Improving Function Space Flow Matching with Kernel Optimal Transport

Fred Xu
University of California, Los Angeles
fredxu@cs.ucla.edu

Thomas Markovich
Block, Inc.
tmarkovich@squareup.com

Barbora Barancikova
Imperial College London
b.barancikova23@imperial.ac.uk

Yizhou Sun
University of California, Los Angeles
yzsun@cs.ucla.edu

Abstract

Generative models for function-valued data, such as time series and solutions of partial differential equations, must learn distributions over infinite-dimensional spaces. Functional Flow Matching (FFM) extends Flow Matching to this setting, learning a velocity field whose flow transports a Gaussian prior to the data distribution, but it inherits the independent endpoint pairing of standard Flow Matching: in each batch, prior and data samples are matched arbitrarily, so the conditional bridge must traverse both the shared global structure of the dataset and instance-specific residuals. In function space this limitation is harder to fix than in finite dimensions, since optimal transport (OT) on function spaces is delicate to formulate and a flat Euclidean surrogate ignores the geometry that distinguishes function-valued data. We propose kernel Functional Flow Matching (kFFM), which replaces the independent pairing by entropic OT under a kernel-induced cost, the coupling underlying the Hilbert Sinkhorn Divergence (HSD), while leaving the FFM neural-operator architecture unchanged. Theoretically, we prove that the kernel cost and the HSD objective are uniformly bounded and well-posed on Banach ambient spaces, derive an explicit error decomposition against quadratic-cost OT on compact metric spaces that isolates an irreducible kernel-cost mismatch term, and prove a discretization-invariance bound whose rate is governed by Sobolev regularity. Empirically, kFFM improves distributional matching over FFM, diffusion, adversarial, and finite-dimensional OT baselines on time-series and PDE benchmarks, with statistically significant paired-seed gains over FFM and improvements that persist under non-kernel and physics-based diagnostics, including a turbulent Navier–Stokes benchmark. Bounded kernel costs already outperform raw L^2 Sinkhorn, and function-space-aware kernels (signature, Sobolev RBF) give further gains on rough or path-valued data.

1 Introduction

Flow matching [35] is a powerful framework for generative modeling that learns the velocity field of a probability path between a reference and a data distribution. Time series and PDE snapshots are naturally represented as functions, motivating Kerrigan et al. [25] to extend this framework to function space through Functional Flow Matching (FFM), which defines a path of Gaussian measures in a Hilbert space and parameterizes the velocity field with a neural operator. FFM enjoys resolution-invariance and competitive performance on sequence and PDE datasets, but inherits the independent endpoint coupling of base FM. Under this coupling, each Gaussian prior sample is

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

paired with an arbitrary data function in a minibatch, and the conditional bridge must traverse both the shared global structure of the dataset and instance-specific residuals; the velocity field is then forced to discover the global structure and match each particular instance simultaneously. A natural fix is geometry-aware coupling: pairing each prior sample with a data sample that is already close in a meaningful geometry, so the bridge primarily transports the residual. This idea has improved finite-dimensional flow matching [49], but transferring it to function space is non-trivial: direct optimal transport (OT) on function spaces is delicate to formulate and compute [8, 9, 22, 38], and a flat L^2 surrogate ignores the function-space geometry (Sobolev structure, path roughness) that distinguishes one function-valued sample from another.

We resolve this through a kernel-induced Hilbert space surrogate. Working in a Reproducing Kernel Hilbert Space [47], the entropic coupling underlying the Hilbert Sinkhorn Divergence [31] provides a tractable pairing rule whose kernel encodes prior knowledge about function-space geometry: Sobolev structure for PDE snapshots via the RBF kernel on H^k, and p-variation roughness for time series via the signature kernel [5, 6]. Our theory certifies this surrogate coupling (well-posedness with uniform bounds, an error decomposition with an irreducible kernel-mismatch term, and discretization invariance); it makes no straightness or sampling-efficiency claim in function space (Section 4, Appendix C.14), and the sample-quality gains are an empirical claim that we test with paired-seed statistics, non-kernel metrics, physics diagnostics, and a turbulent Navier–Stokes stress test. The kernel-induced surrogate also outperforms raw L^2 Sinkhorn: relative to a flat-L^2 minibatch-OT baseline (CFM-OT(L^2)) with the same Sinkhorn solver and neural-operator backbone, kFFM lowers the distributional error by 75% on Gene Expression and 66% on Economy. Function-space-aware kernels (Sobolev RBF, signature) add further gains on rough or path-valued data. Code for every experiment is publicly available.[1] Our main contributions are the following:

1. We propose kernel Functional Flow Matching (kFFM), to our knowledge the first method to perform OT endpoint coupling in function-space flow matching: a kernel-induced cost couples prior and data samples before flow matching, while the FFM neural-operator architecture is unchanged.
2. We establish three theoretical results for the function-space setting: uniform boundedness and well-posedness of the kernel cost and HSD on Banach ambient spaces; an explicit error decomposition against quadratic-cost OT on compact metric spaces (kernel mismatch, entropic bias, covering); and a discretization-invariance theorem with an $N^{-2(\alpha-k)/d}$ rate that ties the function-space objective to the truncated grid representation the model actually computes.
3. We show empirically that kFFM improves population-level distributional quality on time series and PDE snapshots over functional generative baselines and a finite-dimensional OT coupling baseline, with paired-seed significance, gains that survive non-kernel and physics-based metrics including a turbulent Navier–Stokes benchmark, and a training overhead of only 5–7% at the largest resolution we train. The coupling geometry matters differently across data classes: signature kernels for paths, RBF-type costs for fields.

2 Background

For the FFM construction we follow Kerrigan et al. [25] and work in a separable Hilbert space $\mathcal{F}$ of functions $f : \Omega \to \mathbb{R}$, with $\Omega \subset \mathbb{R}^d$ compact, equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{F})$; $\nu \in \mathcal{P}(\mathcal{F})$ denotes the data measure. Section 2.1 works directly with $\mathcal{F}$, while later subsections and Section 3 use $\mathcal{X}$ for the metric sample space on which the OT coupling is defined. Additional background is deferred to Appendix A.

2.1 Flow Matching in Function Space

Flow Matching [35] learns the velocity field whose flow transports a prior distribution to the data distribution along a given probability path. In the Gaussian-measure formulation of FFM, a vector field $v : [0, 1] \times \mathcal{F} \to \mathcal{F}$ induces the flow $\partial_t \psi_t(g) = v_t(\psi_t(g))$, $\psi_0(g) = g$, and the path $\mu_t = [\psi_t]_{\#}\mu_0$ from $\mu_0 \in \mathcal{P}(\mathcal{F})$. FFM constructs this path by averaging endpoint-conditioned Gaussian bridges over $f \sim \nu$: if $\mu_0^f = \mu_0$ and μ_1^f is concentrated around f, then $\mu_t(A) = \int \mu_t^f(A)\, d\nu(f)$ satisfies

[1] https://github.com/heraclixus/KFFM

$\mu_1 = \nu$. For Gaussian bridges with $\mu_0 = \mathcal{N}(0, C_0)$ and $\mu_t^f = \mathcal{N}(m_t^f, (\sigma_t^f)^2 C_0)$, the conditional velocity is

$$v_t^f(g) = \frac{(\sigma_t^f)'}{\sigma_t^f}(g - m_t^f) + \frac{d}{dt} m_t^f, \qquad g \sim \mu_t^f,$$

and the standard FFM parametrization $m_t^f = tf$, $\sigma_t^f = 1 - (1 - \sigma_{\min})t$ corresponds to Gaussian OT between the bridge marginals, while the endpoint pairing itself remains independent [49]. A neural operator $v_\theta(g, t)$, here a Fourier Neural Operator (FNO) [32], is trained with the Conditional Flow Matching (CFM) loss

$$\mathcal{L}_{CFM}(\theta) := \mathbb{E}_{t\sim U[0,1], f\sim\nu, g\sim\mu_t^f}\left[\|v_t^f(g) - v_\theta(g,t)\|_{\mathcal{F}}^2\right], \tag{1}$$

which under suitable regularity conditions [25, 35] has the same minimizers as the unconditional loss $\mathcal{L}_{FM}(\theta) := \mathbb{E}_{t\sim U[0,1], g\sim\mu_t}[\|v_t(g) - v_\theta(g,t)\|_{\mathcal{F}}^2]$; the absolute-continuity assumption behind this equivalence is spelled out in Appendix B.2. In Section 3 we keep the same probability path but replace the independent pairing of its endpoints by an OT-coupled pairing.

2.2 Optimal Transport and Kernel-based Optimal Transport

For $(\mathcal{X}, d)$ a Polish space, $\mu, \nu \in \mathcal{P}(\mathcal{X})$, and a cost c, the static OT problem is $\mathcal{W}(\mu, \nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int c\, d\pi$ [50], and its entropic regularization [39]

$$W_\epsilon(\mu, \nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int c(x, y)\, d\pi(x, y) + \epsilon\, \mathrm{KL}(\pi \,\|\, \mu \otimes \nu) \tag{2}$$

underlies Sinkhorn methods; the debiased Sinkhorn divergence subtracts the corresponding self-costs [16, 17, 48]. OT-based couplings improve finite-dimensional flow matching by computing a static coupling between reference and data samples before interpolation [28, 49]. Adapting this idea to function-valued data is non-trivial: no Lebesgue-like reference measure exists on function space [11, 38], so densities are ill-posed, and Gaussian-measure and Cameron–Martin assumptions [8] are limiting when data are non-Gaussian (PDE states, rough paths). This motivates embedding the measures into a Reproducing Kernel Hilbert Space (RKHS) [47], where the coupling is computed under a kernel-induced cost.

Definition 1 (Hilbert Sinkhorn Divergence [31])**.** Let $(\mathcal{X}, d)$ be a metric space, $\mu, \nu \in \mathcal{P}(\mathcal{X})$, and $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a measurable symmetric positive-definite kernel with RKHS $\mathcal{H}_\kappa$ and canonical feature map $\phi : \mathcal{X} \to \mathcal{H}_\kappa$. The entropic OT functional on the embedded space is

$$\mathsf{OT}_\epsilon^{\mathcal{H}_\kappa}(\phi_\#\mu, \phi_\#\nu) := \inf_{\pi_\phi\in\Pi(\phi_\#\mu,\phi_\#\nu)} \Big[\int \|u - v\|_{\mathcal{H}_\kappa}^2\, d\pi_\phi(u, v) + \epsilon\, \mathrm{KL}(\pi_\phi \,\|\, \phi_\#\mu \otimes \phi_\#\nu)\Big], \tag{3}$$

and the Hilbert Sinkhorn divergence is the debiased quantity $S_\epsilon(\phi_\#\mu, \phi_\#\nu) := \mathsf{OT}_\epsilon^{\mathcal{H}_\kappa}(\phi_\#\mu, \phi_\#\nu) - \frac{1}{2}\mathsf{OT}_\epsilon^{\mathcal{H}_\kappa}(\phi_\#\mu, \phi_\#\mu) - \frac{1}{2}\mathsf{OT}_\epsilon^{\mathcal{H}_\kappa}(\phi_\#\nu, \phi_\#\nu)$.

By the kernel trick, the functional in Equation (3) can be written directly on $\mathcal{X}$ with the RKHS-induced cost $c_\kappa(x, y) := \|\phi(x) - \phi(y)\|_{\mathcal{H}_\kappa}^2 = \kappa(x, x) + \kappa(y, y) - 2\kappa(x, y)$ [31, restated as Proposition 3 in Appendix B]:

$$\mathsf{OT}_{\kappa,\epsilon}(\mu, \nu) := \inf_{\pi\in\Pi(\mu,\nu)} \Big[\int_{\mathcal{X}\times\mathcal{X}} c_\kappa(x, y)\, d\pi(x, y) + \epsilon\, \mathrm{KL}(\pi \,\|\, \mu \otimes \nu)\Big], \tag{4}$$

with minimizers corresponding under pushforward, and $S_{\kappa,\epsilon}(\mu, \nu) := \mathsf{OT}_{\kappa,\epsilon}(\mu, \nu) - \frac{1}{2}\mathsf{OT}_{\kappa,\epsilon}(\mu, \mu) - \frac{1}{2}\mathsf{OT}_{\kappa,\epsilon}(\nu, \nu) = S_\epsilon(\phi_\#\mu, \phi_\#\nu)$. Previous work takes $\mathcal{X} \subset \mathbb{R}^d$, where common kernels apply directly, the divergence enjoys statistical consistency with finite-sample bounds (restated in Appendix B), and the discrete entropic OT problem is convex [16]. Section 3 generalizes $\mathcal{X}$ to function-valued sample spaces.

2.3 Path Signature and Signature Kernel

Let $C_p(\mathbb{R}^d) = C_p([0, T], \mathbb{R}^d)$ denote the continuous paths from $[0, T]$ to $\mathbb{R}^d$ with finite p-variation, $p \in [1, 2)$. The signature of $x \in C_p(\mathbb{R}^d)$ is the sequence of iterated Young integrals $S(x) = (1, S^{(1)}(x), S^{(2)}(x), \ldots)$ with $S^{(m)}(x) = \int_{0<t_1<\cdots<t_m<T} dx_{t_1} \otimes \cdots \otimes dx_{t_m} \in (\mathbb{R}^d)^{\otimes m}$, an element of the free tensor algebra $T((\mathbb{R}^d)) = \prod_{n=0}^\infty (\mathbb{R}^d)^{\otimes n}$.

Definition 2 (Signature Kernel [5])**.** The signature kernel $\kappa_{\mathrm{sig}} : C_p(\mathbb{R}^d) \times C_p(\mathbb{R}^d) \to \mathbb{R}$ is $\kappa_{\mathrm{sig}}(x, y) = \langle S(x), S(y) \rangle := \sum_{m=0}^{\infty} \langle S^{(m)}(x), S^{(m)}(y) \rangle_{(\mathbb{R}^d)^{\otimes m}}$, where $\langle \cdot, \cdot \rangle_{(\mathbb{R}^d)^{\otimes m}}$ is the Hilbert–Schmidt inner product.

The signature kernel is symmetric and positive definite and hence induces a unique RKHS $\mathcal{H}_{\mathrm{sig}}$; on compact subsets, the time-augmented signature kernel is universal and characteristic [5]. We compute it with `pySigLib` [45].

3 Functional Flow Matching with Kernel Optimal Transport

Throughout this section $\mathcal{X}$ is a general metric space, the sample space on which data and reference measures live and on which the kernel OT coupling acts.

3.1 Coupling with Kernel Optimal Transport

kFFM first couples prior and data samples through entropic OT under a kernel-induced cost, Equation (4). The kernel reflects the metric space where the data live. On the Sobolev space $\mathcal{X} = H^k(\Omega)$ (functions with k weak derivatives in L^2, the natural home of PDE solutions [12]), the RBF kernel on H^k rewards pairs close in a derivative-aware norm; on path space $\mathcal{X} \subset C_p([0, T], \mathbb{R}^d)$, $p \in [1, 2)$ (smooth and rough time series [5]), the signature kernel rewards reparametrization-invariant path similarity. Well-posedness for each choice is established in Section 3.2 and validated in Section 4; with empirical measures the minibatch problem is solved by the Sinkhorn algorithm [17].

Given a plan π between prior draws $f_0 \sim \mu_0$ and data samples $f_1 \sim \mu_1$, kFFM uses the same probability path as FFM, with the paired prior sample in the role of the base noise:

$$g_t^{f_0,f_1} = t f_1 + \sigma_t f_0, \qquad \sigma_t = 1 - (1 - \sigma_{\min})t, \qquad (f_0, f_1) \sim \pi,$$

$$v_t^{f_0,f_1}\big(g_t^{f_0,f_1}\big) = \partial_t g_t^{f_0,f_1} = f_1 - (1 - \sigma_{\min}) f_0 = \frac{f_1 - (1 - \sigma_{\min})\, g_t^{f_0,f_1}}{\sigma_t},$$

so that $g_0 = f_0 \sim \mu_0$, $g_1 = f_1 + \sigma_{\min} f_0$, and the conditional velocity is the FFM conditional field of Section 2.1 evaluated with the coupled pair; the FFM path $tf + \sigma_t \xi$ is recovered exactly for the independent coupling $\pi = \mu_0 \otimes \mu_1$.

What the coupling changes. Up to the $\sigma_{\min}$ term, the regression target is the displacement $u := f_1 - f_0$, the *residual* between the paired data and prior samples, which carries all the endpoint information. Under independent coupling this residual is the difference of two unrelated draws, a large and high-variance regression target; under kernel-OT coupling each f_1 is paired with a geometrically close f_0, so the network regresses smaller, structured residuals. The neural operator (an FNO) is unchanged relative to FFM; only the minibatch coupling, and through it the law of the conditional paths, is modified. Algorithm 1 summarizes training; implementation defaults and good practices are collected in Appendix C.8. The reference distribution μ_0 is a Gaussian process (GP) or Gaussian random field [19], or white noise on the grid where the data are rougher than a smooth GP prior (the base measure is part of the selected configuration, Section 4.1); in practice both measures are replaced by their empirical estimators, with the coupling computed from a kernel-induced cost matrix [17, 31]:

$$\mu_n = \sum_{i=1}^{n} \hat{\mu}_i \delta_{x_i}, \qquad \nu_n = \sum_{j=1}^{n} \hat{\nu}_j \delta_{y_j}, \qquad C_{ij}^{\kappa} = \kappa(x_i, x_i) + \kappa(y_j, y_j) - 2\kappa(x_i, y_j). \tag{5}$$

Where the kernel enters. The kernel enters at exactly one place: the cost c_κ used by minibatch Sinkhorn to *pair* prior and data samples. No sample or model input is projected into the RKHS; the FNO acts on the original discretized functions, the paths are constructed in the original space, and generation happens there. The training coupling is the entropic plan induced by c_κ, computed per minibatch with log-domain Sinkhorn on a median-normalized cost and sampled from the plan (Appendix C.8); the debiasing self-terms of the HSD (Definition 1) do not depend on π and cannot change the argmin plan, so HSD is our *analysis* object (what Theorems 1–2 control), not the training objective. Finally, at a fixed batch size b minibatch OT induces the expected minibatch plan $\bar{\pi}_b = \mathbb{E}[\hat{\pi}_b]$ [13, 14], a valid coupling of the true marginals for every b, so the CFM consistency argument applies and b only changes which conditional paths are used; as $b \to \infty$, $\bar{\pi}_b$ converges to the population entropic kernel-OT plan.

Algorithm 1 Functional Flow Matching with Kernel Optimal Transport

Require: Prior distribution $\mu_0 = \mathcal{N}(0, C_0)$, data distribution μ_1, kernel function κ, entropy regularization $\epsilon > 0$, batch size b, initial network v_θ, $\sigma_{\min} > 0$.

Training loop (repeat until convergence):

Sample batches of size b: $f_0 \sim \mu_0$, $f_1 \sim \mu_1$

Solve the entropic kernel OT problem in Equation (4) with κ, ϵ (cost matrix in Equation (5)) to get coupling π.

$(f_0, f_1) \sim \pi$

$t \sim \mathcal{U}(0,1), \quad \sigma_t \leftarrow 1 - (1 - \sigma_{\min})t$

$g \leftarrow t f_1 + \sigma_t f_0$ *// the paired prior sample plays the role of the base noise*

$v_t^{f_0,f_1}(g) \leftarrow \big(f_1 - (1 - \sigma_{\min})\, g\big)/\sigma_t \;=\; f_1 - (1 - \sigma_{\min}) f_0$

$\mathcal{L}_{CFM}(\theta) \leftarrow \left\| v_t^{f_0,f_1}(g) - v_\theta(g,t) \right\|_{\mathcal{F}}^2$

$\theta \leftarrow \mathrm{Update}(\theta, \nabla_\theta \mathcal{L}_{\mathrm{CFM}}(\theta))$

Output: v_θ.

3.2 Theoretical Results for Sobolev and Path Space

The theory certifies the surrogate coupling; it does not by itself explain why sample quality improves, which Section 4 establishes empirically. It answers three questions (proofs in Appendix B): is the kernel-OT objective well-defined and well-conditioned on the function spaces used by kFFM; how far is the RKHS-induced cost from quadratic-cost OT, and which parts of the gap are reducible; and does the continuum formulation survive the Fourier truncation used by the FNO and the Sinkhorn solver?

Theorem 1 (Well-posedness and uniform boundedness)**.** *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a real Banach space and let $d_{\mathcal{X}}(f,g) := \|f - g\|_{\mathcal{X}}$. Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be measurable, symmetric, and positive definite with RKHS $\mathcal{H}_\kappa$. Assume additionally that κ is bounded: $B_\kappa := \sup_{f \in \mathcal{X}} \kappa(f,f) < \infty$. Then the kernel cost satisfies $0 \le c_\kappa(f,g) \le 4B_\kappa$ for all $f, g \in \mathcal{X}$, and for every $\varepsilon > 0$ the Hilbert Sinkhorn divergence $S_{\kappa,\varepsilon}(\mu,\nu)$ (Definition 1) is finite for all $\mu, \nu \in \mathcal{P}(\mathcal{X})$, with $|S_{\kappa,\varepsilon}(\mu,\nu)| \le 8B_\kappa$.*

Finiteness per se is not the binding constraint: under a second-moment assumption on ν the L^2 cost is already integrable, and for the Gaussian prior $\mathbb{E}\|f_0\|^2 = \mathrm{tr}(C_0) < \infty$. Theorem 1 delivers more. (a) *Uniform* boundedness with no assumption on the measures, which conditions the entropic solver: the Gibbs kernel $e^{-c_\kappa/\varepsilon}$ stays bounded away from 0 and 1 across batches, so Sinkhorn converges independently of outlier pairs, unlike unbounded L^2 costs on heavy-tailed batches (Section 4.3 quantifies the resulting speed-up). (b) Generality: the statement holds on Banach spaces without Hilbert structure, which licenses the signature kernel on path space, our strongest configuration. (c) Geometry injection: the kernel is the interface through which domain structure enters the coupling. For Sobolev data in $H^k(\Omega)$ we use $\kappa^{(k)}_{\mathrm{RBF}}(x,y) := \exp(-\|x - y\|^2_{H^k}/2\sigma^2)$, which is bounded and positive definite [7, 55]; for paths we use the signature kernel (Definition 2), universal and characteristic on compact sets [5], bounded there by continuity, and used for two-sample tests on path data [6], applied to time-augmented paths (Appendix A, Definition 14) to avoid its tree-like equivalence.

The entropic OT functional transports with the cost induced by the feature map rather than the original metric cost. On compact subsets of the ambient space ($B_R^{H^\alpha}$ is compact in H^k for $\alpha > k$; compact sets of paths for the signature kernel, with truncation discussed in Appendix B), the next theorem separates the resulting kernel-cost mismatch from entropic and covering effects.

Theorem 2 (Error decomposition)**.** *Let $(\mathcal{X}, d)$ be a compact metric space and let $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Define the target quadratic cost $c(x,y) = d(x,y)^2$. Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded positive-definite kernel with RKHS-induced cost $c_\kappa(x,y) := \kappa(x,x) + \kappa(y,y) - 2\kappa(x,y)$, and let $\Delta_\kappa := \sup_{x,y \in \mathcal{X}} |c_\kappa(x,y) - c(x,y)|$. Let $\mathsf{OT}_{\kappa,\varepsilon}$ denote the entropic OT functional in Equation* (4) *with cost c_κ, and let $M := \mathrm{diam}(\mathcal{X}) < \infty$. Then for every $\varepsilon > 0$ and every $\delta \in (0, M]$:*

$$|\mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) - \mathcal{W}(\mu,\nu)| \le \Delta_\kappa + 2\varepsilon \log \mathcal{N}(\mathcal{X}, \delta; d) + 8M\delta, \tag{6}$$

where $\mathcal{N}(\mathcal{X}, \delta; d)$ is the δ-covering number of $\mathcal{X}$ in the metric d.

Theorem 2 is an error decomposition, not an approximation guarantee: the entropic and covering terms vanish as $\varepsilon, \delta \to 0$, but Δ_κ is *irreducible* and does not shrink with more samples or finer grids. For the RBF kernel, $c_{\mathrm{RBF}} = 2(1 - e^{-\|f-g\|^2/2\sigma^2})$ is a rescaled L^2 cost up to relative error $O((D/\sigma)^2)$, with D the batch diameter; with median-normalized costs Δ_κ is small at large bandwidth, and on smooth Navier–Stokes fields the induced plans agree across costs (total variation ≤ 0.02; Appendix C.12). For the signature kernel Δ_κ is large by design: the theorem then controls the deviation from Wasserstein with the *chosen* cost, not proximity to L^2 OT, and the pair rankings differ measurably from L^2 (rank correlation 0.33–0.37 on AEMET and gene expression), which is where the largest gains occur. In low-regularity regimes the bound degrades on two fronts: Δ_κ grows because a smooth kernel cost saturates on rough functions, and the covering number blows up as the effective smoothness approaches the embedding threshold, making the entropic term large for any practical ε. The benchmarks of Section 4.1 sit in the smooth regime where the bound is informative; Section 4.2 stress-tests the turbulent regime empirically.

Theorems 1 and 2 are stated at the continuum level, while the model operates on a finite-rank spectral representation: PDE snapshots are represented in a Fourier subspace of dimension proportional to N, which both the FNO and the Sinkhorn solver consume. Our final theorem quantifies the resulting gap at a rate governed by Sobolev regularity; the path-space analogue for signature truncation is discussed in Appendix B.

Theorem 3 (Discretization invariance)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded periodic domain and let $\alpha > k \geq 0$. Set $\mathcal{X} := B_R^{H^\alpha} := \{f \in H^\alpha(\Omega) : \|f\|_{H^\alpha} \leq R\}$, viewed as a compact subset of $H^k(\Omega)$, and let $\mu, \nu \in \mathcal{P}(\mathcal{X})$. For the Fourier basis $\{e_n\}_{n\in\mathbb{Z}^d}$, let $V_N := \operatorname{span}\{e_n : |n|^2 \leq N^{2/d}\}$ and let $P_N : H^\alpha(\Omega) \to V_N$ denote the orthogonal projection; hence $\dim V_N \asymp N$. Write $\mu_N := (P_N)_\# \mu$ and $\nu_N := (P_N)_\# \nu$. Define the Sobolev-RBF kernel $\kappa^{(k)}_{\mathrm{RBF}}(f, g) := \exp(-\|f - g\|^2_{H^k}/2\sigma^2)$ on H^k, and write $\mathsf{KOT}^{(k)}_\varepsilon := \mathsf{OT}_{\kappa^{(k)}_{\mathrm{RBF}},\varepsilon}$ for the corresponding entropic kernel-OT functional from Theorem 2. Then for every $\varepsilon > 0$,*

$$\left|\mathsf{KOT}^{(k)}_\varepsilon(\mu, \nu) - \mathsf{KOT}^{(k)}_\varepsilon(\mu_N, \nu_N)\right| \;\leq\; \frac{4R^2}{\sigma^2}\, N^{-2(\alpha-k)/d}.$$

Equivalently, in terms of grid spacing $h = N^{-1/d}$, the discretization error decays as $O(h^{2(\alpha-k)})$.

Combining Theorems 2 and 3 bounds the gap between the computed finite-rank kernel-OT objective and continuum quadratic OT in the H^k metric by four interpretable terms: kernel-cost mismatch, entropic regularization, covering, and discretization (Corollary 1 in Appendix B.6). Only the discretization term vanishes with projection rank N, at rate $N^{-2(\alpha-k)/d}$ driven by the regularity gap $\alpha - k$, the formal counterpart of the empirical resolution-invariance in Section 4; the mismatch term depends only on kernel design (zero for the linear kernel, finite for RBF and controlled by data scale and bandwidth σ). The theorem relies on compact Sobolev embedding [12]; our smooth PDE benchmarks sit in that regime [4].

4 Experiments

Datasets. We use two classes of data. *Sequence datasets*: AEMET [15], gene expression [37], and the economics dataset [3] from Kerrigan et al. [25], plus financial time series from the Heston stochastic volatility model [21], treated as realizations of measures on path space $C_p(\mathbb{R})$. *PDE snapshots*: the 1D Korteweg–de Vries (KdV) equation, the incompressible 2D Navier–Stokes equation [33], and their stochastic counterparts [44], treated as measures on Sobolev space $H^k(\Omega)$; Section 4.2 adds a turbulent 2D Navier–Stokes benchmark.

Baselines. Denoising Diffusion Operator (DDO) with NCSN noise scale [34], functional DDPM [24], Generative Adversarial Neural Operator (GANO) [41], and the original FFM [25]. We also include a direct finite-dimensional OT baseline, CFM-OT(L^2) [49], which uses the same minibatch Sinkhorn solver and FNO backbone as our method but applies entropic OT to flattened grid vectors with the unbounded L^2 cost and the white-noise base of finite-dimensional flow matching; kFFM's Euclidean–RBF configuration instead uses the bounded RBF cost on Euclidean distances. kFFM variants are named by cost, kFFM-Sig (signature), kFFM-Sob (Sobolev-RBF), kFFM-RBF (Euclidean–RBF), kFFM-Euc (raw L^2), and tagged by base measure, gp (GP prior) or wn (white noise). Datasets, architectures, and training budgets, identical across methods, are detailed in Appendix C.

Table 1: Population-level distributional quality measured by MMD-RBF (lower is better; unbiased estimator, so small negative values occur). "kFFM (selected)" is the per-dataset configuration chosen on a held-out validation split by sliced Wasserstein distance (Appendix C.5); each cell is tagged (cost/base) with Sig = signature kernel, RBF = Euclidean–RBF kernel, Sob-RBF = Sobolev-RBF kernel, gp = Gaussian-process prior, wn = white noise. The kFFM and FFM columns are the matched-seed runs of Table 2 (n as given there); the remaining baselines are averaged over 10 seeds. "Gain (%)" is relative to the strongest non-kFFM baseline in the row and is omitted (–) where that baseline sits at the noise floor of the unbiased estimator (non-positive MMD), where ratios are not meaningful; Table 2 reports paired tests instead.

Dataset	kFFM (selected)	CFM-OT(L^2)	FFM	DDPM	DDO/NCSN	GANO	Gain (%)
Heston	**kFFM (RBF/wn)**: -1.30×10^{-4}	-8.26×10^{-5}	3.10×10^{-4}	1.10×10^{-3}	5.53×10^{-2}	3.92×10^{-2}	–
AEMET	**kFFM (Sig/gp)**: -5.10×10^{-3}	-4.11×10^{-3}	-3.24×10^{-3}	-3.99×10^{-3}	2.44×10^{-1}	5.24×10^{-1}	–
KdV	**kFFM (Sob-RBF/gp)**: -2.56×10^{-3}	-1.32×10^{-3}	-1.16×10^{-3}	4.84×10^{-2}	2.37×10^{-1}	4.12×10^{-1}	–
Economy	**kFFM (Sig/gp)**: 2.89×10^{-3}	8.46×10^{-3}	7.30×10^{-3}	9.97×10^{-3}	1.57×10^{-1}	8.98×10^{-1}	**60.4**
Gene Expr.	**kFFM (Sig/gp)**: 7.11×10^{-3}	2.80×10^{-2}	1.78×10^{-2}	5.32×10^{-2}	3.42×10^{-2}	4.17×10^{-1}	**60.1**
Stoch. KdV	**kFFM (Sob-RBF/gp)**: 1.00×10^{-4}	6.58×10^{-4}	8.80×10^{-4}	3.84×10^{-4}	4.48×10^{-2}	2.39×10^{-2}	**74.0**
Stoch. NS	**kFFM (RBF/wn)**: 2.80×10^{-3}	4.22×10^{-3}	9.68×10^{-2}	1.20×10^{-1}	2.95×10^{-1}	2.52×10^{-1}	**33.6**
Navier–Stokes	**kFFM (RBF/wn)**: 1.63×10^{-3}	1.78×10^{-3}	1.31×10^{-1}	3.80×10^{-1}	7.35×10^{-1}	3.00×10^{-1}	**8.4**

Metrics. The maximum mean discrepancy with an RBF kernel (MMD-RBF) is the primary population-level metric, computed with the unbiased estimator (so small negative values occur). To avoid judging an RBF-type coupling by an RBF-MMD alone, we also report non-kernel metrics (sliced Wasserstein distance and marginal W_1), the pointwise diagnostics of Kerrigan et al. [25] (autocorrelation error on sequences, log-spectral distance on PDEs, mean/variance summaries in Appendix C), and, on Navier–Stokes, physics diagnostics that capture non-Gaussian structure: enstrophy-distribution W_1, vorticity-PDF W_1, skewness and excess-kurtosis errors of the pooled vorticity, and log-spectral error. Results aggregate 10 seeds (20 where marked); all methods share seeds, so tests are paired by seed.

4.1 Distributional Quality

Table 1 reports MMD-RBF against all baseline families. The "kFFM (selected)" column reports, per dataset, the configuration (OT cost and Gaussian base measure) selected on a held-out validation split by sliced Wasserstein distance, a non-kernel criterion (Appendix C.5); pointwise diagnostics by kernel variant appear in Appendix C.6. kFFM is the strongest method on every row, and the selected configurations follow a practitioner-usable pattern: the signature kernel with a GP base on path-space datasets, where function-space geometry matters most (except short-horizon Heston, where the top configurations are statistically indistinguishable and the rule picks the bounded RBF cost with a white-noise base); the Sobolev-RBF cost with a GP base on the KdV family; and the bounded Euclidean–RBF cost with a white-noise base on the Navier–Stokes family, whose rough fields are poorly matched by a smooth GP prior.

Uncertainty and paired tests. Since unbiased MMD values can be tiny or negative, percentage gains are uninformative near the noise floor. Table 2 therefore reports per-seed mean$\pm$std and seed-paired tests against FFM at matched architecture and training budget: kFFM improves on FFM on every dataset at level 0.05 under both a paired t-test and a Wilcoxon signed-rank test, and wins on every shared seed for the four largest-gap datasets. In the GP-base rows only the coupling changes; in the white-noise rows (Heston, Navier–Stokes family) the base measure changes too, and at that base the effect of the kernel cost is isolated by the comparison against CFM-OT(L^2) in Table 1.

Non-kernel metrics. The gains survive the non-kernel metrics (Table 3): on the same runs, kFFM improves sliced Wasserstein distance and marginal W_1 over FFM on every dataset, by 18–20% on Gene Expression and Economy, where each comparison is significant when paired by seed, and by 2.9–$4.1\times$ on Navier–Stokes and Stoch. NS; for the signature configurations the coupling and evaluation kernels already differ. The pointwise diagnostics also expose what MMD understates: CFM-OT(L^2) buys its pointwise-marginal fit with $105\times$ and $66\times$ worse autocorrelation error than kFFM-Sig on Economy and Gene Expression (Table 9); function-level structure is where the kernel coupling is strongest.

Bounded cost or function-space geometry? Both effects are real and dominate on different data classes. On smooth fields the RBF coupling is a monotone transform of the L^2 cost (rank correlation

Table 2: kFFM (selected) vs. FFM with uncertainty, at matched architecture and training budget. MMD-RBF$\times 10^3$ as mean$\pm$std over n shared seeds (fixed before outcomes); p-values from a seed-paired t-test and a Wilcoxon signed-rank test. In the gp rows only the coupling changes at a fixed prior; in the wn rows the selected kFFM also uses a white-noise base whereas FFM keeps its GP prior (at that base, the comparison isolating the kernel cost is against CFM-OT(L^2) in Table 1). Economy averages the population and GDP series within each seed. kFFM wins on every shared seed for the first four rows.

Dataset	n	FFM	kFFM (selected)	paired t p-value	Wilcoxon p-value
Gene Expr.	10	17.8 ± 3.4	$\mathbf{7.11 \pm 2.0}$ (Sig/gp)	2.0×10^{-5}	0.002
Economy	10	7.30 ± 1.8	$\mathbf{2.89 \pm 0.97}$ (Sig/gp)	2.5×10^{-5}	0.002
Stoch. NS	10	96.8 ± 21	$\mathbf{2.80 \pm 1.3}$ (RBF/wn)	1.7×10^{-7}	0.002
Navier–Stokes	10	131 ± 60	$\mathbf{1.63 \pm 2.6}$ (RBF/wn)	8.6×10^{-5}	0.002
AEMET	20	-3.24 ± 2.1	$\mathbf{-5.10 \pm 2.6}$ (Sig/gp)	0.015	0.014
Heston	20	0.31 ± 0.80	$\mathbf{-0.13 \pm 0.30}$ (RBF/wn)	0.021	0.011
KdV	20	-1.16 ± 2.1	$\mathbf{-2.56 \pm 1.0}$ (Sob-RBF/gp)	0.039	0.027
Stoch. KdV	20	0.88 ± 1.1	$\mathbf{0.10 \pm 0.33}$ (Sob-RBF/gp)	0.037	0.037

Table 3: Non-kernel metrics for kFFM (selected configuration in parentheses) vs. FFM on the runs of Table 2; lower is better. Sequence rows: mean$\pm$std over 10 shared seeds (Economy averages the population and GDP series within each seed), each comparison significant when paired by seed (Wilcoxon $p = 0.002$); Navier–Stokes rows are means. Appendix Table 16 adds the autocorrelation and log-spectrum errors.

	Sliced-W		Marginal-W_1	
Dataset	FFM	kFFM	FFM	kFFM
Gene Expr. (Sig/gp)	0.106 ± 0.006	$\mathbf{0.087 \pm 0.006}$	0.074 ± 0.005	$\mathbf{0.059 \pm 0.004}$
Economy (Sig/gp)	0.0291 ± 0.0019	$\mathbf{0.0233 \pm 0.0012}$	0.0171 ± 0.0016	$\mathbf{0.0137 \pm 0.0009}$
Navier–Stokes (RBF/wn)	0.315	**0.081**	0.239	**0.058**
Stoch. NS (RBF/wn)	1.06	**0.365**	0.795	**0.249**

≈ 1 on training batches; Appendix C.12), so it reuses the L^2 geometry while compressing outliers, which buys conditioning, ϵ-robustness, and faster Sinkhorn, and matches or improves the unbounded-L^2 coupling at the same base measure (Stoch. NS: 2.80 vs. 4.22×10^{-3}, Table 1). On paths, no Euclidean cost, raw or bounded, recovers the signature gains (matched runs, MMD$\times 10^3$; Gene Expr.: kFFM-Euc 20.0, kFFM-RBF 20.7 vs. kFFM-Sig 7.1; Economy: 7.55/7.60 vs. 2.9), and the signature kernel genuinely re-ranks pairs (rank correlation 0.33–0.37 with L^2); a synthetic example in Appendix C.13 isolates both failure modes of L^2 pairing. Qualitative samples and super-resolution results are shown in Appendix C (Figures 2 and 3).

Kernel choice and sensitivity. Re-aggregating the one-at-a-time sweeps of Appendix C.16: results are indistinguishable within seed-to-seed variation for RBF bandwidths $\sigma \in [0.1, 2]$ on every dataset, flat across three orders of magnitude of the Sinkhorn regularization ($\epsilon \in [0.001, 1]$) for bounded costs (the raw L^2 cost degrades by $\approx 10\times$ at $\epsilon \geq 0.5$ on long volatility paths), and unchanged across the signature hyperparameters (dyadic order 0–3, static bandwidth 0.1–5); over batch sizes $b \in \{64, \ldots, 512\}$, kFFM tracks FFM across the $8\times$ range (Appendix C.8). The kernel *family* is the only choice that matters, and the rule we recommend is simple and consistent with the theory: match the kernel to the data class (signature for path-like data; Sobolev or Euclidean RBF for smooth fields, where the Euclidean RBF is a reasonable default; RBF or L^2 on very long paths, where signature-kernel memory grows with path length), set the bandwidth by the median heuristic, and validate with a non-kernel metric on a held-out split (Appendix C.5). The properties the theorems require (boundedness, characteristicness) hold for every kernel we consider, so the theory constrains the family only weakly and the selection rule does the rest.

4.2 Turbulent Regime and Physics Diagnostics

The Navier–Stokes benchmark above ($\nu = 10^{-3}$) is smooth, the regime where our theory is informative. To test the low-regularity regime where Theorem 2 is weakest, we run the canonical turbulent

Table 4: Turbulent 2D Navier–Stokes ($\nu = 10^{-5}$, 1200 trajectories, $\mathrm{Re} \approx 2000$): distributional and physics diagnostics, lower is better, mean±std over 5 seeds with identical FNO backbone and training budget. Rows marked † use the GP base measure; kFFM and CFM-OT use the white-noise base chosen by the validation rule (Appendix C.5). Best per column in bold (on MMD, kFFM and CFM-OT tie within std). Unscaled values and full configurations are in Appendix Table 17.

Method	MMD ($\times 10^{-3}$)	Sliced-W	Vort-PDF W_1 ($\times 10^{-2}$)	Enstrophy W_1	Skew err. ($\times 10^{-2}$)	Kurt err. ($\times 10^{-2}$)	log-Spec ($\times 10^{-2}$)
kFFM (selected)	$\mathbf{1.48 \pm 0.6}$	$\mathbf{0.193 \pm 0.01}$	$\mathbf{3.8 \pm 1}$	$\mathbf{0.155 \pm 0.02}$	$\mathbf{0.27 \pm 0.3}$	7.8 ± 1	$\mathbf{2.0 \pm 0.8}$
CFM-OT(L^2)	$\mathbf{1.47 \pm 1.0}$	0.194 ± 0.01	4.1 ± 1	0.157 ± 0.01	0.40 ± 0.2	7.5 ± 1	2.2 ± 0.7
kFFM-Sob†	111 ± 20	0.612 ± 0.05	10.4 ± 3	0.194 ± 0.05	4.6 ± 2	$\mathbf{6.8 \pm 5}$	6.5 ± 1
kFFM-Euc†	109 ± 20	0.623 ± 0.03	10.4 ± 4	0.197 ± 0.06	5.0 ± 3	7.7 ± 4	6.7 ± 1
FFM†	108 ± 20	0.601 ± 0.02	10.7 ± 4	0.201 ± 0.06	4.9 ± 3	7.2 ± 5	6.5 ± 1
DDO/NCSN	569 ± 9	1.39 ± 0.06	122 ± 1	1.07 ± 0.02	0.83 ± 1	129 ± 1	93.7 ± 2
DDPM	380 ± 20	1.18 ± 0.1	84.8 ± 3	0.889 ± 0.02	7.9 ± 5	337 ± 200	11.7 ± 3

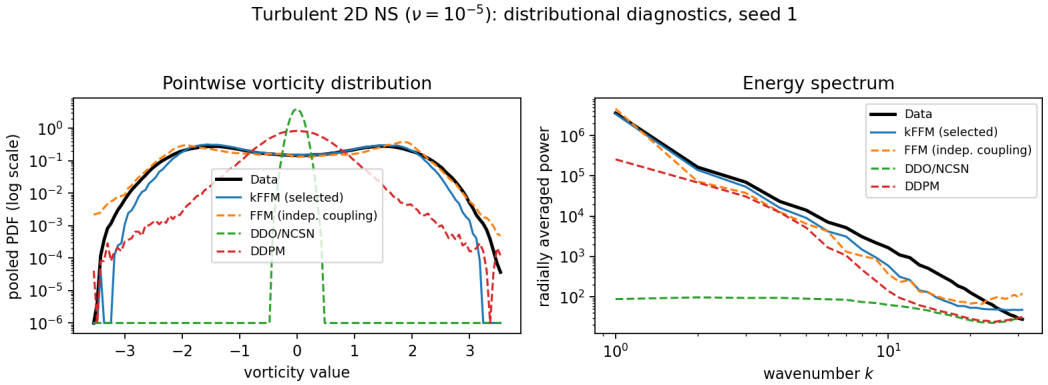


Figure 1: Turbulent 2D Navier–Stokes ($\nu = 10^{-5}$, seed 1): pooled vorticity PDF (log scale) and radially averaged energy spectrum. The data PDF is bimodal and heavy-tailed; kFFM (selected) tracks it, DDO/NCSN collapses to a narrow near-Gaussian spike, and DDPM is unimodal with the wrong shape. The flow-matching spectra track the data across scales; DDO/NCSN's is nearly flat.

FNO benchmark [32]: 2D Navier–Stokes at $\nu = 10^{-5}$ (1200 trajectories, from which we take the spun-up snapshots at $t \geq 10$; $\mathrm{Re} \approx 2000$; seven methods, five seeds, identical FNO backbone and training budget). Table 4 reports MMD, sliced-W, and the physics diagnostics; Figure 1 shows the pooled vorticity PDF and energy spectrum, and vorticity samples are in Appendix C.11. (i) kFFM (selected) attains the best value on five of seven metrics and ties the best on MMD; held-out sliced-W rejects the GP-base arms († rows; 0.19 vs. ≥ 0.60), and among the tied white-noise arms the rule keeps the Euclidean-RBF configuration selected on standard NS; CFM-OT(L^2) at the same base overlaps kFFM within std, with means favoring kFFM on five of seven metrics. (ii) OT-coupled flow matching is far more robust than the score-based and diffusion alternatives: 23–32× better vorticity-PDF W_1, 5.7–6.9× better enstrophy W_1, and 17–43× better vorticity-kurtosis error (Figure 1). (iii) Against FFM run with its GP prior, kFFM improves every metric except kurtosis (a tie), with seed-paired significance (MMD 73× lower, $p = 3 \times 10^{-4}$; sliced-W 3.1×, $p = 1.3 \times 10^{-6}$; log-spectrum 3.2×, $p = 0.005$; vorticity-PDF 2.9×, $p = 0.007$; skewness 18×, $p = 0.016$). This gain combines the base-measure change and the coupling: the three GP-base arms are indistinguishable, so in this regime the white-noise base does most of the work, and at that base the kernel and raw-L^2 costs tie, as in (i). All methods, ours included, still leave headroom on sliced-W in this regime (Section 6).

4.3 Cost, Efficiency, and Convergence

Compute. Measured on every benchmark (Appendix C.9), the end-to-end training overhead of the per-minibatch Sinkhorn solve over FFM is 5–7% at the largest resolution we train (128^2 Navier–Stokes); the $O(b^2)$ coupling cost is independent of model size and resolution, so its relative weight

shrinks as the backbone grows. Peak coupling memory is 80–151 MB (1.5–9% of the FNO's forward-plus-backward peak), and per-batch coupling time is comparable to one FNO step. The bounded RBF cost is about $8\times$ faster than unbounded-L^2 Sinkhorn at $b = 512$ and trains 15–25% faster end-to-end, the conditioning effect of Theorem 1; giving FFM the extra budget as additional epochs does not close the gap.

No straightness claim. We claim no straightness theorem: Appendix C.14 (Proposition 6, Corollary 2) shows that kFFM changes only the endpoint coupling while the affine interpolation template is fixed, and Hilbert-space Gaussian geometry does not inherit the Euclidean straightness picture that motivates finite-dimensional OT-CFM [53]. We therefore make no efficiency claim; Appendix C.15 reports only the training-time convergence of the target metric, where kFFM ends training at a better MMD-RBF than FFM on every dataset.

5 Related Work

Minibatch OT coupling for flow matching was introduced by Tong et al. [49] and Pooladian et al. [40], with follow-ups refining the transport objective [28, 52] and analyses of minibatch plans [13, 14]; all assume $\mathbb{R}^d$ sample spaces. In function space, FFM [25] lifts flow matching to Hilbert spaces via Gaussian measures, alongside functional diffusion [24, 34] and adversarial neural operators [41]. Concurrent function-space flow matching changes the objective, the task, or the prior, but not the coupling: Functional Rectified Flow [54] straightens flows by re-training on model-generated endpoint pairs across training rounds, a mechanism complementary to and composable with our within-batch, data-geometry coupling; Operator Flow Matching [30] and flow matching with GP priors [27] target conditional time-series forecasting with an independent prior-data coupling. Kernel embeddings of measures [20, 47] underlie MMD and two-sample testing, recently on path space with the signature kernel [6]; Sinkhorn divergences interpolate between OT and MMD [16], and the Hilbert Sinkhorn Divergence [31] kernelizes them. To our knowledge, ours is the first work to perform OT endpoint coupling in function-space flow matching; Appendix D gives an extended discussion, including MMD gradient flows and Stein variational particle methods [36].

6 Conclusion and Limitations

kFFM replaces the independent endpoint pairing of FFM with a geometry-aware kernel OT coupling while keeping the neural-operator backbone. Theoretically, the kernel-OT surrogate is uniformly bounded and well-posed on Banach spaces, admits an explicit error decomposition against quadratic-cost OT that isolates an irreducible kernel-cost mismatch, and is discretization-invariant at rate $N^{-2(\alpha-k)/d}$. Empirically, kFFM improves distributional matching on time-series and PDE benchmarks with paired-seed significance, and the gains survive non-kernel metrics, physics diagnostics, and a turbulent Navier–Stokes stress test, at a training overhead of 5–7% at the largest resolution we train.

Limitations. Several deserve plain statement. First, our theory concerns the surrogate coupling: it certifies well-posedness, an error decomposition, and discretization invariance, but says nothing about straightness or sampling efficiency in function space, on which we make no claim, and it does not by itself explain the sample-quality gains, which remain an empirical finding. Second, the theory is least informative in low-regularity regimes: on the turbulent benchmark the prior-data L^2 distances more than double relative to the standard one (median 91 vs. 40), pushing the RBF cost into saturation, where Δ_κ grows and the covering term of Theorem 2 is large for any practical ϵ; there the empirical results carry the method. Third, turbulent small-scale statistics remain an open problem for function-space generative models. On current evidence the failure sits mainly with the score-based and diffusion baselines (vorticity-kurtosis errors of 1.3–3.4 and enstrophy $W_1 \approx 1$), but all methods, ours included, leave headroom on sliced-W (0.19 for the best method), and in this regime the white-noise base rather than the coupling does most of the work; OT-coupled flow matching is currently the strongest option. Finally, signature-kernel memory grows with path length (0.63 GB at 100 time points vs. 12.5 GB at 512), so very long paths fall back to RBF or L^2 costs. Future work includes stronger function-space OT formulations [28, 53] and base measures and costs designed for rough, turbulent data.

Acknowledgments and Disclosure of Funding

This work was partially supported by NSF 2303037, NSF 2312501, NSF 2531008, NSF 2553619, SRC JUMP 2.0 Center, CDSC Industry Partners, Amazon Research Awards, Snapchat, and Google Gifts. Barbora Barancikova is supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. Fred Xu is partially supported by NSF grant 2531008 and was supported by Block, Inc. during an internship.

References

[1] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/944a5ae3483ed5c1e10bbccb7942a279-Abstract.html.

[2] Barbora Barancikova, Zhuoyue Huang, and Cristopher Salvi. SigDiffusions: Score-based diffusion models for time series via log-signature embeddings. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Y8KK9kjgIK.

[3] Jutta Bolt and Jan Luiten van Zanden. Maddison style estimates of the evolution of the world economy. A new 2020 update. Maddison-Project Working Paper WP-15, The Maddison Project, October 2020. URL https://www.rug.nl/ggdc/historicaldevelopment/maddison/publications/wp15.pdf.

[4] Guher Camliyurt, Igor Kukavica, and Vlad Vicol. Analyticity up to the boundary for the Stokes and the Navier–Stokes systems. *Transactions of the American Mathematical Society*, 373(5): 3375–3422, 2020. doi: 10.1090/tran/7990. URL https://doi.org/10.1090/tran/7990.

[5] Thomas Cass and Cristopher Salvi. Lecture notes on rough paths and applications to machine learning, 2024. URL https://arxiv.org/abs/2404.06583.

[6] Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *Journal of Machine Learning Research*, 23(176):1–42, 2022. URL https://jmlr.org/papers/v23/20-1466.html.

[7] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/4e0cb6fb5fb446d1c92ede2ed8780188-Paper.pdf.

[8] Giuseppe Da Prato. *An Introduction to Infinite-Dimensional Analysis*. Universitext. Springer Berlin Heidelberg, 1st edition, 2006. ISBN 978-3-540-29020-9. doi: 10.1007/3-540-29021-4. URL https://link.springer.com/book/10.1007/3-540-29021-4.

[9] Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Optimal transport map estimation in general function spaces. *The Annals of Statistics*, 53(3):963–988, 2025. doi: 10.1214/24-AOS2482. URL https://doi.org/10.1214/24-AOS2482.

[10] R. M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2nd edition, 2002. ISBN 978-0-521-00754-2. URL https://www.cambridge.org/core/books/real-analysis-and-probability/26DDF2D09E526185F2347AA5658B96F6.

[11] Nathaniel Eldredge. Analysis and probability on infinite-dimensional spaces, 2016. URL https://arxiv.org/abs/1607.03591.

[12] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2nd edition, 2010. ISBN 978-0-8218-4974-3. URL https://bookstore.ams.org/gsm-19-r.

[13] Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein: asymptotic and gradient properties. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[14] Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

[15] Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012. doi: 10.18637/jss.v051.i04. URL https://www.jstatsoft.org/article/view/v051i04.

[16] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

[17] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL https://jmlr.org/papers/v22/20-451.html.

[18] Alexandre Galashov, Valentin De Bortoli, and Arthur Gretton. Deep MMD gradient flow without adversarial training. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pf85K2wtz8.

[19] Jacob R. Gardner, Geoff Pleiss, Kilian Q. Weinberger, David Bindel, and Andrew Gordon Wilson. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/27e8e17134dd7083b050476733207ea1-Abstract.html.

[20] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/hash/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Abstract.html.

[21] Steven L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343, April 1993. doi: 10.1093/rfs/6.2.327. URL https://doi.org/10.1093/rfs/6.2.327.

[22] Bamdad Hosseini, Alexander W. Hsu, and Amirhossein Taghvaei. Conditional optimal transport on function spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):304–338, 2025. doi: 10.1137/23M1618922. URL https://epubs.siam.org/doi/10.1137/23M1618922.

[23] Zacharia Issa, Blanka Horvath, Maud Lemercier, and Cristopher Salvi. Non-adversarial training of neural SDEs with signature kernel scores. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2460396f2d0d421885997dd1612ac56b-Abstract-Conference.html.

[24] Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 9538–9563. PMLR, April 2023. URL https://proceedings.mlr.press/v206/kerrigan23a.html.

[25] Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Functional flow matching. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3934–3942. PMLR, May 2024. URL https://proceedings.mlr.press/v238/kerrigan24a.html.

[26] Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pages 6696–6707, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

[27] Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with Gaussian process priors for probabilistic time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=UJ5WggV1xG.

[28] Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3310. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bc8f76d9caadd48f77025b1c889d2e2d-Abstract-Conference.html.

[29] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*. Springer-Verlag, Berlin, Heidelberg, 1991. URL https://link.springer.com/book/10.1007/978-3-642-20212-4.

[30] Yolanne Yi Ran Lee and Kyriakos Flouris. Operator flow matching for timeseries forecasting, 2025. URL https://arxiv.org/abs/2510.15101.

[31] Qian Li, Zhichao Wang, Gang Li, Jun Pang, and Guandong Xu. Hilbert Sinkhorn divergence for optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3835–3844, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Hilbert_Sinkhorn_Divergence_for_Optimal_Transport_CVPR_2021_paper.html.

[32] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

[33] Zongyi Li, Miguel Liu-Schiaffini, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Learning chaotic dynamics in dissipative systems. In *Advances in Neural Information Processing Systems*, volume 35, pages 16768–16781, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6ad68277e27b42c60ac228c9859fc1a2-Abstract-Conference.html.

[34] Jae Hyun Lim, Nikola B. Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, Christopher Pal, Arash Vahdat, and Anima Anandkumar. Score-Based diffusion models in function space. *Journal of Machine Learning Research*, 26(158):1–62, 2025. URL https://jmlr.org/papers/v26/23-1472.html.

[35] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

[36] Safa Messaoud, Skander Charni, Elaa Bouazza, Ali Pourghasemi, and Halima Bensmail. Particles don't care about Z: Towards scaling entropy estimation of unnormalized densities. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=nVAU1aVLtY.

[37] David A. Orlando, Charles Y. Lin, Allister Bernard, Jean Y. Wang, Joshua E. S. Socolar, Edwin S. Iversen, Alexander J. Hartemink, and Steven B. Haase. Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature*, 453(7197):944–947, June 2008. doi: 10.1038/nature06955. URL https://www.nature.com/articles/nature06955.

[38] John C. Oxtoby. Invariant measures in groups which are not locally compact. *Transactions of the American Mathematical Society*, 60(2):215–237, September 1946. ISSN 0002-9947. doi: 10.1090/S0002-9947-1946-0018188-5. URL https://doi.org/10.1090/S0002-9947-1946-0018188-5.

[39] Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends® in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073.

[40] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28100–28127. PMLR, 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

[41] Md Ashiqur Rahman, Manuel A. Florez, Anima Anandkumar, Zachary E. Ross, and Kamyar Azizzadenesheli. Generative adversarial neural operators. *Transactions on Machine Learning Research*, October 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=X1VzbBU6xZ.

[42] Walter Rudin. *Functional Analysis*. McGraw-Hill Series in Higher Mathematics. McGraw-Hill, New York, 1973. ISBN 978-0-07-054225-9. URL https://books.google.com/books/about/Functional_Analysis.html?id=BB_vAAAAMAAJ.

[43] Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The signature kernel is the solution of a Goursat PDE. *SIAM Journal on Mathematics of Data Science*, 3 (3):873–899, 2021. doi: 10.1137/20M1366794. URL https://epubs.siam.org/doi/10.1137/20M1366794.

[44] Cristopher Salvi, Maud Lemercier, and Andris Gerasimovics. Neural stochastic PDEs: Resolution-invariant learning of continuous spatiotemporal dynamics. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/091166620a04a289c555f411d8899049-Abstract-Conference.html.

[45] Daniil Shmelev and Cristopher Salvi. pySigLib – fast signature-based computations on CPU and GPU, 2025. URL https://arxiv.org/abs/2509.10613.

[46] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

[47] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010. URL https://jmlr.org/papers/v11/sriperumbudur10a.html.

[48] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport, 2019. URL https://arxiv.org/abs/1910.12958.

[49] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, pages 1–34, March 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW.

[50] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9. URL https://link.springer.com/book/10.1007/978-3-540-71050-9.

[51] Yasuo Yamasaki. *Measures on Infinite Dimensional Spaces*, volume 5 of *Series in pure mathematics*. World Scientific, 1985. ISBN 9789971978525. URL https://books.google.com/books?id=F243SflQYbQC.

[52] Angxiao Yue, Anqi Dong, and Hongteng Xu. OAT-FM: Optimal acceleration transport for improved flow matching, 2025. URL https://arxiv.org/abs/2509.24936.

[53] Ho Yun and Yoav Zemel. Gaussian optimal transport beyond Brenier's theorem, 2025. URL https://arxiv.org/abs/2512.21464.

[54] Jianxin Zhang and Clayton Scott. Flow straight and fast in hilbert space: Functional rectified flow, 2025. URL https://arxiv.org/abs/2509.10384.

[55] Johanna Ziegel, David Ginsbourger, and Lutz Dümbgen. Characteristic kernels on Hilbert spaces, Banach spaces, and on sets of measures, 2022. URL https://arxiv.org/abs/2206.07588.

A Additional Technical Background

This appendix collects the mathematical background on which the paper relies: probability measures on function spaces, reproducing kernel Hilbert spaces and kernel embeddings, path signatures, and optimal transport.

A.1 Probability Measure in Function Space

In this section we provide the definitions and necessary theoretical results for probability measure in function space, highlighting the analytical difficulties in general. We give high-level intuition and omit the proofs; for details we refer the reader to the textbooks of Rudin [42] and Da Prato [8].

Definition 3 (Hilbert Space)**.** A Hilbert space $\mathcal{H}$ is an inner-product space that is complete in the norm induced by its inner product, defined as $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{F}$, where $\mathbb{F} = \mathbb{R}$ for real Hilbert space.

Importantly, the Hilbert space's inner product structure gives it a metric space with the norm $||x|| := \sqrt{\langle x, x \rangle}$ and metric $d(x, y) = ||x - y||, \quad \forall x, y \in \mathcal{H}$. Let $B(x, r)$ be the open balls defined using the metric d, then the Borel σ-algebra of $\mathcal{H}$, denoted $\mathcal{B}(\mathcal{H})$, is generated by the open balls defined in this way:

$$\mathcal{B}(\mathcal{H}) := \sigma(\{B(x, r) : x \in \mathcal{H}, r > 0\}).$$

Another desirable property of the Hilbert space is eigenfunction spectral decomposition for bounded linear operators:

Theorem 4. *Let $T : \mathcal{H} \to \mathcal{H}$ be a compact, self-adjoint, bounded linear operator. Then there exists an orthonormal set of eigenvectors $\{e_n\}_{n \in I}$ and real eigenvalues $\{\lambda_n\}_{n \in I}$ such that:*

1. $Te_n = \lambda_n e_n$
2. $\lambda_n \to 0$ *if I is infinite.*
3. $\lambda_i \neq 0$ *has finite multiplicity.*
4. *The closure of span of non-zero eigenvectors produces closure of $Ran(T)$.*
5. *The spectral decomposition of T is:*

$$Tx = \sum_{i \in I} \lambda_i \langle x, e_i \rangle e_i, \qquad x \in \mathcal{H}$$

In the main text, we adopt a measure-theoretic view of probability, which is crucial for function spaces. This is due to the following result, known as "no Lebesgue measure in infinite dimensions." The statement usually takes the form: If $\mathcal{X}$ is an infinite-dimensional separable normed linear space, then there is no nontrivial, translation-invariant, σ-additive Borel measure μ on $\mathcal{X}$ that is finite on every nonempty open ball [51]. The implication of this result is that there is no universal definition of **probability density** as in $\mathbb{R}^d$, and only the more general Radon-Nikodym derivative can be defined between two measures that are absolutely continuous.

Definition 4 (Singular and Absolute Continuity)**.** For a measurable space $(\Omega, \mathcal{F})$, two measures μ, ν are mutually singular, written $\mu \perp \nu$, if there exists $A \in \mathcal{F}$, such that $\mu(A) = 0$ and $\nu(\Omega \setminus A) = 0$. μ is said to be absolutely continuous w.r.t ν, denoted $\mu \ll \nu$, if $\nu(A) = 0 \implies \mu(A) = 0$. If $\mu \ll \nu$ and $\nu \ll \mu$, then we say that $\mu \sim \nu$, or that they are equivalent.

Definition 5 (Radon-Nikodym Derivative)**.** If $\mu \ll \nu$ and μ, ν are $\sigma-$finite, then there exists a measurable function $f : \Omega \to [0, \infty]$ such that for all $A \in \mathcal{F}$:

$$\mu(A) = \int_A f d\nu,$$

where f is called the Radon-Nikodym derivative of μ w.r.t ν, denoted $\frac{d\mu}{d\nu}$. When the measure space is $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and ν is the Lebesgue measure, f is also called the probability density function of distribution μ.

Though there is no infinite-dimensional Lebesgue measure, and hence no universal way to define density, the most feasible and well-understood way to have a Lebesgue-like measure is to define Gaussian measures:

Definition 6 (Gaussian Measure)**.** Let $a \in \mathcal{H}$ and $Q \in L_1^+(\mathcal{H})$, the space of positive, symmetric, and trace class linear operators, then the Gaussian measure $\mu := N_{a,Q}$ on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ with mean a and covariance operator Q is defined by the Fourier transform:

$$\widehat{N_{a,Q}}(h) = \exp\{i\langle a, h\rangle - \frac{1}{2}\langle Qh, h\rangle\}$$

It can be shown that for every choice of a, Q, there exists a unique Gaussian measure. In Da Prato [8], Chapter 1, the interesting argument for constructing this measure is by the theory of infinite product measure and the isomorphism between the sequence space l^2 and the Hilbert space $\mathcal{H}$, where the projection map defined using the indices of the spectral decomposition is mapped with Gaussian measures in $\mathbb{R}$ (also known as 1D Gaussian distribution).

Still, for one Gaussian measure to be absolutely continuous with respect to another, strong conditions have to be imposed on their means and covariance operators: the Feldman–Hájek theorem and, for a shifted mean, the Cameron–Martin formula.

Definition 7 (Cameron–Martin Space)**.** Let $\mathcal{H}$ be a Hilbert space and $\mu = N_{0,Q}$ a centered Gaussian measure, then the Cameron–Martin space induced by $N_{0,Q}$ is:

$$H_\mu := \text{Ran}(Q^{1/2}),$$

with inner product $\langle h_1, h_2\rangle := \langle Q^{-1/2}h_1, Q^{-1/2}h_2\rangle_{\mathcal{H}}$.

The general dichotomy for Gaussian measures is the Feldman–Hájek theorem:

Theorem 5 (Feldman–Hájek)**.** *Let $\mathcal{H}$ be a separable Hilbert space and $N_{a,P}, N_{b,Q}$ be two Gaussian measures, then either $N_{a,P} \sim N_{b,Q}$ or $N_{a,P} \perp N_{b,Q}$, and $N_{a,P} \sim N_{b,Q}$ iff:*

1. $P^{1/2}(\mathcal{H}) = Q^{1/2}(\mathcal{H}) =: \mathcal{H}_0$.
2. $a - b \in \mathcal{H}_0$.
3. $(P^{-1/2}Q^{1/2})(P^{-1/2}Q^{1/2})^* - I$ *is Hilbert-Schmidt on* $\mathcal{H}_0$

Theorem 6 (Cameron–Martin Formula)**.** *Let $\mu = N_Q, \nu = N_{a,Q}$ be two Gaussian measures, then:*

1. *If $a \notin Q^{1/2}(\mathcal{H})$, then $\mu \perp \nu$.*
2. *If $a \in Q^{1/2}(\mathcal{H})$, then $\mu \sim \nu$.*
3. *If $\mu \sim \nu$, then with $h := Q^{-1/2}a$,*

$$\frac{d\nu}{d\mu}(x) = \exp\left\{\widehat{h}(x) - \frac{1}{2}\|h\|_{\mathcal{H}}^2\right\},$$

where $\widehat{h}$ denotes the Cameron–Martin linear functional associated with h.

These are the conditions enforced on the FFM model [25], which relies on the absolute continuity between the marginal and the conditional measure at every time t in the probability flow.

For a general Banach space, which is a complete normed space but need not have an inner product structure or spectral decomposition, the analysis becomes more abstract. For instance, without a coordinate map, the preceding product-measure construction of Gaussian measures no longer applies directly. We refer readers to Ledoux and Talagrand [29] for this broader probability theory, and instead explore Hilbert-space embeddings of probability measures in Section A.3.

A.2 Reproducing Kernel Hilbert Space

A Reproducing Kernel Hilbert Space is a Hilbert space of functions with a particular structure that makes each point evaluation well-behaved.

Definition 8 (Reproducing Kernel Hilbert Space)**.** Let $\mathcal{X}$ be a non-empty set. A Hilbert space, $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is called a Reproducing Kernel Hilbert Space if for every $x \in \mathcal{X}$, the point evaluation functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by $\delta_x(f) = f(x)$ is continuous and bounded.

By the Riesz representation theorem, there exists a unique element $\kappa_x \in \mathcal{H}$ such that

$$f(x) = \langle f, \kappa_x \rangle_{\mathcal{H}} \ \ \forall \ \ f \in \mathcal{H}.$$

This is the reproducing property that is the defining characteristic of an RKHS.

Since κ_x is both a function defined on $\mathcal{X}$ and in $\mathcal{H}$, we have that

$$\kappa_x(y) = \delta_y(\kappa_x) = \langle \kappa_x, \kappa_y \rangle_{\mathcal{H}},$$

where $\kappa_y \in \mathcal{H}$ is the element in $\mathcal{H}$ associated with δ_y. This allows us to define the reproducing kernel, $\kappa(x, y)$ as:

$$\kappa(x, y) = \langle \kappa_x, \kappa_y \rangle_{\mathcal{H}}.$$

From this, it is easy to see that $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is both symmetric and positive definite.

Arguably the core reason to use kernels is that they implicitly define feature maps in high (and potentially infinite) dimensional spaces, which is known colloquially as the "kernel trick". We next formalize this property using the properties of the RKHS:

Definition 9 (Canonical Feature Maps)**.** Let κ be a kernel with RKHS $\mathcal{H}_\kappa$. The canonical feature map $\phi : \mathcal{X} \to \mathcal{H}_\kappa$ is defined by

$$\phi(x) = \kappa(\cdot, x),$$

which satisfies the fundamental identity $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle$, which is known as the "kernel trick". This allows us to compute inner products in the feature space without explicit construction.

In the context of our work, we can understand the cost function in Proposition 3 by making use of the kernel trick.

For some of our kernel choices, it is useful to know when the RKHS is rich enough to approximate arbitrary continuous functions and when the induced mean embedding distinguishes probability distributions. Kernels with the former property are referred to as "universal", defined as follows:

Definition 10 (Universal Kernel)**.** Let $\mathcal{X}$ be a compact metric space. A continuous kernel on $\mathcal{X}$ is called universal if its associated RKHS, $\mathcal{H}_\kappa$ is dense in $C(\mathcal{X})$ with respect to the supremum norm, $|| \cdot ||_\infty$. That is to say that for any $f \in C(\mathcal{X})$ and for any $\epsilon > 0$, there exists $g \in \mathcal{H}_\kappa$ such that

$$\sup_{x \in \mathcal{X}} |f(x) - g(x)| < \epsilon.$$

Universality ensures that a kernel can approximate continuous functions, while characteristicness ensures that its mean embedding can distinguish probability distributions. A kernel with the latter property is called "characteristic":

Definition 11 (Characteristic Kernel)**.** Let $\mathcal{X}$ be a compact metric space. A kernel κ on $\mathcal{X}$ is characteristic if the mean embedding map

$$\mu \mapsto \int_{\mathcal{X}} \kappa(\cdot, x) \, d\mu(x)$$

is injective on the space of probability measures $\mathcal{P}(\mathcal{X})$. Equivalently, κ is characteristic if

$$\int \int \kappa(x, y) d\mu(x) d\mu(y) - 2 \int \int \kappa(x, y) d\mu(x) d\nu(y) + \int \int \kappa(x, y) d\nu(x) d\nu(y) = 0$$

implies that $\mu = \nu$.

Universal kernels on compact spaces are characteristic, but the converse does not hold in general. These properties are useful for interpreting kernel discrepancies: characteristicness makes the kernel mean embedding injective, while universality gives a stronger approximation property on compact sets. The main theorems in Section 3.2 only require the specific assumptions stated there, namely bounded positive-definite kernels, compactness for the error decomposition, and Sobolev regularity for discretization invariance.

The radial basis function kernel is used in our Sobolev-space experiments. On a Sobolev space $H^k(\Omega)$ it is defined as

$$\kappa_{\mathrm{RBF}}^{(k)}(f,g) = \exp\left\{-\frac{\|f-g\|_{H^k}^2}{2\sigma^2}\right\},$$

where $\sigma > 0$ is the bandwidth parameter. This Gaussian kernel is symmetric, positive definite, and characteristic on separable Hilbert spaces [55]. It is also bounded, $\kappa_{\mathrm{RBF}}^{(k)}(f,f) = 1$, satisfying the boundedness condition used in Theorems 1 and 2.

A.3 Kernel Embedding of Probability Measure

We next turn our attention to the concept of embedding probability measures into a RKHS, following the framework developed by Sriperumbudur et al. [47]. This is central to our approach because it gives a tractable way to define kernel-induced transport costs on function spaces. As previously discussed, infinite-dimensional spaces lack a Lebesgue-like reference measure, making density-based approaches unavailable. The key insight is that we can embed probability measures as elements into a RKHS, where we compute distances between distributions without requiring explicit density estimation.

For this embedding to be well defined, we require an integrability condition on the kernel. The following proposition shows that bounded kernels ensure the embedding is valid for all probability measures [47]:

Proposition 1. *Let κ be a measurable kernel on $\mathcal{X}$. Then, $\int_{\mathcal{X}} \sqrt{\kappa(x,x)}\, dP(x) < \infty \ \forall\ P \in \mathcal{P}(\mathcal{X})$ if and only if κ is bounded.*

With the mean embedding in hand, we can define a natural notion of distance between probability measures. This construct relies on the concept of characteristic kernel (Definition 11). If the kernel is characteristic, then the metrics γ_κ can be defined such that:

$$\gamma_\kappa(\mathbb{P},\mathbb{Q}) = 0 \Longleftrightarrow \mathbb{P} = \mathbb{Q}, \mathbb{P}, \mathbb{Q} \in \mathcal{P}$$

The bounded condition and the characteristic condition gives the following result, which is Theorem 1 in Sriperumbudur et al. [47]:

Theorem 7. *Let $\mathcal{P}_\kappa := \{\mathbb{P} \in \mathcal{P} : \int_M \sqrt{\kappa(x,x)} d\mathbb{P}(x) < \infty\}$, where κ is measurable on M, then for any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_\kappa$:*

$$\gamma_\kappa(\mathbb{P},\mathbb{Q}) = \left|\left|\int_M \kappa(\cdot,x) d\mathbb{P}(x) - \int_M \kappa(\cdot,x) d\mathbb{Q}(x)\right|\right|_{\mathcal{H}} := ||\mathbb{P}\kappa - \mathbb{Q}\kappa||_{\mathcal{H}}$$

Below we show that the bounded RBF kernel defined on a Sobolev space H^k is characteristic, to provide additional support for the kernel choices in the main text. First we state a lemma proved in Sriperumbudur et al. [47]:

Lemma 1. *Let κ be an integrally strictly positive definite kernel on a topological space M. Then κ is characteristic to $\mathcal{P}$.*

Definition 12. Let $M(H)$ be the finite signed Borel measures on H. A bounded measurable kernel κ on H is *integrally strictly positive definite (ISPD)* if for every $\mu \in M(H) \setminus \{0\}$,

$$\int\int_{H\times H} \kappa(x,y) d\mu(x) d\mu(y) > 0,$$

or equivalently, if $m_\kappa(\mu) := \int \kappa(\cdot,x) d\mu(x)$ is the *kernel mean element* (KME), then:

$$\int\int \kappa(x,y) d\mu(x) d\mu(y) = ||m_\kappa(\mu)||_{\mathcal{H}_\kappa}^2.$$

One can interpret ISPD as the statement that a zero KME implies a zero measure: $m_\kappa(\mu) = 0 \implies \mu = 0$.

We then reference Theorem 3.1 from Ziegel et al. [55], which shows that RBF / Gaussian kernel is ISPD on separable Hilbert space:

Theorem 8. *Let H be a separable Hilbert space, then the Gaussian kernel on H is ISPD with respect to $M(H)$.*

This theorem and Lemma 1 therefore imply that the RBF kernel is characteristic on our Sobolev spaces of interest.

A.4 Path Space and Signature Kernel

Denote by $\mathcal{C}_p$ the space of continuous paths $x : [a, b] \to V$ of finite p-variation (see Definition 1.1.4 in Cass and Salvi [5]), with $p \in [1, 2)$. V is a finite-dimensional Banach space. For time series datasets, we usually have $x : [0, T] \to \mathbb{R}^d$. For discrete observations, one may work with a piecewise-linear interpolation, which is again of finite p-variation. Let $\mathcal{T}$ be the tensor algebra (written $T((\mathbb{R}^d))$ in Section 2.3, where $V = \mathbb{R}^d$)

$$\mathcal{T} \;:=\; \prod_{m=0}^{\infty} V^{\otimes m}.$$

Definition 13 (The signature transform)**.** The *signature transform* $S : \mathcal{C}_p \to \mathcal{T}$ maps a path x to the sequence of tensors

$$S(x) \;=\; \big(1, S^{(1)}(x), S^{(2)}(x), \ldots\big),$$

where the m-th tensor is given by the iterated integral

$$S^{(m)}(x) \;:=\; \int_{a \le t_1 < \cdots < t_m \le b} dx_{t_1} \otimes \cdots \otimes dx_{t_m} \;\in\; V^{\otimes m}, \qquad m \in \mathbb{N}.$$

Here $\otimes$ denotes the standard tensor product.

One can view $S(x)$ as an infinite-dimensional feature representation of the path x, with higher signature levels $S^{(m)}(x)$ (the level-m term of $S(x)$; we use this notation throughout) encoding increasingly high-order interactions between the d channels.

Definition 14 (Time augmentation)**.** The signature is invariant under continuous and non-decreasing time reparametrizations [5, Lemma 1.2.1]. In many machine learning settings, this invariance is not desirable, as we want to capture time series in a way that also depends on their time parametrisation. Following prior work [2, 26], we therefore use the standard *time augmentation* trick. For a path $x : [a, b] \to V$, we define the augmented path as

$$\bar{x} : [a, b] \to \mathbb{R} \times V, \qquad \bar{x}(t) := (t, x(t)).$$

Throughout, we apply the signature transform to $\bar{x}$ rather than x. For notational simplicity, we suppress the bar in the definitions below. Note the signature is also invariant under constant translations of x (see Section 1.4 in Cass and Salvi [5]). Throughout, we either fix $x(a) = 0$ or treat paths modulo translation.

Remark 1 (Tree-like equivalence without time augmentation)*.* Without the additional time coordinate, the signature is injective only up to *tree-like equivalence* [5, Section 1.4]. Accordingly, statements such as universality and characteristicness of the signature kernel on a set K should be understood as holding on the corresponding quotient space of equivalence classes when time augmentation is not used.

When working directly with signatures, one typically truncates $S(x)$ at level L, defining the truncated signature $S_{\le L}(x) = (1, S^{(1)}(x), \ldots, S^{(L)}(x))$. However, if one is interested in comparing paths $x, y \in \mathcal{C}_p$ and defining the corresponding kernel-induced discrepancies between distributions, the *signature kernel* provides a way to do so without truncation by working with inner products on the full tensor algebra. Concretely, equip V with an inner product $\langle \cdot, \cdot \rangle_V$. This induces canonical inner products $\langle \cdot, \cdot \rangle_m$ on each tensor space $V^{\otimes m}$, and hence an inner product $\langle \cdot, \cdot \rangle_{\mathcal{T}}$ on $\mathcal{T}$ by linearity across levels.

Definition 15 (Signature kernel)**.** For two paths $x, y \in \mathcal{C}_p$, the signature kernel $\kappa_{\mathrm{sig}} : \mathcal{C}_p \times \mathcal{C}_p \to \mathbb{R}$ is defined as the inner product of their signatures,

$$\kappa_{\mathrm{sig}}(x, y) \;:=\; \langle S(x), S(y) \rangle_{\mathcal{T}}.$$

It is symmetric and positive semidefinite, and thus defines an RKHS on $\mathcal{C}_p$.

Crucially, κ_{sig} can be computed without explicitly forming the signature tensors, via the signature kernel trick introduced in Salvi et al. [43, Theorem 2.5]. In particular, one can evaluate

$$\kappa_{\mathrm{sig}}(x, y) \;=\; K(b, b),$$

where $K : [a, b]^2 \to \mathbb{R}$ is the signature kernel between two path prefixes, $K(s, t) := \langle S(x_{[a,s]}), S(y_{[a,t]}) \rangle$, which satisfies the integral equation

$$K(s, t) \;=\; 1 + \int_a^s \int_a^t K(u, v) \, \langle dx_u, dy_v \rangle_V .$$

When x and y are continuously differentiable, this reduces to a linear hyperbolic Goursat-type PDE with boundary conditions $K(a, \cdot) = K(\cdot, a) = 1$, for which standard finite-difference solvers apply. See Salvi et al. [43, Section 3.1] for one possible finite-difference scheme. We use the PYSIGLIB library from Shmelev and Salvi [45] for efficient signature kernel computations.

Proposition 2 (Universality of the signature kernel [5])**.** *Let $\mathcal{C}$ be a path space equipped with a topology such that the signature transform $S : \mathcal{C} \to \mathcal{T}$ is continuous into a tensor Hilbert space $\mathcal{T}$ for which the signature kernel*

$$\kappa_{\mathrm{sig}}(x, y) := \langle S(x), S(y) \rangle_{\mathcal{T}}$$

is well-defined. Then on compact $K \subset \mathcal{C}$, the restriction $\kappa_{\mathrm{sig}}|_{K \times K}$ is universal in the sense of Definition 10, and hence also characteristic to $\mathcal{P}(K)$.

A precise statement of universality and its proof are given in Cass and Salvi [5, Theorem 2.2.3]. Characteristicness on compact K follows from the relation between universality and characteristicness established in Cass and Salvi [5, Theorem 2.2.7].

For a more detailed overview of signatures and signature kernels, we refer the reader to Cass and Salvi [5]. The compact-set universality and characteristicness statements used here can be found specifically in Chapter 2.2.

A.5 Optimal Transport and Sinkhorn Divergence

This section introduces the static optimal transport problem and its entropy-regularized version, which connects to the theory of statistical divergences.

Let X, Y be metric spaces with Borel σ-algebras $\mathcal{B}(X), \mathcal{B}(Y)$, let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ be probability measures, and let $c : X \times Y \to \mathbb{R}$ be a measurable cost function.

Definition 16 (Coupling)**.** Let $\mathrm{pr}_X(x, y) = x$ and $\mathrm{pr}_Y(x, y) = y$ be the coordinate projections. Then the set of couplings (transport plans) is:

$$\Pi(\mu, \nu) := \Big\{ \pi \in \mathcal{P}(X \times Y) \;:\; (\mathrm{pr}_X)_{\#}\pi = \mu,\; (\mathrm{pr}_Y)_{\#}\pi = \nu \Big\}.$$

Definition 17 (Kantorovich Problem)**.** The static optimal transport cost associated with c is

$$\mathsf{OT}_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y) \, \pi(dx, dy).$$

Any minimizer $\pi^\star \in \Pi(\mu, \nu)$ (if it exists) is called an *optimal transport plan*.

Definition 18 (Wasserstein-p metrics)**.** If $X = Y$ is a metric space (X, d) and $c(x, y) = d(x, y)^p$ with $p \geq 1$, define

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x, y)^p \, d\pi(x, y) \right)^{1/p},$$

for μ, ν in $\mathcal{P}_p(X) := \big\{ \rho \in \mathcal{P}(X) : \int d(x, x_0)^p \, d\rho(x) < \infty \big\}$.

In the main text, $\mathcal{W}(\mu,\nu)$ denotes $\mathsf{OT}_c(\mu,\nu)$ with the quadratic cost $c = d^2$, that is, $W_2^2(\mu,\nu)$. Next we present the Kantorovich dual form, which clarifies what is required of the ambient space for the infimum in the Kantorovich problem to be attained. First, we define Polish spaces:

Definition 19 (Polish space)**.** A topological space (X,τ) is called *Polish* if there exists a metric $d : X \times X \to [0,\infty)$ such that:

1. d generates the topology τ, i.e. $\tau = \tau_d$, where τ_d is the metric topology induced by d;
2. (X,d) is *complete*, i.e. every d-Cauchy sequence in X converges (with respect to d) to a point in X;
3. (X,d) is *separable*, i.e. there exists a countable dense subset $D \subset X$ such that $\overline{D} = X$.

Equivalently, a metric space (X,d) is Polish if it is complete and separable.

Theorem 9 (Kantorovich duality on Polish spaces)**.** *Let (X,d_X) and (Y,d_Y) be Polish spaces, and let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$. Assume $c : X \times Y \to (-\infty,+\infty]$ is lower semicontinuous and bounded from below, and that there exists at least one coupling $\pi_0 \in \Pi(\mu,\nu)$ with $\int c\, d\pi_0 < \infty$. Then*

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\, d\pi(x,y) \;=\; \sup\left\{ \int_X \varphi\, d\mu + \int_Y \psi\, d\nu \;:\; \varphi(x)+\psi(y) \le c(x,y)\ \forall (x,y) \right\},$$

where the supremum is taken over all Borel functions $\varphi : X \to \mathbb{R}\cup\{-\infty\}$ and $\psi : Y \to \mathbb{R}\cup\{-\infty\}$ such that $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$. Moreover, the infimum is attained by some optimal plan $\pi^\star \in \Pi(\mu,\nu)$.

This duality theorem can be found in Villani [50], Chapter 5, and the attainment of the infimum on Polish spaces in Villani [50], Theorem 4.1. For Theorem 2, attainment of the infimum holds on both function spaces used in our experiments: a separable Hilbert space with its norm topology is Polish, and so is path space with its product topology [5, Lemma 1.4.13].

The theory of entropy-regularized static OT can be found in Peyré and Cuturi [39], Chapter 4, and its connection with statistical divergences in Chapter 8. We state only the definitions needed for the Hilbert Sinkhorn Divergence of the main text.

The entropy-regularized problem adds to the transport cost a multiple of the Kullback–Leibler (KL) divergence of the plan from the product of its marginals.

Definition 20 (Kullback–Leibler (KL) divergence)**.** Let $(X,\mathcal{F})$ be a measurable space and let P, Q be probability measures on it. The *KL divergence* of P from Q is defined by

$$\mathrm{KL}(P\|Q) := \begin{cases} \displaystyle\int_X \log\!\left(\frac{dP}{dQ}\right) dP, & \text{if } P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

Equivalently, if $P \ll Q$ then

$$\mathrm{KL}(P\|Q) = \int_X \frac{dP}{dQ}\, \log\!\left(\frac{dP}{dQ}\right) dQ,$$

where $\frac{dP}{dQ}$ denotes the Radon–Nikodym derivative (defined Q-a.e.).

Definition 21 (Entropic OT Functional)**.**

$$\mathsf{OT}_{c,\varepsilon}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{X\times Y} c(x,y)\, d\pi(x,y) + \varepsilon\, \mathrm{KL}(\pi \,\|\, \mu\otimes\nu) \right\}$$

The Sinkhorn divergence is the debiased version of the entropic OT functional:

Definition 22 (Sinkhorn-Divergence)**.**

$$\mathsf{S}_{c,\varepsilon}(\mu,\nu) := \mathsf{OT}_{c,\varepsilon}(\mu,\nu) - \frac{1}{2}\,\mathsf{OT}_{c,\varepsilon}(\mu,\mu) - \frac{1}{2}\,\mathsf{OT}_{c,\varepsilon}(\nu,\nu).$$

This is the entropic OT functional used in the proof of Theorem 2; it is well defined in this generality because our ambient spaces are Polish. The Hilbert Sinkhorn divergence in the main text and in Li et al. [31] is obtained by applying the same self-cost debiasing after restricting the ambient space to an RKHS with its norm-induced topology.

B Technical Proofs

B.1 Inherited Results from Li et al. (2021)

For transparency, we separate prior results from our original proofs. Proposition 3 (the equivalence behind Equation (4)) and the two propositions below are all inherited from Li et al. [31]. We restate them here because they are standard ingredients used by our function-space analysis, but we do not claim them as new contributions. Their proofs are therefore referenced directly to the original source.

Proposition 3 (Li et al. [31])**.** *The entropic OT functional underlying Definition 1 can be equivalently formulated as in Equation* (4)*, in the sense that if π^* minimizes Equation* (4)*, then its pushforward $(\phi \otimes \phi)_{\#}\pi^*$ minimizes Equation* (3)*; applying the same self-cost subtraction on both sides yields $S_{\kappa,\epsilon}(\mu,\nu) = S_\epsilon(\phi_{\#}\mu, \phi_{\#}\nu)$.*

Proof: The proof is given as Theorem 1 in the supplementary material of Li et al. [31] [2].

Proposition 4. *Let μ_n, ν_n be the empirical estimators in Equation* (5) *and $\epsilon, \eta > 0$. Then there exists $N > 0$ such that*

$$\forall n \geq N, \qquad \mathbb{P}\left(|S_\epsilon(\mu_n,\nu_n) - S_\epsilon(\mu,\nu)| \leq \epsilon\eta\right) = 1.$$

Proof: The proof is given as Theorem 2 in the supplementary material of Li et al. [31], same link as before [2].

Proposition 5. *Let $||f||_{\mathcal{H}} \leq M$ and consider $B_\eta = \{f \in \mathcal{H} : ||f||_{\mathcal{H}} < \eta\}$. Let m be the number of basis spanning functions f in B_η. Then for $M, \eta, \delta, \epsilon > 0$,*

$$\mathbb{P}\Big(|S_\epsilon(\mu_n,\nu_n) - S_\epsilon(\mu,\nu)| \leq \epsilon\eta\Big) \geq 1 - \delta,$$

whenever

$$n \geq \frac{1}{\eta^2}\Big(2M^2(\log(2/\delta) + m\log(24M/\eta))\Big).$$

Proof: The proof is given as Proposition 3 in the supplementary material of Li et al. [31], same link as before [2].

B.2 Absolute Continuity Behind the CFM/FM Equivalence

The equivalence between the conditional and unconditional flow-matching losses in Section 2.1 rests on an absolute-continuity assumption that is worth making explicit in infinite dimensions. Two FFM conditionals $\mathcal{N}(tf, \sigma_t^2 C_0)$ and $\mathcal{N}(t'f', \sigma_{t'}^2 C_0)$ are typically mutually singular (Feldman–Hájek), so the relevant structure is not conditional-versus-conditional but *conditional-versus-mixture at each fixed t*: one needs $\mu_t^f \ll \mu_t$ for ν-a.e. f. A sufficient condition, used implicitly by FFM [25], is that the data measure ν is supported in the Cameron–Martin space $H(C_0) = \mathrm{Ran}(C_0^{1/2})$ of the prior, together with σ_t bounded below ($\sigma_{\min} > 0$, which Algorithm 1 enforces). Then, for fixed t, all conditionals $\{\mathcal{N}(tf, \sigma_t^2 C_0) : f \in \operatorname{supp}\nu\}$ are mutually equivalent by the Cameron–Martin formula, hence absolutely continuous with respect to the mixture μ_t, and the exchange of marginal and conditional velocities is justified at each t. When the data are rougher than the prior ($\operatorname{supp}\nu \not\subset H(C_0)$) the equivalence degrades, so the prior roughness should be matched to the data; this is one reason the white-noise base measure is selected on the Navier–Stokes family (Appendix C.5).

How kFFM inherits this structure. At the population level, write $G_t(f_0, f_1) := tf_1 + \sigma_t f_0$. The kFFM path marginal is $\mu_t^\pi = (G_t)_{\#}\pi$, the FFM marginal is $\mu_t = (G_t)_{\#}(\mu_0 \otimes \nu)$, and the kFFM conditional given f_1 is $\mu_t^{f_1,\pi} = G_t(\cdot, f_1)_{\#}\,\pi(\cdot \mid f_1)$, while the FFM conditional is $\mathcal{N}(tf_1, \sigma_t^2 C_0) = G_t(\cdot, f_1)_{\#}\,\mu_0$. For the entropic problem with a bounded cost (Theorem 1) the optimal plan has the Gibbs form $d\pi/d(\mu_0 \otimes \nu) = \exp\{(\varphi \oplus \psi - c_\kappa)/\varepsilon\}$ with bounded potentials, so its density is bounded above and away from zero. Hence $\pi \sim \mu_0 \otimes \nu$ and $\pi(\cdot \mid f_1) \sim \mu_0$ for

[2] https://openaccess.thecvf.com/content/CVPR2021/supplemental/Li_Hilbert_Sinkhorn_Divergence_CVPR_2021_supplemental.pdf

ν-a.e. f_1; since equivalence of measures is preserved under pushforward, $\mu_t^{f_1,\pi} \sim \mathcal{N}(tf_1, \sigma_t^2 C_0)$ and $\mu_t^\pi \sim \mu_t$, and every absolute-continuity relation that FFM relies on transfers verbatim to kFFM. Two remarks. First, this is exactly where entropic regularization matters: an unregularized OT plan can be singular with respect to the product measure, and the argument would fail. Second, the minibatch plans used in training are discrete surrogates of this population object (Section 3.1); the statement concerns the coupling they approximate.

B.3 Proof of Well-posedness and Uniform Boundedness (Theorem 1)

Theorem 1 (Well-posedness and uniform boundedness)**.** *Let $(\mathcal{X}, \|\cdot\|_\mathcal{X})$ be a real Banach space and let $d_\mathcal{X}(f,g) := \|f-g\|_\mathcal{X}$. Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be measurable, symmetric, and positive definite with RKHS $\mathcal{H}_\kappa$. Assume additionally that κ is bounded: $B_\kappa := \sup_{f\in\mathcal{X}} \kappa(f,f) < \infty$. Then the kernel cost satisfies $0 \le c_\kappa(f,g) \le 4B_\kappa$ for all $f, g \in \mathcal{X}$, and for every $\varepsilon > 0$ the Hilbert Sinkhorn divergence $S_{\kappa,\varepsilon}(\mu,\nu)$ (Definition 1) is finite for all $\mu, \nu \in \mathcal{P}(\mathcal{X})$, with $|S_{\kappa,\varepsilon}(\mu,\nu)| \le 8B_\kappa$.*

Proof: The argument is elementary once one uses the boundedness of the RKHS cost induced by a bounded kernel. We start by writing the Hilbert Sinkhorn divergence in the general form

$$S_{\kappa,\varepsilon}(\mu,\nu) := \mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) - \tfrac{1}{2}\mathsf{OT}_{\kappa,\varepsilon}(\mu,\mu) - \tfrac{1}{2}\mathsf{OT}_{\kappa,\varepsilon}(\nu,\nu).$$

By the reproducing property, the canonical feature map ϕ of Definition 1 satisfies

$$\|\phi(f)\|^2_{\mathcal{H}_\kappa} = \langle \kappa(\cdot,f), \kappa(\cdot,f)\rangle_{\mathcal{H}_\kappa} = \kappa(f,f) \le B_\kappa \qquad \text{for all } f \in \mathcal{X}.$$

Therefore, for any $f, g \in \mathcal{X}$,

$$c_\kappa(f,g) = \|\phi(f) - \phi(g)\|^2_{\mathcal{H}_\kappa} \le 2\|\phi(f)\|^2_{\mathcal{H}_\kappa} + 2\|\phi(g)\|^2_{\mathcal{H}_\kappa} \le 4B_\kappa.$$

Fix $\varepsilon > 0$. Since $c_\kappa \ge 0$ and $\mathrm{KL} \ge 0$, we have $\mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) \ge 0$. To upper bound $\mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu)$, use the feasible coupling $\pi_0 := \mu \otimes \nu \in \Pi(\mu,\nu)$. Then $\mathrm{KL}(\pi_0 \| \mu\otimes\nu) = 0$ and hence

$$\mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) \le \int_{\mathcal{X}\times\mathcal{X}} c_\kappa(f,g)\, d(\mu\otimes\nu)(f,g) \le 4B_\kappa.$$

Since $\mu\otimes\nu$ is a probability measure, it follows that $0 \le \mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) \le 4B_\kappa < \infty$. The same argument applies to $\mathsf{OT}_{\kappa,\varepsilon}(\mu,\mu)$ and $\mathsf{OT}_{\kappa,\varepsilon}(\nu,\nu)$. Since each of the three terms in the definition of $S_{\kappa,\varepsilon}(\mu,\nu)$ lies in $[0, 4B_\kappa]$, the expression is finite. Moreover,

$$\begin{aligned} |S_{\kappa,\varepsilon}(\mu,\nu)| &\le \mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) + \tfrac{1}{2}\mathsf{OT}_{\kappa,\varepsilon}(\mu,\mu) + \tfrac{1}{2}\mathsf{OT}_{\kappa,\varepsilon}(\nu,\nu) \\ &\le 4B_\kappa + \tfrac{1}{2}(4B_\kappa) + \tfrac{1}{2}(4B_\kappa) = 8B_\kappa. \qquad \square \end{aligned}$$

Interpretation. Theorem 1 is the well-posedness step needed in the function-space setting: it guarantees that the kernel cost is uniformly bounded by $4B_\kappa$ and that the HSD objective is finite for all probability measures on the ambient Banach space whenever the kernel is bounded, with no moment assumption on the measures. The uniform bound is what conditions the entropic solver (Section 3.2). This is precisely the hypothesis that is left implicit in the inherited equivalence proposition from Li et al. [31].

B.4 Proof of the Error Decomposition (Theorem 2)

Theorem 2 (Error decomposition)**.** *Let $(\mathcal{X}, d)$ be a compact metric space and let $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Define the target quadratic cost $c(x,y) = d(x,y)^2$. Let $\kappa : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a bounded positive-definite kernel with RKHS-induced cost $c_\kappa(x,y) := \kappa(x,x) + \kappa(y,y) - 2\kappa(x,y)$, and let $\Delta_\kappa := \sup_{x,y\in\mathcal{X}} |c_\kappa(x,y) - c(x,y)|$. Let $\mathsf{OT}_{\kappa,\varepsilon}$ denote the entropic OT functional in Equation* (4) *with cost c_κ, and let $M := \mathrm{diam}(\mathcal{X}) < \infty$. Then for every $\varepsilon > 0$ and every $\delta \in (0, M]$:*

$$|\mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) - \mathcal{W}(\mu,\nu)| \le \Delta_\kappa + 2\varepsilon \log \mathcal{N}(\mathcal{X},\delta;d) + 8M\delta, \tag{6}$$

where $\mathcal{N}(\mathcal{X},\delta;d)$ is the δ-covering number of $\mathcal{X}$ in the metric d.

Proof. Let $c(x,y) = d(x,y)^2$ and let c_κ be the RKHS-induced cost in Theorem 2. By definition of Δ_κ, for every coupling $\pi \in \Pi(\mu,\nu)$,

$$\left| \int c_\kappa \, d\pi - \int c \, d\pi \right| \le \Delta_\kappa.$$

Therefore the unregularized OT values for the two costs differ by at most Δ_κ.

For the lower bound, non-negativity of the KL term gives

$$\mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) \ge \inf_{\pi \in \Pi(\mu,\nu)} \int c_\kappa \, d\pi \ge \mathcal{W}(\mu,\nu) - \Delta_\kappa.$$

For the upper bound, let $m := \mathcal{N}(\mathcal{X}, \delta; d)$ and choose centers $x_1, \dots, x_m$ whose closed δ-balls cover $\mathcal{X}$. A measurable quantization map $Q : \mathcal{X} \to \{x_1, \dots, x_m\}$ with $d(x, Q(x)) \le \delta$ is obtained by assigning each point to the first center whose ball contains it, giving a finite Borel partition $C_i = Q^{-1}(\{x_i\})$. Since $\mathcal{X}$ is compact and $c = d^2$ is continuous, an optimal coupling $\pi^\star$ for $\mathcal{W}(\mu,\nu)$ exists.

Set $p_{ij} := \pi^\star(C_i \times C_j)$, $a_i := \mu(C_i)$, and $b_j := \nu(C_j)$. For cells with positive mass, let μ_i and ν_j be the conditional restrictions of μ and ν to C_i and C_j, respectively; zero-mass cells may be assigned arbitrary probability measures on $\mathcal{X}$, since their weights vanish. Define

$$\widehat{\pi} := \sum_{i,j} p_{ij} \, \mu_i \otimes \nu_j.$$

This construction preserves the marginals, so $\widehat{\pi} \in \Pi(\mu,\nu)$. On $C_i \times C_j$ with $a_i b_j > 0$, the density of $\widehat{\pi}$ with respect to $\mu \otimes \nu$ is $p_{ij}/(a_i b_j)$, and cells with $a_i b_j = 0$ have $p_{ij} = 0$. Hence

$$\mathrm{KL}(\widehat{\pi} \| \mu \otimes \nu) = \mathrm{KL}(p \| a \otimes b) \le 2 \log m,$$

where the final inequality follows from $\mathrm{KL}(p \| a \otimes b) = H(a) + H(b) - H(p) \le H(a) + H(b) \le 2 \log m$. Moreover, if $x \in C_i$ and $y \in C_j$, then

$$|d(x,y)^2 - d(x_i,x_j)^2| \le (d(x,y) + d(x_i,x_j)) \, |d(x,y) - d(x_i,x_j)| \le 4M\delta.$$

Applying this once under $\widehat{\pi}$ and once under $\pi^\star$ gives

$$\int c \, d\widehat{\pi} \le \mathcal{W}(\mu,\nu) + 8M\delta.$$

Using again $\int c_\kappa \, d\widehat{\pi} \le \int c \, d\widehat{\pi} + \Delta_\kappa$, feasibility of $\widehat{\pi}$ for the entropic problem yields

$$\mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) \le \mathcal{W}(\mu,\nu) + \Delta_\kappa + 8M\delta + 2\varepsilon \log m.$$

Combining the lower and upper bounds and substituting $m = \mathcal{N}(\mathcal{X}, \delta; d)$ proves the result. □

Euclidean covering-rate special case. If $\mathcal{X} \subset \mathbb{R}^d$ has bounded radius M, then its covering number satisfies $\log \mathcal{N}(\mathcal{X}, \delta) \le C d \log(M/\delta)$ for a universal constant $C > 0$. Substituting this into Theorem 2 gives

$$|\mathsf{OT}_{\kappa,\varepsilon}(\mu,\nu) - \mathcal{W}(\mu,\nu)| \le \Delta_\kappa + C\big(\varepsilon d \log(M/\delta) + M\delta\big),$$

after absorbing constants. When $c_\kappa = c$, this recovers the usual entropic-regularization rate up to constants.

B.5 Proof of Discretization Invariance (Theorem 3)

Theorem 3 (Discretization invariance)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded periodic domain and let $\alpha > k \ge 0$. Set $\mathcal{X} := B_R^{H^\alpha} := \{f \in H^\alpha(\Omega) : \|f\|_{H^\alpha} \le R\}$, viewed as a compact subset of $H^k(\Omega)$, and let $\mu, \nu \in \mathcal{P}(\mathcal{X})$. For the Fourier basis $\{e_n\}_{n \in \mathbb{Z}^d}$, let $V_N := \mathrm{span}\{e_n : |n|^2 \le N^{2/d}\}$ and let $P_N : H^\alpha(\Omega) \to V_N$ denote the orthogonal projection; hence $\dim V_N \asymp N$. Write $\mu_N := (P_N)_\# \mu$ and $\nu_N := (P_N)_\# \nu$. Define the Sobolev-RBF kernel $\kappa^{(k)}_{\mathrm{RBF}}(f,g) := \exp(-\|f-g\|^2_{H^k}/2\sigma^2)$ on H^k, and write $\mathsf{KOT}^{(k)}_\varepsilon := \mathsf{OT}_{\kappa^{(k)}_{\mathrm{RBF}},\varepsilon}$ for the corresponding entropic kernel-OT functional from Theorem 2. Then for every $\varepsilon > 0$,*

$$\left|\mathsf{KOT}^{(k)}_\varepsilon(\mu,\nu) - \mathsf{KOT}^{(k)}_\varepsilon(\mu_N,\nu_N)\right| \;\le\; \frac{4R^2}{\sigma^2} \, N^{-2(\alpha-k)/d}.$$

Equivalently, in terms of grid spacing $h = N^{-1/d}$, the discretization error decays as $O(h^{2(\alpha-k)})$.

Notation. Throughout this proof, $\Omega \subset \mathbb{R}^d$ is the bounded periodic domain, $\{e_n\}_{n\in\mathbb{Z}^d}$ is the Fourier basis on Ω, and the Sobolev norm is

$$\|f\|_{H^s}^2 = \sum_{n\in\mathbb{Z}^d} (1+|n|^2)^s \, |\hat{f}_n|^2, \qquad s \geq 0.$$

P_N denotes the orthogonal projection onto $V_N := \operatorname{span}\{e_n : |n|^2 \leq N^{2/d}\}$, so $\dim V_N \asymp N$ and the highest retained mode has frequency $\asymp N^{1/d}$. We write $\mathsf{KOT}_\varepsilon^{(k)} := \mathsf{OT}_{\kappa_{\mathrm{RBF}}^{(k)},\varepsilon}$ and

$$c_{\mathrm{RBF}}^{(k)}(f,g) := \|\phi(f)-\phi(g)\|^2_{\mathcal{H}_{\kappa_{\mathrm{RBF}}^{(k)}}} = 2\big(1-\kappa_{\mathrm{RBF}}^{(k)}(f,g)\big)$$

for the RKHS-induced cost associated with the RBF kernel on H^k. Since $B_R^{H^\alpha}$ is compact in H^k by the compact Sobolev embedding on bounded periodic domains [12] and $c_{\mathrm{RBF}}^{(k)}$ is bounded and continuous, the entropic kernel-OT functionals below attain minimizers.

Step 1: Sobolev approximation on the truncation tail. For any $f \in H^\alpha(\Omega)$ with $\|f\|_{H^\alpha} \leq R$ and any $0 \leq k < \alpha$,

$$\begin{aligned}
\|f - P_N f\|_{H^k}^2 &= \sum_{|n|^2 > N^{2/d}} (1+|n|^2)^k \, |\hat{f}_n|^2 \\
&= \sum_{|n|^2 > N^{2/d}} (1+|n|^2)^{-(\alpha-k)} \, (1+|n|^2)^\alpha \, |\hat{f}_n|^2 \\
&\leq (1+N^{2/d})^{-(\alpha-k)} \, \|f\|_{H^\alpha}^2 \;\leq\; R^2 \, N^{-2(\alpha-k)/d}.
\end{aligned}$$

Hence

$$\|f - P_N f\|_{H^k} \;\leq\; R\, N^{-(\alpha-k)/d} \qquad \text{for all } f \in B_R^{H^\alpha}. \tag{7}$$

Step 2: Stability of the squared H^k distance under simultaneous projection. For $f, g \in B_R^{H^\alpha}$, since P_N is an orthogonal projection on H^k,

$$\begin{aligned}
\|f-g\|_{H^k}^2 - \|P_N(f-g)\|_{H^k}^2 &= \|(I-P_N)(f-g)\|_{H^k}^2 \\
&\leq \|f-g\|_{H^\alpha}^2 \cdot N^{-2(\alpha-k)/d} \\
&\leq 4R^2 \, N^{-2(\alpha-k)/d},
\end{aligned}$$

where we used $\|f-g\|_{H^\alpha} \leq 2R$ and the Sobolev tail estimate from Step 1 applied to $f-g$.

Step 3: Cost stability for the RBF kernel. Let $h(s) := e^{-s/(2\sigma^2)}$ for $s \geq 0$, so $\kappa_{\mathrm{RBF}}^{(k)}(f,g) = h(\|f-g\|_{H^k}^2)$ and $|h'(s)| \leq \frac{1}{2\sigma^2}$. Then

$$\begin{aligned}
\big|c_{\mathrm{RBF}}^{(k)}(f,g) - c_{\mathrm{RBF}}^{(k)}(P_N f, P_N g)\big| &= 2\big|h(\|f-g\|_{H^k}^2) - h(\|P_N(f-g)\|_{H^k}^2)\big| \\
&\leq \frac{1}{\sigma^2}\big|\|f-g\|_{H^k}^2 - \|P_N(f-g)\|_{H^k}^2\big| \;\leq\; \frac{4R^2}{\sigma^2}\, N^{-2(\alpha-k)/d}.
\end{aligned}$$

The bound is uniform over $(f,g) \in B_R^{H^\alpha} \times B_R^{H^\alpha}$.

Step 4: Upper bound. Let $\pi^\star \in \Pi(\mu,\nu)$ be a feasible coupling for the continuum problem and define $\widetilde{\pi}_N := (P_N \otimes P_N)_\# \pi^\star \in \Pi(\mu_N, \nu_N)$. By Step 3,

$$\int_{V_N \times V_N} c_{\mathrm{RBF}}^{(k)} \, d\widetilde{\pi}_N \;=\; \int_{\mathcal{X}\times\mathcal{X}} c_{\mathrm{RBF}}^{(k)}(P_N f, P_N g)\, d\pi^\star(f,g) \;\leq\; \int c_{\mathrm{RBF}}^{(k)}\, d\pi^\star + \frac{4R^2}{\sigma^2}\, N^{-2(\alpha-k)/d}.$$

By the data-processing inequality for KL divergence applied to the measurable map $P_N \otimes P_N$,

$$\mathrm{KL}(\widetilde{\pi}_N \| \mu_N \otimes \nu_N) \;\leq\; \mathrm{KL}(\pi^\star \| \mu \otimes \nu).$$

Taking $\pi^\star$ to be an optimizer of $\mathsf{KOT}_\varepsilon^{(k)}(\mu,\nu)$ and using $\widetilde{\pi}_N$ as a feasible point for $\mathsf{KOT}_\varepsilon^{(k)}(\mu_N,\nu_N)$ gives

$$\mathsf{KOT}_\varepsilon^{(k)}(\mu_N,\nu_N) \;\leq\; \mathsf{KOT}_\varepsilon^{(k)}(\mu,\nu) + \frac{4R^2}{\sigma^2}\, N^{-2(\alpha-k)/d}. \tag{8}$$

Step 5: Lower bound via gluing. Let $\pi_N^\star \in \Pi(\mu_N, \nu_N)$ be an optimizer for $\mathsf{KOT}_\varepsilon^{(k)}(\mu_N, \nu_N)$. Disintegrate μ along P_N: since compact metric spaces are standard Borel, the standard regular conditional probability theorem applies [10]. Thus there exist Borel kernels $\mu(\cdot \,|\, u)$, $u \in V_N$, such that

$$\mu(A) = \int_{V_N} \mu(A \,|\, u)\, d\mu_N(u), \qquad A \subset \mathcal{X} \text{ Borel},$$

with $\mu(\cdot \,|\, u)$ supported on the fibre $P_N^{-1}(u)$ for μ_N-a.e. u, and similarly for $\nu(\cdot \,|\, v)$. Define the lifted coupling

$$\widehat{\pi}(A \times B) := \int_{V_N \times V_N} \mu(A \,|\, u)\, \nu(B \,|\, v)\, d\pi_N^\star(u, v), \qquad A, B \subset \mathcal{X} \text{ Borel}.$$

Direct verification gives $\widehat{\pi} \in \Pi(\mu, \nu)$ and $(P_N \otimes P_N)_\# \widehat{\pi} = \pi_N^\star$. Since $\mu_N \otimes \nu_N$ is feasible with finite KL, the optimizer $\pi_N^\star$ has finite KL and hence $\pi_N^\star \ll \mu_N \otimes \nu_N$. On each fibre product, the Radon–Nikodym derivative of $\widehat{\pi}$ with respect to $\mu \otimes \nu$ is exactly the derivative of $\pi_N^\star$ with respect to $\mu_N \otimes \nu_N$ evaluated at $(u, v) = (P_N f, P_N g)$. Therefore, by the chain rule for KL divergence,

$$\mathrm{KL}(\widehat{\pi} \| \mu \otimes \nu) = \mathrm{KL}(\pi_N^\star \| \mu_N \otimes \nu_N).$$

For the cost integral, Step 3 applied in reverse gives

$$\int c_{\mathrm{RBF}}^{(k)}\, d\widehat{\pi} \;\le\; \int_{V_N \times V_N} c_{\mathrm{RBF}}^{(k)}\, d\pi_N^\star + \frac{4R^2}{\sigma^2}\, N^{-2(\alpha-k)/d},$$

using that $(f, g) \sim \widehat{\pi}$ projects to $(u, v) \sim \pi_N^\star$ and that the cost difference is bounded uniformly. Hence

$$\mathsf{KOT}_\varepsilon^{(k)}(\mu, \nu) \;\le\; \mathsf{KOT}_\varepsilon^{(k)}(\mu_N, \nu_N) + \frac{4R^2}{\sigma^2}\, N^{-2(\alpha-k)/d}. \tag{9}$$

Step 6: Conclusion. Combining (8) and (9) yields

$$\left|\mathsf{KOT}_\varepsilon^{(k)}(\mu, \nu) - \mathsf{KOT}_\varepsilon^{(k)}(\mu_N, \nu_N)\right| \;\le\; \frac{4R^2}{\sigma^2}\, N^{-2(\alpha-k)/d}. \qquad \square$$

B.6 Composed End-to-End Bound

Corollary 1 (Composed end-to-end bound)**.** *Let $\mathcal{X} = B_R^{H^\alpha}$ be as in Theorem 3, equipped with $d_k(f, g) := \|f - g\|_{H^k}$, and let*

$$\mathcal{W}_k(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \|f - g\|_{H^k}^2\, d\pi(f, g)$$

denote quadratic-cost OT in the H^k metric. Write $M_k := \mathrm{diam}(\mathcal{X}; d_k)$ and let $\mathcal{N}_k(\delta) := \mathcal{N}(\mathcal{X}, \delta; d_k)$ be the corresponding covering number. For the Sobolev-RBF kernel in Theorem 3, define the mismatch $\Delta_{\mathrm{RBF}} := \Delta_{\kappa_{\mathrm{RBF}}^{(k)}}$, i.e.,

$$\Delta_{\mathrm{RBF}} := \sup_{f, g \in \mathcal{X}} \left| 2\left(1 - e^{-\|f-g\|_{H^k}^2/(2\sigma^2)}\right) - \|f - g\|_{H^k}^2 \right| \;\le\; M_k^2 + 2,$$

where the bound follows because the RBF-induced cost lies in $[0, 2]$ and $\|f - g\|_{H^k}^2 \in [0, M_k^2]$ on $\mathcal{X}$. Then for every $\varepsilon > 0$, $\delta \in (0, M_k]$ and every projection rank $N \in \mathbb{N}$,

$$\begin{aligned}\left|\mathsf{KOT}_\varepsilon^{(k)}(\mu_N, \nu_N) - \mathcal{W}_k(\mu, \nu)\right| \;\le\; & \underbrace{\Delta_{\mathrm{RBF}}}_{\text{kernel-cost mismatch}} + \underbrace{2\varepsilon\, \log \mathcal{N}_k(\delta)}_{\text{entropic regularization}} \\ & + \underbrace{8 M_k \delta}_{\text{covering}} + \underbrace{\frac{4R^2}{\sigma^2}\, N^{-2(\alpha-k)/d}}_{\text{discretization}}.\end{aligned}$$

Proof. By the triangle inequality, $|\mathsf{KOT}_\varepsilon^{(k)}(\mu_N, \nu_N) - \mathcal{W}_k(\mu, \nu)| \le |\mathsf{KOT}_\varepsilon^{(k)}(\mu_N, \nu_N) - \mathsf{KOT}_\varepsilon^{(k)}(\mu, \nu)| + |\mathsf{KOT}_\varepsilon^{(k)}(\mu, \nu) - \mathcal{W}_k(\mu, \nu)|$. The first term is bounded by Theorem 3; the second by Theorem 2 applied on the compact metric space $(\mathcal{X}, d_k)$ with the Sobolev-RBF kernel, whose mismatch is Δ_{RBF}. $\square$

Remark. Of the four terms, only the discretization term vanishes with projection rank N (rate $N^{-2(\alpha-k)/d}$, driven by the regularity gap $\alpha - k$), the formal counterpart of the empirical resolution-invariance in Section 4. The corollary relies on compact Sobolev embedding [12], and our PDE benchmarks are used in the corresponding smooth-solution regime; such Sobolev and analytic regularity is standard in the PDE settings considered here, including Navier–Stokes in appropriate regimes [4]. The kernel-cost mismatch Δ_{RBF}, by contrast, depends only on kernel design: zero for the linear kernel, finite for RBF (controlled by data scale and bandwidth σ).

Remark (path space). For data on path space $C_p([0,T],\mathbb{R}^d)$ with $p \in [1,2)$, the analogue of spectral truncation is signature truncation at level L, $S_{\le L}(x) := (1, S^{(1)}(x), \dots, S^{(L)}(x))$. Known factorial-decay bounds for signatures of finite-p-variation paths [5, Proposition 1.4.6] imply that, on compact sets with uniformly bounded p-variation control, the tail $\sum_{\ell > L} \|S^{(\ell)}(x)\|^2$ decays faster than any polynomial in L. By Cauchy–Schwarz, the full and truncated signature kernels are then uniformly close on such compact sets. Consequently, the corresponding entropic kernel-OT objectives differ by at most the uniform signature-kernel cost error, since the KL term is unchanged. This gives the same qualitative discretization-invariance message for path space, but with a factorial tail in the signature truncation level rather than a Sobolev Fourier-tail rate.

C Additional Experimental Details

This appendix provides implementation details and additional experimental results that complement the main text.

C.1 Model Parameterization

We implement Functional Flow Matching with the official implementation provided by its authors[3] without modifying any provided hyperparameter of the neural architecture, including $\sigma_{\min} = 10^{-4}$. The baseline models use the following hyperparameters, also following Kerrigan et al. [25]:

1. DDPM: the noise schedule has $\beta_0 = 10^{-4}$, $\beta_T = 0.02$, and $T = 1000$.
2. DDO (NCSN): the time interval is set to $T = 10$, and the noise schedule interpolates geometrically between $\sigma_{10} = 10^{-3}$ and $\sigma_1 = 1$ on the 1D datasets and between $\sigma_{10} = 10^{-2}$ and $\sigma_1 = 100$ on the 2D datasets.
3. GANO: the generator is trained every 5 epochs, and the gradient penalty is $\lambda = 0.1$ on the 1D and $\lambda = 10$ on the 2D datasets.

Training. The baseline models are trained with Adam, with learning rate 10^{-3} on 1D and 5×10^{-4} on 2D data, except that the rate is 10^{-4} for GANO. This configuration also follows Kerrigan et al. [25]; no hyperparameter search was performed on the optimizer.

C.2 Baseline Description

We describe the baseline generative models used for the empirical comparisons. All operate in function space, and the roster covers the diffusion, flow-matching, adversarial, and score-based paradigms; every baseline uses the same neural-operator backbone and data pipeline as kFFM.

1. **DDPM [24]**: The DDPM used as a baseline is the functional DDPM model proposed by Kerrigan et al. [24], which generalizes the diffusion model in $\mathbb{R}^d$ using Gaussian measure theory in Hilbert space. Its formulation is analogous to the base DDPM but performs diffusion directly on Gaussian measures.
2. **FFM [25]**: Functional Flow Matching is the base model that we extend. It generalizes flow matching in $\mathbb{R}^d$ to infinite dimensions by formulating a path of Gaussian measures in a Hilbert space, in analogy with Kerrigan et al. [24]. In the terminology of Tong et al. [49], it couples prior and data samples independently, which leaves room for OT-based coupling.

[3] `https://github.com/GavinKerrigan/functional_flow_matching`

3. **CFM-OT(L^2) [49]**: This is the direct finite-dimensional OT baseline used in the main text. It uses the same minibatch Sinkhorn solver and FNO backbone as our method, but applies OT to discretized vectors with the raw L^2 cost and the white-noise (i.i.d. Gaussian) base measure, rather than a function-space kernel cost.
4. **GANO [41]**: The Generative Adversarial Neural Operator (GANO) is a generative adversarial network in function space whose generator maps Gaussian random fields to samples.
5. **DDO/NCSN [34]**: The Denoising Diffusion Operator (DDO) is the analogue of score-based generative models [46] in infinite dimensions. We follow Kerrigan et al. [25] and use the NCSN noise scale; the loss and noise schedule, including preconditioning, follow the code released with Lim et al. [34].

C.3 Dataset Description

Of all the datasets we experiment on, AEMET, Gene expression, Economics are from Kerrigan et al. [25], Navier–Stokes, KdV, stochastic Navier–Stokes, stochastic KdV are from Salvi et al. [44], and Heston is from Issa et al. [23]. The descriptions below follow those sources; the architectures and training budgets of our runs are given in Appendix C.4.

AEMET Dataset: This dataset comprises functional observations representing the mean annual profile of average daily temperature (°C) over 1980–2009, collected from 73 weather stations in Spain. Each curve is sampled on a uniform grid of length 365.

Gene Expression: The full dataset contains 10,928 time series sampled at 20 evenly spaced time points, measuring gene-expression amplitudes for four genes. To produce a periodic spiking appearance while preserving the underlying structure, we concatenate the gene-specific sequences in time. Before training, the data are log-transformed and mean-centered. We then focus on 156 high-variability trajectories, selected by requiring the time-averaged standard deviation of each centered series to exceed 0.3.

Economic Dataset: We use the three economic datasets (population, GDP, labor) of Kerrigan et al. [25]; a separate model is trained on each, and the reported Economy metrics are averaged over the component series within each seed (the matched-seed comparisons of Tables 1 and 2 use the population and GDP series). As an example, the population dataset consists of time series tracking population changes for 169 countries worldwide from 1950 to 2018, sampled at 69 discrete time points. For visualization, each country's series is normalized by its own mean, so the curves show population relative to that country's average over the 69-year period. Overall, the trajectories display approximately linear growth with a common change point shared across countries.

Heston Dataset: The Heston model [21] is a stochastic volatility model with diffusion-driven, non-smooth sample paths. It models the asset price S_t and its variance ν_t as (in the standard Heston notation, which is local to this paragraph):

$$dS_t = \mu_t S_t dt + \sqrt{\nu_t} S_t dW_t^S$$
$$d\nu_t = \kappa(\theta - \nu_t)dt + \xi\sqrt{\nu_t} dW_t^\nu,$$

where W_t^S, W_t^ν are Wiener processes with correlations.

The Heston dataset is generated using the simulation code provided in the official repository of `sigker-nsde` [23] [4]. The script generates 5000 univariate time series of length 100 on $[0, 1]$; we model the normalized log-variance paths. A long-horizon variant with 1000 time steps (Heston-Long) is used only in the sensitivity study (the "long volatility paths" of Appendices C.8 and C.16) and in the convergence diagnostic of Figure 6.

KdV and Stochastic KdV Equation: The stochastic Korteweg–de Vries equation is a higher-order SPDE given by:

$$\partial_t u + \gamma\partial_x^3 u = 6u\partial_x u + \xi, \quad u(t,0) = u(t,1), \quad u(0,x) = u_0(x), \quad (t,x) \in [0,T] \times [0,1],$$

[4] `https://github.com/issaz/sigker-nsdes`

and the base KdV equation simply removes the Q-Wiener process ξ. The KdV dataset is a single trajectory on a grid of 512 spatial points with 201 time steps, and each time slice is one sample; the stochastic KdV dataset consists of 1200 trajectories on a grid of 128 spatial points with 101 time steps each, again treated as snapshots. The simulation code for these equations is taken from the `torch-spde` repository [44] [5].

Navier–Stokes and Stochastic Navier–Stokes Equation: The 2D stochastic Navier–Stokes equation for an incompressible flow has the form:

$$\partial_t w - \nu \Delta w = -u \cdot \nabla w + f + \sigma \xi, \quad w(0, x) = w_0(x), \quad (t, x) \in [0, T] \times [0, 1]^2,$$

where the deterministic force is a function of space only, and ξ is a Q-Wiener process colored in space and rescaled by $\sigma = 0.05$. The initial condition is a Gaussian Random Field $w_0 \sim \mathcal{N}(0, 3^{3/2}(-\Delta + 49I)^{-3})$ with periodic boundary conditions, and the viscosity is $\nu = 10^{-3}$. The base Navier–Stokes equation corresponds to the case where $\sigma = 0$ (no stochastic forcing). We again use the simulation code from `torch-spde` for these 2D datasets, on a uniform 64×64 grid. The turbulent benchmark of Section 4.2 uses the $\nu = 10^{-5}$ data of Li et al. [32] on the same grid (1200 trajectories; we keep the snapshots at $t \geq 10$ and train on 9600 of them), and a 128×128 forced Navier–Stokes dataset is used only for the cost measurements of Appendix C.9.

C.4 Training Configuration Description

The following tables give the hyperparameters of our runs. In all cases, the base FFM model follows the configuration in the original codebase of Kerrigan et al. [25]. We keep the neural architectures fixed and tune only the OT-related hyperparameters; for the 2D PDE datasets we additionally study the Gaussian-prior length scale in Appendix C.16.

1. All FFM and kFFM runs use the Adam optimizer with learning rate 10^{-3} and a StepLR scheduler (step size 50, $\gamma = 0.1$).
2. Table 5 gives the FNO configurations on the 1D sequence datasets and Table 6 those on the 1D PDE datasets. We perform no hyperparameter search for the FNO backbone; only the OT-related hyperparameters are varied, so the configurations are dataset-specific. "Batch Size (Sig)" is the batch size used with the signature kernel, whose memory grows with path length.
3. The hyperparameters of the FNO on the 2D PDE datasets are in Table 7.
4. Shared training settings are given in Table 8.

Table 5: Training hyperparameters for 1D sequence datasets (FFM with FNO backbone).

Dataset	Epochs	Batch Size	Batch Size (Sig)	FNO Modes	Width	MLP Width	GP ℓ	GP σ^2
AEMET	300	512	64	64	256	128	0.01	0.1
Gene Expr.	300	512	128	16	256	128	0.01	0.1
Economy	300	512	128	16	128	64	0.01	0.1
Heston	300	512	128	32	256	128	0.01	0.1

Table 6: Training hyperparameters for 1D PDE datasets (FFM with FNO backbone).

Dataset	Epochs	Batch Size	Batch Size (Sig)	FNO Modes	Width	MLP Width	GP ℓ	GP σ^2
KdV	300	16	16	64	256	128	0.01	0.1
Stochastic KdV	100	512	128	32	256	128	0.01	0.1

C.5 Configuration Selection Protocol

kFFM as evaluated in this paper has two per-dataset configuration choices: the OT cost entering the entropic coupling (signature kernel, Euclidean–RBF kernel, Sobolev-RBF kernel, or the raw L^2 cost) and the Gaussian base measure (a smooth GP prior or white noise). The "kFFM (selected)" column of

[5] `https://github.com/crispitagorico/torchspde`

Table 7: Training hyperparameters for 2D spatial PDE datasets (FFM with 2D-FNO backbone).

Dataset	Epochs	Batch Size	Batch Size (Sig)	FNO Modes	Hidden Ch.	Proj. Ch.	GP ℓ	GP σ^2
Navier–Stokes	300	512	512	16	32	64	0.001	1.0
Stoch. NS	100	512	512	16	32	64	0.001	1.0

Table 8: Common training settings for all datasets.

Parameter	Value
Optimizer	Adam
Learning Rate	10^{-3}
LR Scheduler	StepLR (step=50, γ=0.1)
FFM $\sigma_{\min}$	10^{-4}
Number of Seeds	10 (seeds $2^0, \ldots, 2^9$); 20 (seeds $2^0, \ldots, 2^{19}$) where $n = 20$ in Table 2; 5 on the turbulent benchmark
ODE solver tolerances (atol, rtol)	10^{-5}

Table 1 is determined by a validation-based rule rather than by test-set performance: for each dataset, every candidate configuration is trained identically, the configuration with the best sliced Wasserstein distance — a criterion independent of the kernels being compared — on a held-out validation split is selected, and only the selected configuration is then evaluated on the test set. In practice this rule is nearly deterministic and coincides with a simple prior-knowledge heuristic that practitioners can apply without any search: use the signature kernel with a GP base for path-like or rough sequence data, and a bounded RBF cost (Sobolev-RBF with a GP base on the 1D KdV family, Euclidean-RBF on the Navier–Stokes family) for PDE fields, with the white-noise base selected on the Navier–Stokes family, whose fields are rougher than a smooth GP prior. On our benchmark the validated choice coincides with the best-performing configuration on all datasets except short-horizon Heston, where the top configurations are statistically indistinguishable. Pointwise diagnostics by kernel variant are reported in Appendix C.6; the cost and base measure of each selected cell are tagged in Table 1, and the Sinkhorn regularization and kernel bandwidth follow the defaults of Appendix C.8, to which the results are insensitive within the ranges reported in the sensitivity study below.

C.6 Detailed Main-Paper Results

This appendix includes a self-contained summary of the full baseline roster used in the paper, together with structural diagnostics, fuller pointwise comparisons, kernel ablations, and the super-resolution visualizations referenced in Section 4. The additional pointwise tables follow the evaluation convention of Kerrigan et al. [25]: discretization-level mean/variance MSE summaries together with autocorrelation or log-spectrum diagnostics, and higher moments where reported. These tables focus on models for which such pointwise statistics are currently available, while Table 1 in the main text summarizes the complete baseline family comparison including CFM-OT(L^2).

C.7 Full Statistics Report on Sequence Datasets

For the sequence datasets, in addition to the mean/variance/autocorrelation tables above, we also report kurtosis and skewness in Table 14, since these were also reported in Kerrigan et al. [25]. We keep them in the appendix because they are complementary to the main conclusion rather than central to it.

C.8 Implementation Details and Good Practices

Algorithm 1 (Section 3.1) summarizes kFFM training; here we collect the implementation choices behind it.

Unless stated otherwise, the Sinkhorn regularization is $\epsilon = 0.05$ on the median-normalized cost matrix and the RBF bandwidth is $\sigma = 1$ on the same normalized scale (i.e. the median heuristic). The coupling is computed per minibatch with the following choices, which we found to matter for stability and recommend as defaults. (i) *Median-normalize the per-batch cost matrix* before Sinkhorn; this is the most important stability lever and makes the entropic regularization ϵ scale-free. (ii) Use *log-domain Sinkhorn* [39]. (iii) *Sample endpoint pairs from the plan* rather than using the barycentric projection. (iv) *Use a bounded (kernel) cost*: results are then invariant to ϵ across three

Table 9: Interpretable structural diagnostics complementing the full-baseline MMD comparison in Table 1. The first block reports autocorrelation error on sequence datasets; the second block reports log-spectrum error on PDE datasets. Lower is better. As in Tables 11 and 13, the kFFM column reports the best kernel variant per row (the signature kernel on Gene Expression, Economy, and Heston). All results are averaged over 10 seeds.

Dataset	DDO/NCSN	GANO	DDPM	FFM	CFM-OT(L^2)	kFFM
AEMET (Autocorr.)	$4.21\pm0.94\times10^{-3}$	$4.45\pm3.43\times10^{-6}$	$2.84\pm2.98\times10^{-7}$	$4.69\pm1.53\times10^{-7}$	$\mathbf{1.23\pm2.72\times10^{-8}}$	$2.60\pm1.92\times10^{-7}$
Gene Expr. (Autocorr.)	$5.48\pm0.24\times10^{-2}$	$1.16\pm0.68\times10^{-1}$	$3.59\pm3.46\times10^{-4}$	$1.59\pm1.98\times10^{-4}$	$2.78\pm2.32\times10^{-3}$	$\mathbf{4.18\pm4.02\times10^{-5}}$
Economy (Autocorr.)	$2.73\pm0.95\times10^{-1}$	$2.29\pm1.91\times10^{-1}$	$2.05\pm2.85\times10^{-2}$	$1.72\pm1.02\times10^{-4}$	$9.35\pm3.75\times10^{-3}$	$\mathbf{8.90\pm8.39\times10^{-5}}$
Heston (Autocorr.)	$4.36\pm0.25\times10^{-2}$	$8.36\pm8.68\times10^{-3}$	$3.35\pm2.80\times10^{-4}$	$2.78\pm2.05\times10^{-4}$	$1.73\pm1.17\times10^{-4}$	$\mathbf{3.31\pm3.38\times10^{-5}}$
KdV (Log-spectrum)	$4.52\pm0.00\times10^{1}$	$2.02\pm0.30\times10^{1}$	$1.71\pm0.13\times10^{1}$	$1.72\pm0.03\times10^{1}$	$2.73\pm0.05\times10^{1}$	$\mathbf{1.69\pm0.02\times10^{1}}$
Navier–Stokes (Log-spectrum)	$9.93\pm0.08\times10^{-1}$	$2.21\pm0.15\times10^{0}$	$\mathbf{2.58\pm0.63\times10^{-1}}$	$7.92\pm0.89\times10^{-1}$	$3.44\pm0.35\times10^{-1}$	$5.87\pm0.46\times10^{-1}$
Stoch. KdV (Log-spectrum)	$1.84\pm0.00\times10^{1}$	$1.45\pm0.09\times10^{1}$	$\mathbf{1.57\pm0.22\times10^{0}}$	$4.27\pm0.25\times10^{0}$	$9.25\pm0.31\times10^{0}$	$3.97\pm0.28\times10^{0}$
Stoch. NS (Log-spectrum)	$1.06\pm0.28\times10^{0}$	$1.03\pm0.93\times10^{0}$	$8.60\pm2.39\times10^{-1}$	$6.22\pm1.05\times10^{-2}$	$\mathbf{3.11\pm0.93\times10^{-2}}$	$3.22\pm0.63\times10^{-2}$

Table 10: Sequence-data pointwise diagnostics following the evaluation convention of Kerrigan et al. [25]: mean and variance of the discretization-level MSE between target and generated samples, together with average autocorrelation error. We compare baseline FFM (Independent) against the OT-coupled variants: Signature = signature kernel, RBF = Euclidean–RBF kernel, Euclidean = raw L^2 cost with the GP base (kFFM-Euc). Results are mean±std over 10 seeds. Best is **green**; second best is **orange**.

Dataset	Kernel	Mean	Variance	Autocorr.
AEMET	Independent	$7.38 \pm 6.37 \times 10^{-2}$	$1.20 \pm 0.40 \times 10^{-2}$	$4.69 \pm 1.53 \times 10^{-7}$
	Signature	$3.60 \pm 3.66 \times 10^{-2}$	$7.84 \pm 5.59 \times 10^{-3}$	$3.87 \pm 2.45 \times 10^{-7}$
	RBF	$1.51 \pm 1.61 \times 10^{-2}$	$4.70 \pm 2.57 \times 10^{-3}$	$2.60 \pm 1.92 \times 10^{-7}$
	Euclidean	$4.10 \pm 2.69 \times 10^{-2}$	$5.59 \pm 2.65 \times 10^{-3}$	$3.47 \pm 1.99 \times 10^{-7}$
Gene Expr.	Independent	$1.39 \pm 0.30 \times 10^{-3}$	$5.13 \pm 0.74 \times 10^{-3}$	$1.59 \pm 1.98 \times 10^{-4}$
	Signature	$5.21 \pm 0.98 \times 10^{-4}$	$2.23 \pm 0.23 \times 10^{-3}$	$4.18 \pm 4.02 \times 10^{-5}$
	RBF	$1.25 \pm 0.20 \times 10^{-3}$	$5.67 \pm 0.74 \times 10^{-3}$	$4.54 \pm 6.51 \times 10^{-5}$
	Euclidean	$1.25 \pm 0.20 \times 10^{-3}$	$5.56 \pm 0.61 \times 10^{-3}$	$4.45 \pm 4.47 \times 10^{-5}$
Economy	Independent	$4.90 \pm 3.26 \times 10^{-5}$	$1.08 \pm 1.13 \times 10^{-5}$	$1.72 \pm 1.02 \times 10^{-4}$
	Signature	$3.30 \pm 2.22 \times 10^{-5}$	$6.60 \pm 5.62 \times 10^{-6}$	$8.90 \pm 8.39 \times 10^{-5}$
	RBF	$3.96 \pm 2.12 \times 10^{-5}$	$6.82 \pm 3.44 \times 10^{-6}$	$1.38 \pm 0.89 \times 10^{-4}$
	Euclidean	$3.94 \pm 2.07 \times 10^{-5}$	$6.76 \pm 3.48 \times 10^{-6}$	$1.38 \pm 0.89 \times 10^{-4}$
Heston	Independent	$6.66 \pm 4.77 \times 10^{-3}$	$1.07 \pm 0.25 \times 10^{-1}$	$2.78 \pm 2.05 \times 10^{-4}$
	Signature	$3.33 \pm 2.25 \times 10^{-3}$	$8.17 \pm 1.14 \times 10^{-2}$	$3.31 \pm 3.38 \times 10^{-5}$
	RBF	$4.48 \pm 4.18 \times 10^{-3}$	$8.27 \pm 1.53 \times 10^{-2}$	$1.41 \pm 1.03 \times 10^{-4}$
	Euclidean	$3.38 \pm 3.81 \times 10^{-3}$	$8.40 \pm 0.77 \times 10^{-2}$	$9.08 \pm 7.99 \times 10^{-5}$

orders of magnitude and Sinkhorn converges faster (15–25% faster end-to-end training than with the unbounded L^2 cost; Table 15); with the raw L^2 cost keep $\epsilon \leq 0.1$ (we observed a $\approx 10\times$ degradation at $\epsilon \geq 0.5$ on long volatility paths). (v) Per-batch re-coupling was stable on all datasets. (vi) *Batch size.* Minibatch kernel OT induces the expected minibatch plan $\bar{\pi}_b$ (Section 3.1); in a sweep over $b \in \{64, 128, 256, 512\}$ on Heston and Stoch. KdV, kFFM tracks FFM across the $8\times$ range, and the only seed-consistent edge favors kFFM at the largest batch (Stoch. KdV, $b = 512$: $3/3$ seeds, 0.11 vs. 0.52×10^{-3} MMD-RBF).

C.9 Compute and Memory Cost

Table 15 summarizes the measured cost of the coupling step, profiled on Heston, Stoch. KdV, and Navier–Stokes at 64^2 and 128^2 for batch sizes $b \in \{64, 128, 256, 512\}$. The end-to-end overhead is 5–7% at the largest resolution we train (128^2), and because the $O(b^2)$ coupling cost is independent of model size and resolution, its relative weight shrinks as the backbone grows. The bounded RBF cost also conditions the Sinkhorn iterations (the Gibbs kernel $e^{-c_\kappa/\epsilon}$ stays bounded away from 0 and 1 across batches; Theorem 1), which is why it is about $8\times$ faster than unbounded-L^2 Sinkhorn at $b = 512$. Signature-kernel memory grows with path length, which is why the selection rule assigns RBF or L^2 costs to very long sequences.

Table 11: Baseline comparison for time series datasets among models with available pointwise statistics. The full baseline roster, including CFM-OT(L^2), is summarized in Table 1. Following Kerrigan et al. [25], we report mean/variance MSE and autocorrelation error computed on the discretized samples. The kFFM row reports the best of the kernel variants in Table 10 for each metric (an oracle over variants, used only for these diagnostics), whereas Table 1 uses the validation-selected configuration. Results are mean±std over 10 seeds. Best is **green**; second best is **orange**.

Dataset	Model	Mean	Variance	Autocorr.
AEMET	DDO/NCSN	$3.56 \pm 0.96 \times 10^{-1}$	$1.17 \pm 0.03 \times 10^{-1}$	$4.21 \pm 0.94 \times 10^{-3}$
	DDPM	$1.92 \pm 1.29 \times 10^{-1}$	$3.22 \pm 2.98 \times 10^{-2}$	$2.84 \pm 2.98 \times 10^{-7}$
	GANO	$2.47 \pm 0.89 \times 10^{-1}$	$2.35 \pm 0.01 \times 10^{-1}$	$4.45 \pm 3.43 \times 10^{-6}$
	FFM	$7.38 \pm 6.37 \times 10^{-2}$	$1.20 \pm 0.40 \times 10^{-2}$	$4.69 \pm 1.53 \times 10^{-7}$
	kFFM	$1.51 \pm 1.61 \times 10^{-2}$	$4.70 \pm 2.57 \times 10^{-3}$	$2.60 \pm 1.92 \times 10^{-7}$
Gene Expr.	DDO/NCSN	$6.07 \pm 1.06 \times 10^{-3}$	$9.69 \pm 0.70 \times 10^{-3}$	$5.48 \pm 0.24 \times 10^{-2}$
	DDPM	$4.10 \pm 1.24 \times 10^{-3}$	$1.08 \pm 0.08 \times 10^{-2}$	$3.59 \pm 3.46 \times 10^{-4}$
	GANO	$1.89 \pm 0.52 \times 10^{-2}$	$3.14 \pm 0.00 \times 10^{-2}$	$1.16 \pm 0.68 \times 10^{-1}$
	FFM	$1.39 \pm 0.30 \times 10^{-3}$	$5.13 \pm 0.74 \times 10^{-3}$	$1.59 \pm 1.98 \times 10^{-4}$
	kFFM	$5.21 \pm 0.98 \times 10^{-4}$	$2.23 \pm 0.23 \times 10^{-3}$	$4.18 \pm 4.02 \times 10^{-5}$
Economy	DDO/NCSN	$1.70 \pm 1.34 \times 10^{-3}$	$9.91 \pm 2.06 \times 10^{-3}$	$2.73 \pm 0.95 \times 10^{-1}$
	DDPM	$6.08 \pm 7.92 \times 10^{-4}$	$2.27 \pm 5.34 \times 10^{-4}$	$2.05 \pm 2.85 \times 10^{-2}$
	GANO	$3.79 \pm 0.51 \times 10^{-1}$	$7.28 \pm 4.08 \times 10^{-4}$	$2.29 \pm 1.91 \times 10^{-1}$
	FFM	$4.90 \pm 3.26 \times 10^{-5}$	$1.08 \pm 1.13 \times 10^{-5}$	$1.72 \pm 1.02 \times 10^{-4}$
	kFFM	$3.30 \pm 2.22 \times 10^{-5}$	$6.60 \pm 5.62 \times 10^{-6}$	$8.90 \pm 8.39 \times 10^{-5}$
Heston	DDO/NCSN	$1.60 \pm 0.47 \times 10^{-2}$	$7.32 \pm 0.15 \times 10^{-1}$	$4.36 \pm 0.25 \times 10^{-2}$
	DDPM	$9.33 \pm 9.43 \times 10^{-3}$	$1.08 \pm 0.25 \times 10^{-1}$	$3.35 \pm 2.80 \times 10^{-4}$
	GANO	$1.55 \pm 0.90 \times 10^{-2}$	$4.41 \pm 1.15 \times 10^{-1}$	$8.36 \pm 8.68 \times 10^{-3}$
	FFM	$6.66 \pm 4.77 \times 10^{-3}$	$1.07 \pm 0.25 \times 10^{-1}$	$2.78 \pm 2.05 \times 10^{-4}$
	kFFM	$3.33 \pm 2.25 \times 10^{-3}$	$8.17 \pm 1.14 \times 10^{-2}$	$3.31 \pm 3.38 \times 10^{-5}$

Table 12: PDE pointwise diagnostics following the evaluation convention of Kerrigan et al. [25]: mean and variance of the discretization-level MSE between target and generated samples, together with average log-spectrum error. Here we compare the baseline FFM (Independent) against its OT-coupled variants: RBF = Euclidean–RBF kernel, Euclidean = raw L^2 cost with the GP base (kFFM-Euc). Results are mean±std over 10 seeds. Best is **green**; second best is **orange**.

Dataset	Kernel	Mean	Variance	Spectrum (log)
KdV	Independent	$5.20 \pm 3.37 \times 10^{-4}$	$1.39 \pm 0.49 \times 10^{-4}$	$1.72 \pm 0.03 \times 10^{1}$
	RBF	$1.48 \pm 0.95 \times 10^{-4}$	$5.68 \pm 2.58 \times 10^{-5}$	$1.69 \pm 0.02 \times 10^{1}$
	Euclidean	$2.51 \pm 0.74 \times 10^{-4}$	$6.88 \pm 3.30 \times 10^{-5}$	$1.69 \pm 0.03 \times 10^{1}$
Navier–Stokes	Independent	$7.96 \pm 7.82 \times 10^{-2}$	$3.11 \pm 0.80 \times 10^{-2}$	$7.92 \pm 0.89 \times 10^{-1}$
	RBF	$2.01 \pm 2.01 \times 10^{-3}$	$7.51 \pm 3.33 \times 10^{-4}$	$5.87 \pm 0.46 \times 10^{-1}$
	Euclidean	$2.03 \pm 1.68 \times 10^{-2}$	$1.87 \pm 0.57 \times 10^{-2}$	$6.00 \pm 0.57 \times 10^{-1}$
Stoch. KdV	Independent	$1.69 \pm 1.31 \times 10^{-3}$	$3.15 \pm 2.31 \times 10^{-3}$	$4.27 \pm 0.25 \times 10^{0}$
	RBF	$1.17 \pm 0.72 \times 10^{-3}$	$1.62 \pm 0.91 \times 10^{-3}$	$4.05 \pm 0.14 \times 10^{0}$
	Euclidean	$2.34 \pm 1.71 \times 10^{-3}$	$2.51 \pm 1.50 \times 10^{-3}$	$3.97 \pm 0.28 \times 10^{0}$
Stoch. NS	Independent	$3.72 \pm 2.31 \times 10^{-1}$	$5.76 \pm 0.76 \times 10^{0}$	$6.22 \pm 1.05 \times 10^{-2}$
	RBF	$3.38 \pm 1.15 \times 10^{-2}$	$2.59 \pm 0.76 \times 10^{-1}$	$3.22 \pm 0.63 \times 10^{-2}$
	Euclidean	$1.81 \pm 0.66 \times 10^{-1}$	$4.59 \pm 0.81 \times 10^{0}$	$5.25 \pm 0.62 \times 10^{-2}$

C.10 Non-Kernel Metrics

Table 3 (Section 4.1) reports sliced Wasserstein distance and marginal W_1 for kFFM versus FFM on the same runs. Coupling and evaluation kernels already differ for our strongest configurations (signature coupling, RBF-MMD evaluation), and the gains persist under every non-kernel metric. Table 16 adds the autocorrelation and log-spectrum errors on the same matched runs for the sequence datasets; the pointwise diagnostics by kernel variant on the original 10-seed runs are in Appendix C.6.

Table 13: Baseline comparison for PDE datasets among models with available pointwise statistics. The full baseline roster, including CFM-OT(L^2), is summarized in Table 1. Following Kerrigan et al. [25], we compare kFFM against baseline models on discretization-level mean/variance MSE and log-spectrum error; the kFFM row reports the best of the kernel variants in Table 12 for each metric (an oracle over variants, used only for these diagnostics), whereas Table 1 uses the validation-selected configuration. Results are mean±std over 10 seeds. Best is **green**; second best is **orange**.

Dataset	Model	Mean	Variance	Spectrum (log)
KdV	DDO/NCSN	$2.46 \pm 0.46 \times 10^{-1}$	$2.78 \pm 0.43 \times 10^{-2}$	$4.52 \pm 0.00 \times 10^{1}$
	DDPM	$3.67 \pm 0.91 \times 10^{-3}$	$5.81 \pm 0.69 \times 10^{-3}$	$1.71 \pm 0.13 \times 10^{1}$
	GANO	$1.87 \pm 0.94 \times 10^{-1}$	$2.91 \pm 0.16 \times 10^{-2}$	$2.02 \pm 0.30 \times 10^{1}$
	FFM	$5.20 \pm 3.37 \times 10^{-4}$	$1.39 \pm 0.49 \times 10^{-4}$	$1.72 \pm 0.03 \times 10^{1}$
	kFFM	$1.48 \pm 0.95 \times 10^{-4}$	$5.68 \pm 2.58 \times 10^{-5}$	$1.69 \pm 0.02 \times 10^{1}$
Navier–Stokes	DDO/NCSN	$4.79 \pm 0.01 \times 10^{-1}$	$4.39 \pm 0.00 \times 10^{-2}$	$9.93 \pm 0.08 \times 10^{-1}$
	DDPM	$5.27 \pm 0.21 \times 10^{-1}$	$2.17 \pm 0.52 \times 10^{-2}$	$2.58 \pm 0.63 \times 10^{-1}$
	GANO	$6.16 \pm 4.25 \times 10^{-2}$	$2.74 \pm 1.59 \times 10^{-2}$	$2.21 \pm 0.15 \times 10^{0}$
	FFM	$7.96 \pm 7.82 \times 10^{-2}$	$3.11 \pm 0.80 \times 10^{-2}$	$7.92 \pm 0.89 \times 10^{-1}$
	kFFM	$2.01 \pm 2.01 \times 10^{-3}$	$7.51 \pm 3.33 \times 10^{-4}$	$5.87 \pm 0.46 \times 10^{-1}$
Stoch. KdV	DDO/NCSN	$3.59 \pm 0.60 \times 10^{-2}$	$1.55 \pm 0.01 \times 10^{-1}$	$1.84 \pm 0.00 \times 10^{1}$
	DDPM	$3.90 \pm 0.78 \times 10^{-3}$	$7.19 \pm 8.48 \times 10^{-3}$	$1.57 \pm 0.22 \times 10^{0}$
	GANO	$1.80 \pm 0.31 \times 10^{-3}$	$1.42 \pm 1.57 \times 10^{-2}$	$1.45 \pm 0.09 \times 10^{1}$
	FFM	$1.69 \pm 1.31 \times 10^{-3}$	$3.15 \pm 2.31 \times 10^{-3}$	$4.27 \pm 0.25 \times 10^{0}$
	kFFM	$1.17 \pm 0.72 \times 10^{-3}$	$1.62 \pm 0.91 \times 10^{-3}$	$3.97 \pm 0.28 \times 10^{0}$
Stoch. NS	DDO/NCSN	$3.45 \pm 0.01 \times 10^{-1}$	$7.59 \pm 0.02 \times 10^{0}$	$1.06 \pm 0.28 \times 10^{0}$
	DDPM	$3.52 \pm 0.36 \times 10^{-1}$	$4.81 \pm 0.28 \times 10^{0}$	$8.60 \pm 2.39 \times 10^{-1}$
	GANO	$4.38 \pm 0.48 \times 10^{-1}$	$7.05 \pm 0.65 \times 10^{0}$	$1.03 \pm 0.93 \times 10^{0}$
	FFM	$3.72 \pm 2.31 \times 10^{-1}$	$5.76 \pm 0.76 \times 10^{0}$	$6.22 \pm 1.05 \times 10^{-2}$
	kFFM	$3.38 \pm 1.15 \times 10^{-2}$	$2.59 \pm 0.76 \times 10^{-1}$	$3.22 \pm 0.63 \times 10^{-2}$

C.11 Turbulent Navier–Stokes: Samples and Statistics

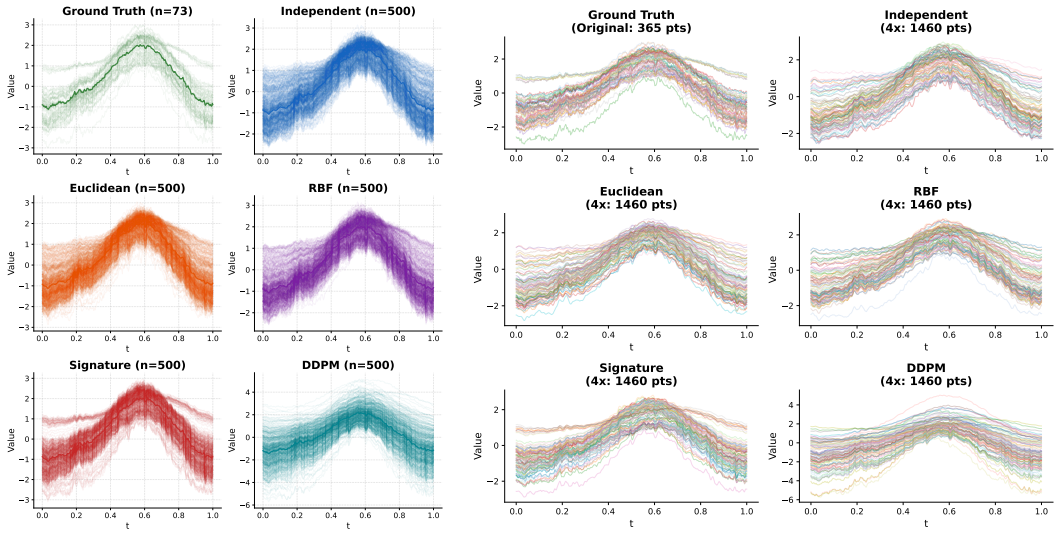
Table 17 gives the turbulent benchmark of Section 4.2 with explicit configurations and unscaled values, and Figure 4 shows vorticity samples (pooled vorticity PDFs and energy spectra are in Figure 1). We also re-ran the standard Navier–Stokes configurations of Section 4.1 with the same physics diagnostics (3 seeds): the flow-matching family leads (enstrophy W_1 of 0.16 vs. 0.28 for DDPM; vorticity-PDF W_1 of 0.14–0.15 vs. 0.39), and kFFM matches FFM within noise on the physics metrics while retaining its distributional edge from Table 1. The kernel-cost mismatch is larger in the turbulent regime: prior-data L^2 distances more than double relative to the standard benchmark (median 91 vs. 40), pushing the RBF cost deeper into saturation, the regime where Theorem 2 is weakest; the empirical results carry the method there.

C.12 Measuring the Kernel-Cost Mismatch

Analytically, $c_{\mathrm{RBF}}(f, g) = 2(1 - e^{-\|f-g\|^2/2\sigma^2})$ is a rescaled L^2 cost up to relative error $O((D/\sigma)^2)$, with D the batch diameter; median normalization cancels the rescaling, so Δ_κ is small at large bandwidth. We measured the mismatch on training-faithful batches through the rank correlation between kernel and L^2 cost matrices and the total variation (TV) distance between the induced plans. (1) Where the RBF cost is not saturated it is a monotone transform of L^2 (rank correlation ≈ 1: Gene Expression 1.000, Heston 0.98), i.e. the same geometry with compressed outliers. (2) The signature kernel genuinely re-ranks pairs (rank correlation with L^2 of only 0.33–0.37 on AEMET and Gene Expression), the data class where its wins are largest and most significant; Δ_κ is large there by design, and Theorem 2 controls the deviation from Wasserstein with the chosen cost rather than proximity to L^2 OT. (3) On smooth Navier–Stokes fields the plans agree across costs (plan TV ≤ 0.02), the small-Δ_κ regime where the bound is tightest and the bounded cost's stability and speed are the operative benefits.

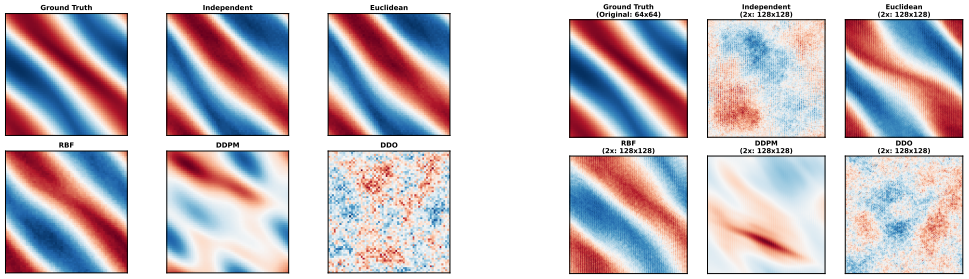
C.13 Synthetic Couplings: Where L^2 Pairing Fails

Figure 5 isolates the two failure modes of L^2 pairing discussed in Section 4.1 on synthetic data at the training coupling protocol (stable over 6 seeds): a conditioning failure under heavy-tailed nuisance



(a) Samples. (b) 4x temporal super-resolution.

Figure 2: Qualitative results on AEMET; panels labeled Euclidean, RBF, and Signature are kFFM variants (kFFM-Euc, kFFM-RBF, kFFM-Sig) and Independent is FFM. Kernel OT better preserves the seasonal spread and amplitude of trajectories both at the original resolution and under temporal super-resolution.



(a) Samples. (b) 2x spatial super-resolution.

Figure 3: Qualitative results on Navier–Stokes; panels labeled Euclidean and RBF are kFFM variants (kFFM-Euc, kFFM-RBF) and Independent is FFM. Kernel OT produces smoother and more coherent flow structures than independent coupling, and this advantage persists under spatial super-resolution.

bursts, cured by any bounded cost, and a geometry failure on paths matched by volatility, cured only by the signature kernel.

C.14 Why Straightness Arguments Do Not Directly Transfer

Finite-dimensional OT-CFM is often motivated by the Euclidean W_2 picture: a better coupling, combined with linear interpolation, can lead to straighter paths and hence lower numerical integration cost. That intuition depends on geometric facts that are specific to the finite-dimensional Wasserstein setting and should not be treated as automatic in function space.

Proposition 6 (Coupling-only role of kernel OT)**.** *Fix a coupling $\pi \in \Pi(\mu_0, \mu_1)$. In kFFM, conditional on $(f_0, f_1) \sim \pi$, the path is*

$$g_t^{f_0,f_1} = tf_1 + \sigma_t f_0, \qquad \sigma_t = 1 - (1 - \sigma_{\min})t,$$

so that

$$\partial_t g_t^{f_0,f_1} = f_1 - (1 - \sigma_{\min})f_0, \qquad \partial_t^2 g_t^{f_0,f_1} = 0 :$$

Table 14: Higher-moment sequence-data statistics (skewness and kurtosis errors) for the models with available pointwise diagnostics, following the evaluation convention of Kerrigan et al. [25]; mean, variance, and autocorrelation errors are in Table 11, and the full baseline roster, including CFM-OT(L^2), in Table 1. The kFFM row uses the best kernel variant per metric. Lower is better. Best is **green**, second best is **orange**.

Dataset	Model	Skewness	Kurtosis
AEMET	DDO/NCSN	$2.73 \pm 0.24 \times 10^{-1}$	$3.57 \pm 0.18 \times 10^{-1}$
	DDPM	$1.61 \pm 0.50 \times 10^{-1}$	$3.95 \pm 1.79 \times 10^{-1}$
	GANO	$2.34 \pm 0.39 \times 10^{-1}$	$3.84 \pm 0.83 \times 10^{-1}$
	FFM	$5.52 \pm 4.06 \times 10^{-2}$	$3.94 \pm 4.16 \times 10^{-1}$
	kFFM	$8.12 \pm 5.35 \times 10^{-3}$	$4.22 \pm 2.61 \times 10^{-2}$
Gene Expr.	DDO/NCSN	$2.02 \pm 0.13 \times 10^{-1}$	$3.75 \pm 0.44 \times 10^{-1}$
	DDPM	$3.72 \pm 1.17 \times 10^{-1}$	$1.57 \pm 1.31 \times 10^{1}$
	GANO	$8.62 \pm 4.56 \times 10^{-1}$	$2.88 \pm 2.57 \times 10^{0}$
	FFM	$1.75 \pm 0.36 \times 10^{-1}$	$1.64 \pm 0.57 \times 10^{0}$
	kFFM	$7.43 \pm 0.76 \times 10^{-2}$	$4.87 \pm 1.17 \times 10^{-1}$
Economy	DDO/NCSN	$4.91 \pm 1.24 \times 10^{-1}$	$4.70 \pm 1.87 \times 10^{0}$
	DDPM	$1.26 \pm 3.23 \times 10^{0}$	$9.00 \pm 34.10 \times 10^{1}$
	GANO	$1.05 \pm 0.57 \times 10^{0}$	$3.78 \pm 1.71 \times 10^{0}$
	FFM	$1.80 \pm 0.80 \times 10^{-1}$	$1.75 \pm 0.69 \times 10^{0}$
	kFFM	$7.65 \pm 2.94 \times 10^{-2}$	$9.77 \pm 4.01 \times 10^{-1}$
Heston	DDO/NCSN	$1.36 \pm 0.01 \times 10^{1}$	$1.58 \pm 0.00 \times 10^{3}$
	DDPM	$3.97 \pm 0.91 \times 10^{0}$	$9.53 \pm 2.04 \times 10^{2}$
	GANO	$6.19 \pm 0.99 \times 10^{0}$	$1.33 \pm 0.06 \times 10^{3}$
	FFM	$5.30 \pm 2.79 \times 10^{0}$	$1.59 \pm 1.12 \times 10^{3}$
	kFFM	$3.31 \pm 0.61 \times 10^{0}$	$8.38 \pm 1.60 \times 10^{2}$

Table 15: Measured training cost of the kernel-OT coupling step across benchmarks. The $O(b^2)$ Sinkhorn solve is independent of model size and resolution.

Quantity	Measurement
End-to-end training overhead vs. FFM	5–7% at the largest trained resolution (128^2 Navier–Stokes)
Peak coupling memory	80–151 MB across benchmarks and $b \in \{64, \dots, 512\}$; 1.5–9% of the FNO forward+backward peak (e.g. 116 MB vs. 7.62 GB on Navier–Stokes 64^2)
Per-batch coupling time	1–80 ms, about one FNO step
Bounded RBF vs. unbounded L^2 Sinkhorn ($b = 512$)	9 vs. 73 ms per batch ($\approx 8\times$); 15–25% faster end-to-end training
Signature-kernel memory	0.63 GB at 100 time points vs. 12.5 GB at 512; the selection rule therefore assigns RBF/L^2 costs to very long sequences
Larger-model control	Giving FFM the +6% budget as extra epochs does not close the gap (loss curves plateau early)

the conditional path is affine in t and deterministic given the pair, and its geometry does not depend on π. Moreover, the training objective can be written as

$$\mathcal{L}_{CFM}^{\pi}(\theta) = \mathbb{E}_{t,\,(f_0,f_1)\sim\pi} \left\| \partial_t g_t^{f_0,f_1} - v_\theta\left(g_t^{f_0,f_1}, t\right) \right\|_{\mathcal{F}}^2 .$$

Therefore, optimizing over π changes only the distribution of endpoint pairs presented to the model; it does not introduce any explicit pathwise straightness, geodesic, or action-minimization term.

Proof. The path and its velocity are exactly those of Section 3.1; they coincide with the FFM path $tf + \sigma_t \xi$ and conditional field $f - (1 - \sigma_{\min})\xi$ of Section 2.1 with the base noise ξ replaced by the paired prior sample f_0. Differentiating the affine path gives the displayed derivatives. Substituting the conditional path into the conditional flow matching objective yields the displayed loss. Since π enters only through the sampling law of (f_0, f_1) and no additional functional of the path $t \mapsto g_t^{f_0,f_1}$ appears in the objective, the method does not explicitly optimize any global path-geometry quantity. □

Table 16: Non-kernel and structural metrics for kFFM-Sig vs. FFM on the sequence datasets (same prior and architecture; lower is better), computed on the matched runs of Table 2: mean±std over 10 shared seeds (Economy averages the population and GDP series within each seed). Every comparison is individually significant when paired by seed (Wilcoxon $p \leq 0.027$; kFFM better on 9–10 of the 10 seeds on both datasets); the wide autocorrelation spread on Gene Expression is across-seed variation shared by both methods, which the pairing removes. These runs are distinct from the 10-seed runs behind Table 10, so the autocorrelation values differ between the two tables; both favor kFFM.

Dataset	Metric	FFM	kFFM-Sig
Gene Expr.	Sliced-W	0.106 ± 0.006	$\mathbf{0.087 \pm 0.006}$
	Marginal-W_1	0.074 ± 0.005	$\mathbf{0.059 \pm 0.004}$
	Autocorr. MSE ($\times 10^{-5}$)	9.3 ± 6.8	$\mathbf{3.3 \pm 3.6}$
	log-Spectrum MSE ($\times 10^{-2}$)	1.76 ± 0.42	$\mathbf{0.86 \pm 0.22}$
Economy	Sliced-W	0.0291 ± 0.0019	$\mathbf{0.0233 \pm 0.0012}$
	Marginal-W_1	0.0171 ± 0.0016	$\mathbf{0.0137 \pm 0.0009}$
	Autocorr. MSE ($\times 10^{-4}$)	1.59 ± 0.39	$\mathbf{1.12 \pm 0.35}$
	log-Spectrum MSE	1.09 ± 0.07	$\mathbf{0.97 \pm 0.04}$

Table 17: Full version of Table 4 with explicit configurations (cost/base measure) and unscaled values. Turbulent 2D Navier–Stokes ($\nu = 10^{-5}$, 1200 trajectories, $\mathrm{Re} \approx 2000$): distributional and physics diagnostics (lower is better), mean±std over 5 seeds, identical FNO backbone and training budget for all flow/diffusion methods. Best per column in bold (on MMD, kFFM and CFM-OT(L^2) tie within std). Vort-PDF W_1: W_1 between pooled vorticity distributions; Enstrophy W_1: W_1 between per-snapshot enstrophy distributions; Skew/Kurt: absolute errors of pooled-vorticity skewness and excess kurtosis; log-Spec: log-spectral error.

Method	MMD	Sliced-W	Vort-PDF W_1	Enstrophy W_1	Skew err.	Kurt err.	log-Spec
kFFM (selected: Euclidean-RBF/wn)	$\mathbf{0.00148 \pm 0.0006}$	$\mathbf{0.193 \pm 0.01}$	$\mathbf{0.0375 \pm 0.01}$	$\mathbf{0.155 \pm 0.02}$	$\mathbf{0.0027 \pm 0.003}$	0.0777 ± 0.01	$\mathbf{0.0204 \pm 0.008}$
CFM-OT(L^2) (raw L^2/wn)	$\mathbf{0.00147 \pm 0.001}$	0.194 ± 0.01	0.0408 ± 0.01	0.157 ± 0.01	0.00404 ± 0.002	0.0752 ± 0.01	0.0223 ± 0.007
kFFM-Sob (Sobolev-RBF/gp)	0.111 ± 0.02	0.612 ± 0.05	0.104 ± 0.03	0.194 ± 0.05	0.0456 ± 0.02	$\mathbf{0.0679 \pm 0.05}$	0.0652 ± 0.01
kFFM-Euc (raw L^2/gp)	0.109 ± 0.02	0.623 ± 0.03	0.104 ± 0.04	0.197 ± 0.06	0.0495 ± 0.03	0.0774 ± 0.04	0.067 ± 0.01
FFM (gp, independent)	0.108 ± 0.02	0.601 ± 0.02	0.107 ± 0.04	0.201 ± 0.06	0.0487 ± 0.03	0.0723 ± 0.05	0.0646 ± 0.01
DDO/NCSN	0.569 ± 0.009	1.39 ± 0.06	1.22 ± 0.01	1.07 ± 0.02	0.0083 ± 0.01	1.29 ± 0.01	0.937 ± 0.02
DDPM	0.38 ± 0.02	1.18 ± 0.1	0.848 ± 0.03	0.889 ± 0.02	0.0789 ± 0.05	3.37 ± 2	0.117 ± 0.03

Corollary 2 (Coupling-invariance of ambient path straightness)**.** *For a twice differentiable path $m : [0,1] \to \mathcal{F}$, define the ambient acceleration and speed-variation functionals*

$$\mathsf{A}(m) := \int_0^1 \|\partial_t^2 m_t\|_{\mathcal{F}}^2 \, dt, \qquad \mathsf{V}(m) := \int_0^1 \left(\|\partial_t m_t\|_{\mathcal{F}} - \int_0^1 \|\partial_s m_s\|_{\mathcal{F}} \, ds \right)^2 dt.$$

Then for every coupling $\pi \in \Pi(\mu_0, \mu_1)$ and every endpoint pair $(f_0, f_1) \sim \pi$, the kFFM conditional path g^{f_0,f_1} satisfies

$$\mathsf{A}(g^{f_0,f_1}) = 0, \qquad \mathsf{V}(g^{f_0,f_1}) = 0.$$

Hence these natural ambient-space straightness measures of the prescribed conditional path are identical for all couplings.

Proof. By Proposition 6, $\partial_t g_t^{f_0,f_1} = f_1 - (1 - \sigma_{\min}) f_0$ is constant in t and $\partial_t^2 g_t^{f_0,f_1} = 0$. Therefore $\mathsf{A}(g^{f_0,f_1}) = 0$. The speed $\|\partial_t g_t^{f_0,f_1}\|_{\mathcal{F}} = \|f_1 - (1 - \sigma_{\min}) f_0\|_{\mathcal{F}}$ is also constant in t, so it coincides with its own time average and $\mathsf{V}(g^{f_0,f_1}) = 0$. □

Proposition 6 and Corollary 2 formalize the internal reason that the usual OT-CFM straightness argument does not transfer directly to kFFM: kernel OT acts only at the level of endpoint coupling, while the interpolation template remains the same affine FFM path. Any improvement in convergence or sampling cost could therefore only arise indirectly through better endpoint pairings and the learned velocity field, not from an explicit path-straightness objective, and we do not claim one.

Our setting is different in a second, geometric way. FFM is formulated on Gaussian measures over separable Hilbert spaces, where absolute continuity and transport already depend on Cameron–Martin and Feldman–Hajek type conditions rather than Lebesgue-density arguments. Even when one restricts

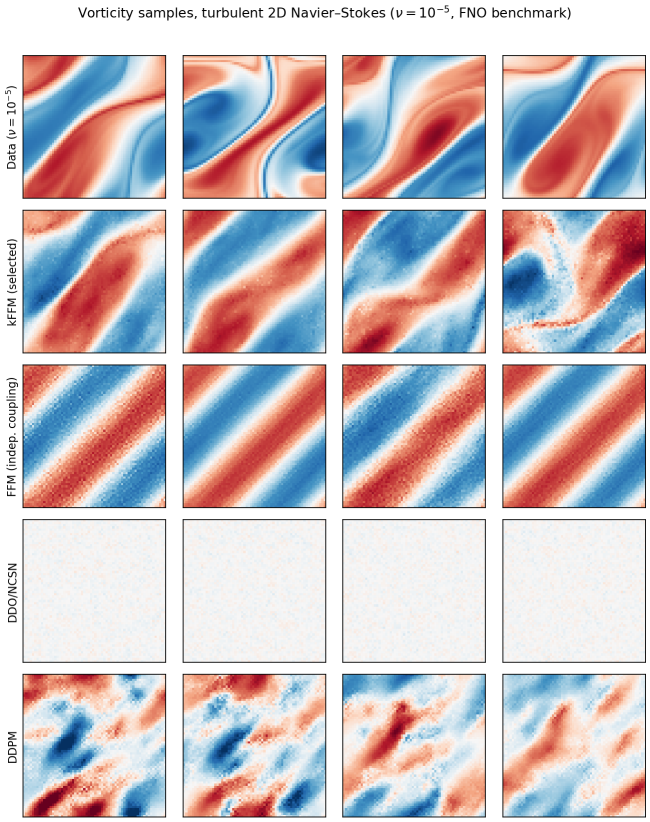


Figure 4: Turbulent 2D Navier–Stokes ($\nu = 10^{-5}$): vorticity samples on a shared color scale (for each method, the samples whose enstrophy is closest to the data median). kFFM (selected) and FFM reproduce the data's large-scale vortical organization, DDO/NCSN produces near-noise fields, and DDPM coarse blobs. The corresponding pooled vorticity PDFs and energy spectra are in Figure 1.

attention to Gaussian optimal transport, the appropriate geometry is the Bures–Wasserstein geometry rather than the flat Euclidean geometry used by standard OT-CFM heuristics.

Recent work by Yun and Zemel [53] makes this distinction explicit. For Gaussian measures on separable Hilbert spaces, they show that the optimal transport map can fail to admit the usual Brenier variational interpretation: formally, it is the subgradient of a convex function that is infinite almost everywhere, and the geodesic structure between degenerate Gaussian measures is richer than the classical finite-dimensional McCann interpolation picture. This does not mean that efficient flows are impossible in function space; rather, it means that the usual "OT implies straighter paths" argument is not a theorem in the present setting.

For this reason, our paper does not claim that kFFM is geodesic or pathwise optimal in a Bures–Wasserstein sense. The theoretical guarantees in this paper concern the well-posedness, error decomposition, and discretization invariance of the kernel OT surrogate, not global straightness of the learned ODE.

This is also why Section 4 makes no efficiency claim: the main supported claim of the paper is better distributional matching from geometry-aware coupling, and Appendix C.15 reports only the training-time convergence of the target metric.

C.15 Target-Metric Convergence During Training

In finite-dimensional OT-CFM, optimal transport is often motivated by improved convergence and straighter flows [49]. As discussed in Appendix C.14, we do not treat such properties as theoretical consequences in function space and make no efficiency claim. We report only the convergence of the target metric during training: Table 18 and Figure 6 show the MMD-RBF of FFM and kFFM over

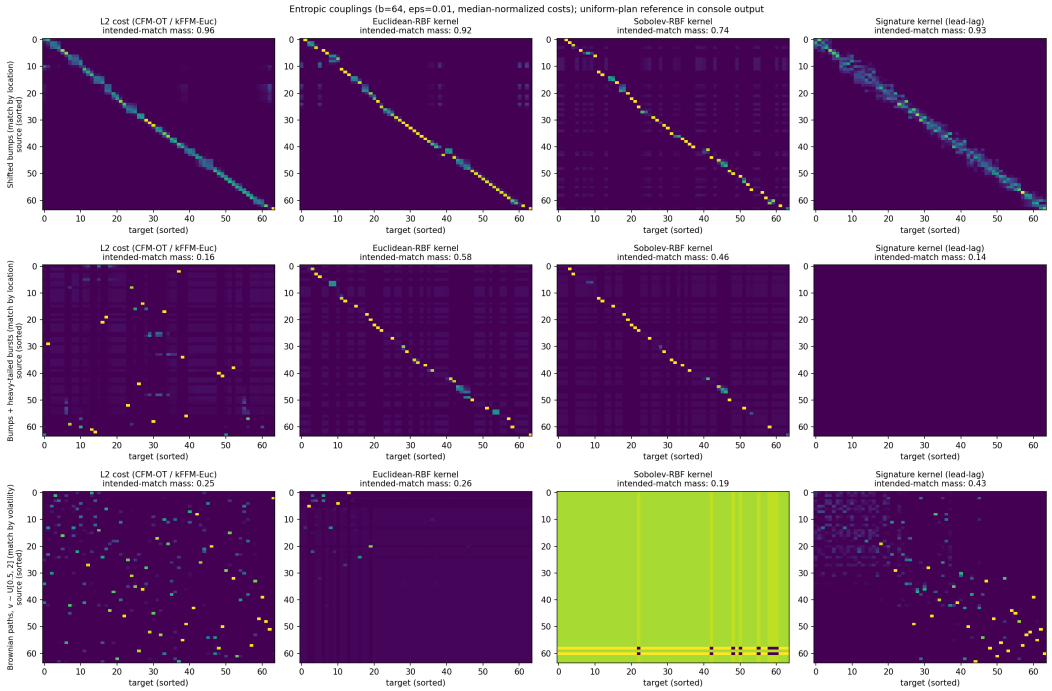


Figure 5: Entropic minibatch couplings ($b = 64$, $\epsilon = 0.01$, median-normalized costs as in training) for three synthetic scenarios (rows) and four costs (columns). Sources and targets are sorted so that the intended matching is the diagonal, and each panel reports the plan mass on intended matches (a uniform plan gives ≈ 0.14–0.19). Top: shifted bumps matched by location; every cost recovers the matching. Middle: bumps where 40% of samples carry large, independent nuisance bursts. Burst pairs dominate the median normalization of the unbounded L^2 cost, collapsing clean-pair contrasts below ϵ, so its plan degenerates to near uniform (0.16); the bounded RBF cost saturates the bursts and recovers the matching (0.58), invariant to burst amplitude (the conditioning effect of Theorem 1). Bottom: Brownian paths $f = vW(t)$, $v \sim U[0.5, 2]$, matched by volatility. The expected L^2 cost is separable, $\mathbb{E}\|f - g\|^2 = (v^2 + w^2) \int t\, dt$, so it carries no assortative signal in expectation and its plan stays near uniform (0.25); the signature kernel (lead-lag) reads volatility through quadratic variation and concentrates on the intended matching (0.43). No single cost wins every scenario, which motivates the per-data-class selection rule of Appendix C.5.

Table 18: Relative MMD-RBF improvement of kFFM over FFM at the end of the convergence curves of Figure 6 (higher is better; values above 100% occur where the unbiased estimate for kFFM is negative). Both methods share the GP prior and FNO backbone, isolating the coupling as the only changed component.

Dataset	Heston	AEMET	KdV	Economy	Gene Expr.	Stoch. KdV	Stoch. NS	Navier–Stokes
ΔMMD-RBF (%)	$+\mathbf{32}\%$	$+\mathbf{53}\%$	$+\mathbf{8}\%$	$+\mathbf{52}\%$	$+\mathbf{64}\%$	$+\mathbf{225}\%$	$+\mathbf{7}\%$	$+\mathbf{9}\%$

the final phase of training, with both models sharing the GP prior and FNO backbone so that the only change is the coupling. The curves are the best-so-far value (cumulative minimum) of the across-seed mean; on KdV, where the unbiased estimate is noisy at the level of single checkpoints, the curves are instead the running mean over the final phase. kFFM ends training at a better MMD-RBF on every dataset, with the largest gains on Stoch. KdV, Gene Expr., AEMET, Economy, and Heston.

C.16 Hyperparameter Sensitivity

Unless otherwise noted, we keep the training setup, neural architectures, Gaussian prior, and remaining OT settings fixed at the default values from the main experiments, and vary only one

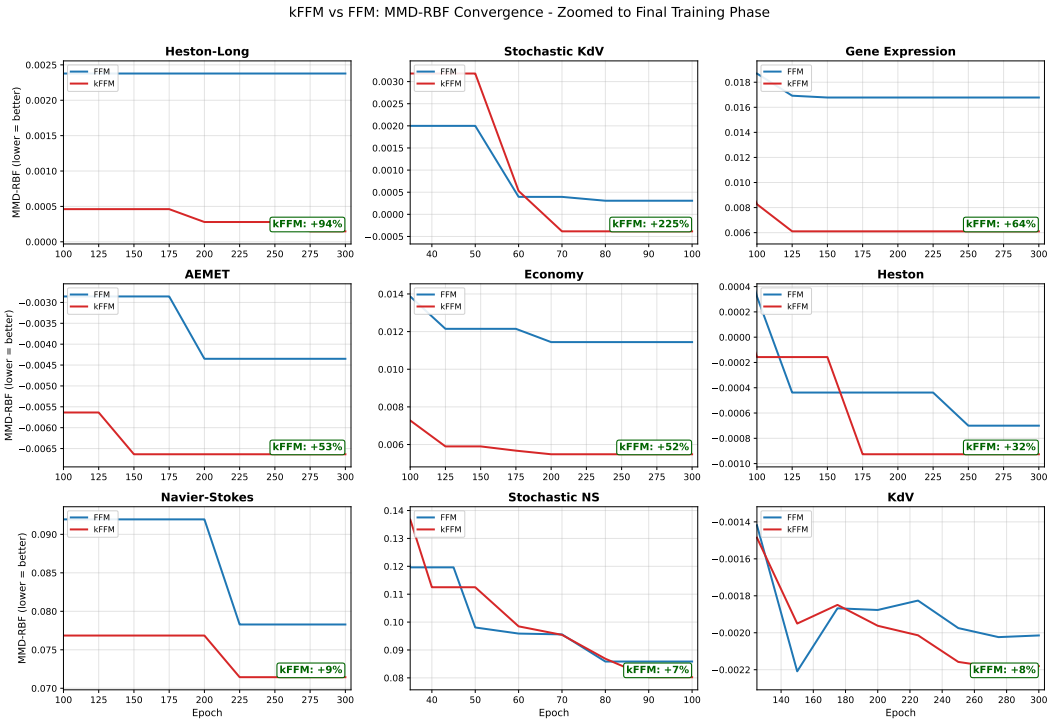


Figure 6: Per-dataset target-metric convergence using MMD-RBF (lower is better) corresponding to the headline summary in Table 18. The plot shows the best-so-far across-seed mean MMD-RBF over the final phase of training (on KdV, the running mean over the final phase, from epoch 125) for kFFM versus the original FFM under matched GP prior and FNO backbone, so the only change is the coupling. The kFFM curve uses the signature kernel on Stoch. KdV, Gene Expr., AEMET, Economy, and Heston, the Sobolev-RBF cost on KdV, the Euclidean–RBF cost on Stoch. NS, and the raw L^2 cost on Navier–Stokes and Heston-Long. kFFM ends training at a better MMD-RBF on every dataset, with especially large gains on Stoch. KdV, Gene Expr., AEMET, Economy, and Heston. The additional Heston-Long panel is a long-horizon variant of Heston (1000 time steps) used only in this diagnostic and in the sensitivity study; it is not one of the eight benchmark datasets.

hyperparameter at a time. We use the `POT`[6] package [17] for optimal transport. The figures below are direct one-at-a-time sweeps: they vary σ or ϵ while holding the other settings fixed. We additionally report the sensitivity to the Gaussian-prior length scale on the 2D PDE datasets, where it has the clearest effect.

The kernel-specific hyperparameters are:

1. RBF kernel: the bandwidth $\sigma \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$.
2. Signature kernel:
   (a) Boolean switches for the lead-lag and time augmentations (`lead_lag`, `time_aug`).
   (b) Dyadic order in $\{0, 1, 2, 3\}$.
   (c) Static-kernel bandwidth between 0.1 and 5.
   (d) Maximum sequence length (subsampling) in $\{32, 50, 64, 128\}$.

The hyperparameters for the Sinkhorn algorithm are:

1. `ot_reg`, the regularization parameter ϵ; the sweeps of Figure 8 use $\{0.01, 0.05, 0.1, 0.5, 1.0\}$, and the full grids extend this range.
2. `ot_method`: Sinkhorn or exact OT; Sinkhorn generally performs better.

[6] `https://pythonot.github.io/`

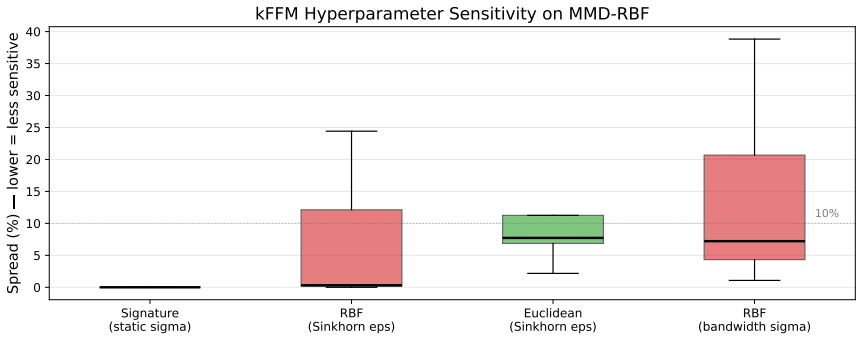


Figure 7: Aggregate MMD-RBF sensitivity of kFFM hyperparameters across datasets; the spread is $(\max - \min)/|\text{mean}|$ over the tested settings of the hyperparameter in parentheses, and lower is less sensitive. Signature-kernel settings are nearly invariant on the tested sequence datasets, the Sinkhorn regularization is typically less influential than the RBF bandwidth, and the results depend more on ϵ under the raw L^2 (Euclidean) cost than under the bounded RBF cost (median spread of about 8% vs. near zero).

3. `ot_coupling`: sampling from the plan (default) or barycentric projection; sampling usually performs better.

Figure 7 summarizes the aggregate MMD-RBF spread across datasets for the main tunable kernel families, where the spread for a given dataset is computed as $(\max - \min)/|\text{mean}|$ over the tested settings. The dominant pattern is that the Sinkhorn regularization is typically quite stable, while the RBF bandwidth can matter more on selected datasets. Figure 8 then shows the direct sweeps on representative datasets, run at the white-noise base measure: the top panel varies the RBF bandwidth, and the bottom panel varies the Sinkhorn regularization. We focus these direct sweeps on the axes that showed the most variation; the signature-kernel settings were substantially flatter on the tested sequence datasets and are summarized in Figure 7. Re-aggregating the full grids: results are indistinguishable within seed-to-seed variation for RBF bandwidths $\sigma \in [0.1, 2]$ on every dataset (the wider sweeps in Figure 8 show where larger bandwidths start to matter on selected datasets); the Sinkhorn regularization is flat across three orders of magnitude ($\epsilon \in [0.001, 1]$) for bounded costs, whereas with the raw L^2 cost $\epsilon \geq 0.5$ degrades long volatility paths by $\approx 10\times$; and the signature hyperparameters (dyadic order 0–3, static bandwidth 0.1–5) leave the results unchanged. The kernel *family* and the base measure are the choices that matter: Euclidean costs, raw or bounded, still lose to the signature kernel by 2.6–$2.9\times$ on Gene Expression and Economy (Section 4.1).

The choice of Gaussian prior matters more on the 2D PDE datasets (Navier–Stokes and Stoch. NS), independently of the kernel. Figure 9 shows the sensitivity of these two datasets to the length scale of the Gaussian prior, which encodes a prior belief about the smoothness of the data.

D Extended Related Work

This section expands the discussion of Section 5. The study of using optimal transport [50] to improve flow matching was first proposed by Pooladian et al. [40], Tong et al. [49], which add an OT-induced coupling of each minibatch before the interpolation and training step. Several follow-up works [28, 52] have aimed to further expand the framework. All have assumed that the sample space is $\mathbb{R}^d$. In order to better capture the resolution-invariance of grid data and continuity of path data, Kerrigan et al. [25] first proposed Functional Flow Matching, a framework to perform flow matching in Hilbert space, relying on Gaussian measure theory. However, OT-based endpoint coupling had not been brought to this space.

The line of research that embeds probability measures in RKHS has been proposed by the pioneering work of Gretton et al. [20], Sriperumbudur et al. [47], where kernel-induced metrics, such as Integral Probability Metrics (IPM), have been studied extensively [10, 47] and used for two-sample hypothesis testing [20], most recently in path space [6] with the signature kernel. The IPM in the form of the Maximum Mean Discrepancy (MMD) has been linked with both probability divergences and

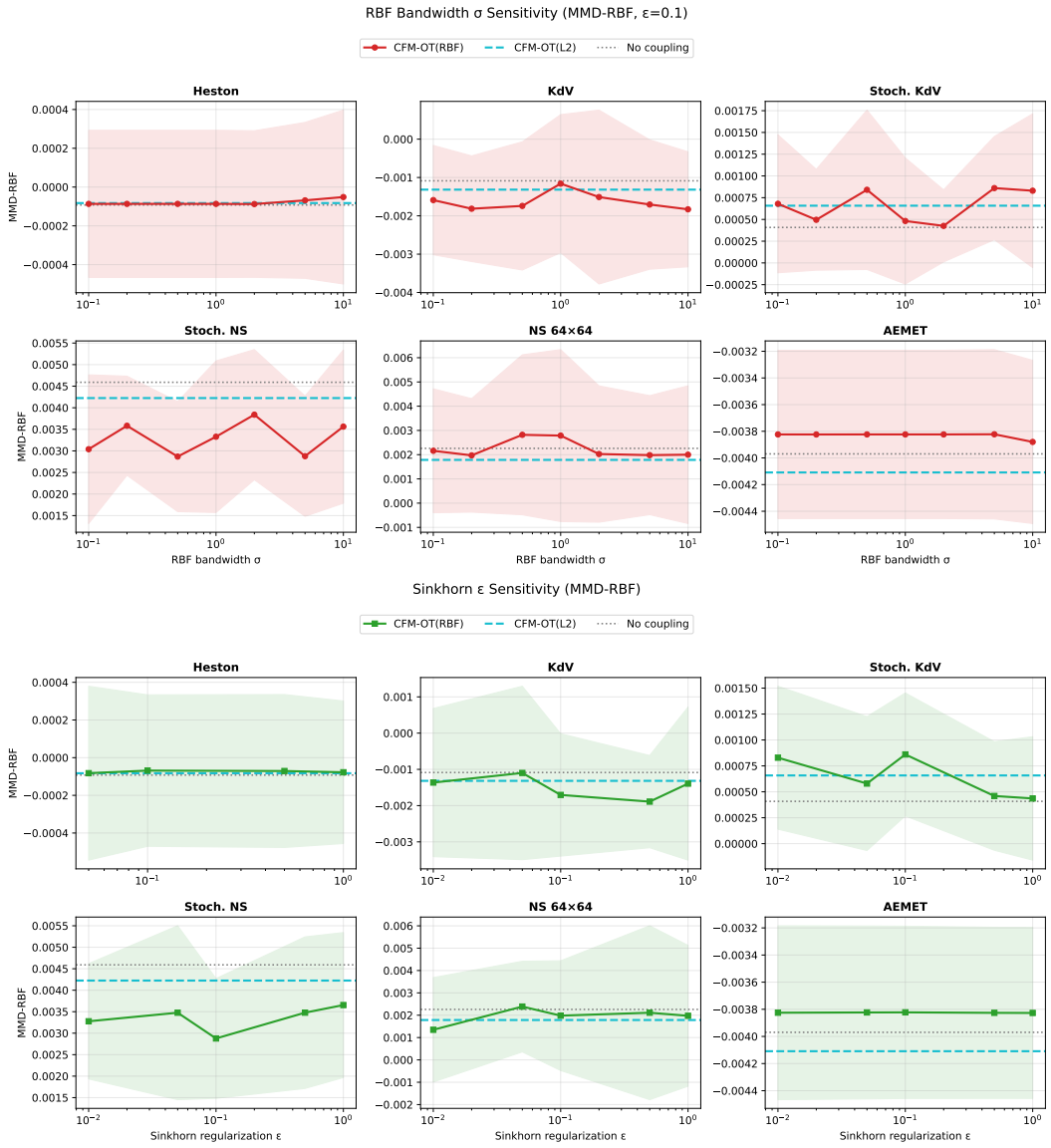


Figure 8: Direct one-at-a-time MMD-RBF sweeps on representative datasets (mean and one standard deviation across seeds). Top: varying the RBF bandwidth σ at fixed Sinkhorn regularization ($\epsilon = 0.1$). Bottom: varying ϵ at fixed bandwidth. All arms in this figure use the white-noise base measure: the solid curve (legend: CFM-OT(RBF)) is the Euclidean–RBF kernel coupling, the dashed line is CFM-OT(L^2), and the dotted line (legend: No coupling) is independent coupling at the same base. Across the tested values the swept arm moves within its seed-to-seed variation.

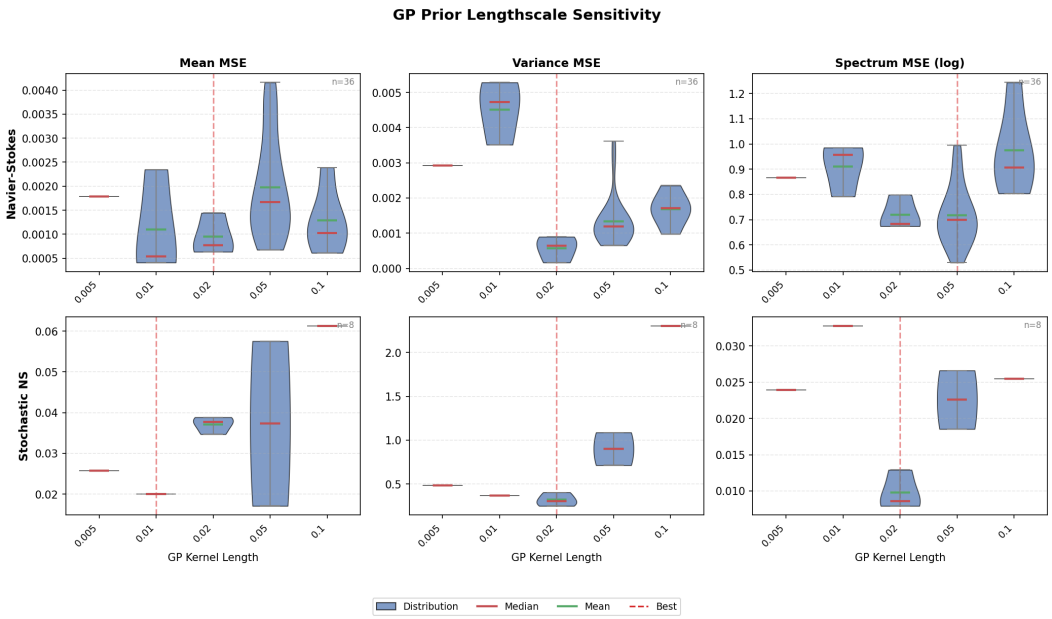


Figure 9: Distribution of the pointwise metrics (mean MSE, variance MSE, log-spectrum error) on the two 2D PDE datasets, Navier–Stokes and Stoch. NS, as a function of the length scale of the Gaussian prior.

Wasserstein distances [16], and it has an entropic OT formulation in the Hilbert Sinkhorn Divergence of Li et al. [31]. It has also been used as the objective of gradient flows [1, 18], which have been adopted in deep generative models. Kernels also define particle dynamics in Stein variational gradient descent (SVGD), which moves particles along the gradient flow of the KL divergence restricted to an RKHS; Messaoud et al. [36] make SVGD-based entropy estimation of unnormalized densities scale to high dimensions, with invertibility and convergence guarantees, and apply it to energy-based image generation and maximum-entropy reinforcement learning. In these particle methods the kernel drives the transport itself, whereas in kFFM it enters only the cost that pairs prior and data samples, and the velocity field is learned by flow matching. The extension of these frameworks to function space is an active area of research, and to our knowledge our work is the first to explore it in the context of flow matching.

Concurrent function-space flow matching. Three concurrent works are closest to ours; none studies the endpoint coupling. Kollovieh et al. [27] use flow matching with GP priors for probabilistic time-series forecasting, a conditional task with an independent prior-data coupling; this is closest in spirit to our GP-prior setup and orthogonal on the axis we study, and our results suggest that framework could itself benefit from kernel-OT coupling. Zhang and Scott [54] straighten flows in Hilbert space via rectification, iteratively re-training on model-generated endpoint pairs; rectification changes pairs across training rounds using the model, whereas kernel OT changes pairs within each batch using data geometry in a single run and extends beyond Hilbert spaces (signature kernel on path space), so the two can in principle be composed. Lee and Flouris [30] propose operator flow matching for time-series forecasting, again conditional prediction with independent coupling. On the finite-dimensional side, Fatras et al. [13, 14] analyze minibatch OT plans and the expected-minibatch coupling that Section 3.1 relies on.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the three theoretical results (Theorems 1–3, Section 3.2) and the empirical claims of Section 4, together with their scope: the theory certifies the surrogate coupling and we make no straightness or sampling-efficiency claim.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 has a dedicated limitations paragraph (scope of the theory, low-regularity regimes, turbulent small-scale statistics, signature-kernel memory); Section 3.2 states where the bounds are uninformative.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: Every theorem is stated with its assumptions in Section 3.2; complete proofs are in Appendix B, with background in Appendix A and the Cameron–Martin assumption behind the CFM/FM equivalence in Appendix B.2.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Datasets, architectures, training budgets, and hyperparameters are in Appendices C.1–C.4; the validation-based configuration selection protocol in Appendix C.5; implementation defaults and good practices in Appendix C.8 (Algorithm 1); and the code is publicly released.

Guidelines:

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

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

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

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

   Answer: [Yes]

   Justification: The complete implementation, including the kernel-OT coupling, the turbulent Navier–Stokes pipeline, the physics diagnostics, the selection protocol, and the scripts for every table and figure, is released at `https://github.com/heraclixus/KFFM`. All datasets are public benchmarks or synthetic simulations whose sources are given in Appendix C.3.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Appendix C gives the model parameterization, baseline configurations, dataset descriptions, per-dataset training configurations (epochs, batch sizes, FNO widths and modes, GP prior parameters), and the selection protocol.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: Tables report mean $\pm$ standard deviation over seeded runs (10 or 20 seeds as marked; 5 on the turbulent benchmark), with all methods sharing seeds. Table 2 reports seed-paired t-tests and Wilcoxon signed-rank tests against FFM, and the turbulent comparisons in Section 4.2 and the non-kernel metrics in Table 3 are paired by seed as well.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Section 4.3 and Appendix C.9 (Table 15) report the per-batch time, peak memory, and end-to-end training overhead of the coupling step relative to FFM on every benchmark; training budgets per dataset are in Appendix C.4.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. The work uses public benchmark datasets and synthetic simulations, involves no human subjects or personal data, and conforms to the Code.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This is foundational methodological work on generative modeling of function-valued data (time series and PDE snapshots); we see no direct societal impact to highlight beyond those general to generative models.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: No pretrained models or datasets with a high risk of misuse are released; the released code trains generative models on public scientific benchmarks.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All datasets and the baseline code bases we build on are cited with their sources in Appendices C.2 and C.3, and used under their respective licenses.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: We release our code (`https://github.com/heraclixus/KFFM`) with instructions for reproducing the experiments; no new datasets are introduced.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were used only for editing and polishing the writing, which does not require declaration under the policy; they are not a component of the method, the theory, the experiments, or the code.

    Guidelines:

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