\cdot, t)$ on a moderate truncation level H_{c,h_0} (e.g. Fourier cutoff K_0) using the same constrained noise model. The goal is not to reproduce trajectories but to reproduce equilibrium statistics under the induced sampler. We therefore compare empirical means and variances of physically relevant quantities such as kinetic energy $\|u\|_{L^2}^2$, enstrophy $\|\nabla \times u\|_{L^2}^2$, and the spectrum $E(k)$ aggregated over shells. We also compare spatial correlation functions and verify that sampled fields exhibit the expected regularity at the learned noise scales.

A key feature in this experiment is *zero-shot super-resolution.* After training at resolution h_0, we sample on a finer Galerkin space H_{c,h_1} (higher cutoff $K_1 > K_0$) by (i) representing the state on H_{c,h_1}, (ii) drawing $\xi_n \sim \mathcal{N}(0, C_{c,h_1})$, and (iii) evaluating the same operator F_θ in a discretization-consistent manner (e.g. via a Fourier neural operator whose parameterization is independent of grid size). The projected dynamics remain on H_{c,h_1} and preserve incompressibility exactly at the finer level. Empirically, we observe that the low-frequency statistics of samples generated at h_1 match the reference distribution at least as well as those generated at h_0, while the additional high-frequency degrees of freedom are populated in a manner consistent with the constrained Gaussian reference at the smallest annealing scales. This behavior is consistent with the error decomposition implicit in the resolution-invariance theorem: the dominant discrepancies are attributable to score approximation and finite-time annealing, rather than to constraint leakage or resolution-induced instability.

(iii) Conditional sampling with sparse observations: posterior summaries and calibration. Finally, we consider conditional generation from partial observations. Let $B : H_c \to \mathbb{R}^m$ be a bounded observation operator (e.g. pointwise velocity probes, sparse Fourier coefficients, or local averages), and let $y = B(u) + \zeta$ with $\zeta \sim \mathcal{N}(0, \Gamma)$ independent. The target is the posterior $\mu(\cdot \mid y)$ on H_c. We implement conditioning by training a conditional drift $F_\theta(u, t; y)$ (realized by concatenating y-dependent features to the neural operator input), while maintaining feasibility by projecting the output: $F_\theta := \Pi \circ \widetilde{F}_\theta(\cdot, \cdot; y)$. Sampling uses the same projected annealed Langevin iterations, now with the conditional drift. Because the constraint is linear and enforced by projection at each step, all conditional samples remain feasible regardless of the observation pattern.

We evaluate conditional performance through posterior mean and variance fields. Given repeated runs of the conditional sampler, we estimate $\widehat{m}_y(x) = \mathbb{E}[\tilde{u}(x) \mid y]$ and $\widehat{s}_y^2(x) = \mathrm{Var}(\tilde{u}(x) \mid y)$ and compare these against Monte Carlo references computed from a baseline method when available (or against withheld ground truth in synthetic tests). We also assess calibration: for linear functionals $\ell \in H_c^*$ (e.g. average flow through a cross-section), we check whether credible intervals based on $\ell(\tilde{u})$ attain nominal coverage over repeated draws of (u, y). Across observation sparsity regimes we find that the conditional sampler appropriately contracts uncertainty near observed locations while retaining physically plausible variability elsewhere, without introducing spurious divergence. These results support the practical claim that enforcing H_c structurally is compatible with both unconditional and conditional generation, and that discretization changes at inference time do not require retraining provided the operator evaluation and noise sampling are consistent.

Collectively, these experiments validate the two structural points emphasized by our analysis: exact feasibility is maintained by construction during both training and sampling, and the learned sampler can be deployed across discretizations with stable behavior governed primarily by score approximation and annealing error rather than by resolution-dependent constraint enforcement.

Discussion: boundary and inequality constraints, and directions beyond linear subspaces. Our presentation has emphasized linear homogeneous constraints encoded by a closed subspace $H_c \subset H$ and enforced by an orthogonal projector Π. Many PDE constraints of interest, however, are affine rather than homogeneous because of boundary conditions or inhomogeneous conservation laws. A simple and robust extension is to reduce affine constraints to the homogeneous setting by *lifting*. Concretely, suppose the admissible set is

$$\mathcal{A} \;=\; \{u \in H :\; Lu = b\},$$

for a bounded linear operator $L : H \to Y$ and given $b \in Y$ (e.g. $u|_{\partial D} = g$ or prescribed flux). Fix any $u_{\mathrm{lift}} \in H$ such that $Lu_{\mathrm{lift}} = b$ and define the homogeneous subspace $H_0 := \ker L$. Then every $u \in \mathcal{A}$ decomposes uniquely as $u = u_{\mathrm{lift}} + w$ with $w \in H_0$, and learning/sampling may be performed on w instead of u. In this representation, the constraint-preserving projector becomes the affine map

$$\Pi^{\mathrm{aff}}(u) \;:=\; u_{\mathrm{lift}} + \Pi_0(u - u_{\mathrm{lift}}),$$

where $\Pi_0 : H \to H_0$ is the orthogonal projector onto H_0. If the original constraint also includes a subspace condition (e.g. incompressibility), one may set $H_c := H_0 \cap H_{\mathrm{div}}$ and take Π as the orthogonal projection onto H_c; the affine update remains $u_{\mathrm{lift}} + \Pi(\cdot - u_{\mathrm{lift}})$. Since all iterates are obtained by applying Π^{aff} to a feasible initial point plus feasible increments (drift and noise), exact feasibility is again invariant under time stepping. At the discrete level, the same idea applies with a discrete lift $u_{\mathrm{lift},h}$ satisfying the discrete boundary operator and a commuting or compatible discrete projector $\Pi_{0,h}$ (e.g. Helmholtz–Hodge projections with boundary conditions, or $H(\mathrm{div})$-conforming mixed methods). The key point is that boundary constraints should be enforced by construction at the same level as the stochastic dynamics, rather than as an additional penalty that must be tuned with resolution.

A second practically important extension concerns *positivity and bound constraints* (densities, concentrations, viscosities), which are not linear and hence do not define a subspace of H. Here a standard remedy is *reparameterization*: we choose an unconstrained latent field $f \in \widetilde{H}$ and a pointwise map $T : \widetilde{H} \to H$ such that $u = T(f)$ satisfies the constraint automatically (for example $T(f) = \exp(f)$ for positivity, $T(f) = \mathrm{softplus}(f)$ for nonnegativity with better conditioning, or $T(f) = a + (b-a)\sigma(f)$ for $u \in [a, b]$). We then train and sample in f–space using our linear constraint machinery (and any additional linear constraints may be imposed on f or on u depending on the application), and finally push forward samples by T. This procedure yields exact positivity at every iteration, regardless of discretization. One must, however, be clear about the target: if μ is a distribution on u, then the corresponding pullback distribution on f is $\mu^f := \mu \circ T$, and the DDO/DSM objective identifies the score of μ^f with respect to the chosen reference on $\widetilde{H}$. When T is smooth and invertible (e.g. exponential), the relationship between scores is given formally by the chain rule with a Jacobian correction:

$$D \log \frac{d\mu^f}{d\mu_0}(f) \;=\; (DT(f))^* \, D \log \frac{d\mu}{d(T_{\#}\mu_0)}(T(f)) \;+\; D \log |\det DT(f)|,$$

where $(DT(f))^*$ is the adjoint derivative. In practice we avoid explicit determinants by defining the model and corruption directly in f–space, training there, and accepting that the induced distribution on $u = T(f)$ is the push-forward of the learned latent sampler. This viewpoint keeps the sampler constraint-preserving (positivity by construction) while retaining resolution invariance provided T is evaluated pointwise and the latent operator network is discretization-consistent.

These two extensions highlight a broader limitation of our current theory: our equivalence and invariance results rest on the linear-subspace geometry of H_c and on Gaussian reference measures supported on H_c. When the constraint set is a *nonlinear manifold* $\mathcal{M} \subset H$ (e.g. unit-length fields, orthonormal frames, determinant constraints, contact constraints, or level-set constraints), there is in general no globally defined bounded projector $\Pi : H \to \mathcal{M}$ with the properties we exploited (idempotence, self-adjointness, and stability under Galerkin refinement). Local projections exist under regularity and reach assumptions, but they are nonlinear and can introduce bias or instability under discretization, and the analogues of Cameron–Martin structure and Gaussian reference measures become substantially more delicate. Moreover, even when one performs a projection step onto $\mathcal{M}$ after each update, it is not immediate that the resulting Markov chain targets a measure with a tractable density or that a DSM identity holds with respect to a convenient reference.

Several concrete directions follow. First, it is natural to develop a *manifold-constrained* version of the present framework in which the state evolves on $\mathcal{M}$ via a Riemannian Langevin diffusion driven by a tangent noise and a drift defined by the intrinsic gradient of an intrinsic potential. One would then seek a score defined relative to a reference measure on $\mathcal{M}$ (e.g. the Riemannian volume or a specified base distribution) and prove an analogue of the denoising identity using intrinsic integration by parts. The algorithmic counterpart would replace the linear projection Π by a retraction and tangent-space projection at each step, preserving the constraint up to retraction accuracy while aiming for a correct invariant measure.

Second, while we have treated Π as known (coming from the PDE constraint), it is appealing to consider *learned projectors* or learned fast approximations of Π when an exact Helmholtz or boundary projection is expensive. Any such approximation must be handled with care: small violations of idempotence or failure to commute with refinement can reintroduce resolution-dependent constraint leakage. A principled goal is therefore to learn operators that remain uniformly bounded on H and converge strongly to the true Π as resolution increases, possibly by parameterizing Π through a constrained architecture (e.g. $\Pi = QQ^*$ with Q a partial isometry) or by learning a constraint solver that is consistent with the continuous operator.

Third, we expect that *non-Gaussian constrained references* will be important in applications where Gaussian small-scale structure is inadequate (e.g. sparsity, intermittency, or heavy tails). Replacing μ_0^c by Besov-type priors, Lévy-driven fields, or Gibbs measures changes the notion of score and the form of the denoising identity; the appropriate objects are then logarithmic derivatives and Stein operators associated with the reference measure rather than Cameron–Martin gradients. Establishing measure equivalence, finiteness of DSM objectives, and resolution-invariant sampling in these settings appears feasible but requires different functional-analytic tools than those used for trace-class Gaussians.

In summary, the core message persists under these extensions: whenever constraints can be encoded by an invariant structure of the dynamics (affine lifting, reparameterization, or intrinsic geometry), feasibility should be enforced by construction rather than by penalties. The main open challenge is to preserve the same level of discretization robustness when moving from linear subspaces with Gaussian references to genuinely nonlinear constraint sets and non-Gaussian base measures.