Neural Optimal Transport in Hilbert Spaces

Jae-Hwan Choi
The Catholic University of Korea
mathjhchoi@catholic.ac.kr

Jiwoo Yoon
Sungkyunkwan University
yoon6019@skku.edu

Dohyun Kwon*
Yonsei University / KIAS
dohyunkwon@yonsei.ac.kr

Jaewoong Choi*
Sungkyunkwan University
jaewoongchoi@skku.edu

Abstract

Functional data such as time series, trajectories, and fields are increasingly central in machine learning and are naturally modeled in infinite-dimensional Hilbert spaces. We study Semi-dual Neural Optimal Transport (SNOT) for learning transport maps between probability measures on separable Hilbert spaces. A critical challenge in this setting is the spurious solution problem: a neural map can globally optimize the SNOT objective while failing to recover an optimal transport map. We show that this failure arises from non-regular source measures, which make the inner minimization in the semi-dual objective non-identifiable. We prove that regular source measures, defined via Gaussian null sets, restore inner-minimizer uniqueness and recover the Monge map. For singular sources, we introduce Gaussian smoothing and establish a necessary and sufficient kernel condition for smoothing to restore regularity. We further prove plan-level consistency: as smoothing vanishes, every Wasserstein accumulation point of the induced optimal plans solves the original Kantorovich problem. Experiments on synthetic singular transports and real-world unpaired time-series imputation benchmarks validate the theory, achieving the best performance on several benchmarks and competitive performance on the remaining ones.

1 Introduction

Functional observations such as PDE solutions, trajectories, time series, and random fields are naturally modeled in infinite-dimensional spaces [18, 20, 35]. Singular or low-dimensional structure and finite discretization complicate identifiability and consistent approximation.

Neural optimal transport (OT) learns maps and potentials for generative modeling [4, 6, 7, 11, 30]. Broader OT applications include time-series imputation [35] and functional domain adaptation [42]. Semi-dual Neural Optimal Transport (SNOT) uses a max–min objective, studied primarily in finite-dimensional Euclidean spaces. We extend it to Hilbert spaces to learn maps between distributions of functions.

We propose *Hilbert Semi-dual Neural Optimal Transport* (HiSNOT) for probability measures on separable Hilbert spaces. We give sufficient conditions for Monge-map identifiability and develop Gaussian regularization for singular sources.

Spurious solution problems. A central theoretical challenge is the *spurious solution* problem [7, 8, 11, 30]. Even given an optimal Kantorovich potential, the inner minimization in the semi-dual formulation may admit global minimizers that achieve the optimal objective value yet deviate from the true Monge map. We interpret this phenomenon as a failure of *inner-minimizer identifiability*.

*Corresponding Authors: Jaewoong Choi and Dohyun Kwon.

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

In infinite-dimensional Hilbert spaces, addressing this issue is challenging due to the absence of a standard Lebesgue measure. To resolve this, our first contribution is an identifiability theorem. Using the classical notion of regular measures defined via Gaussian null sets [1], we prove that this regularity condition restores inner-minimizer identifiability, guaranteeing that the HiSNOT objective uniquely recovers the true Monge map (Theorem 3.2).

Gaussian smoothing for singular functional data. Functional sources may be supported on proper low-dimensional subspaces or exact affine constraints and thus fail to be regular in the ambient Hilbert space. We study how Gaussian smoothing removes this obstruction. The kernel projection of the source must be regular; the converse holds under the Cameron–Martin compatibility condition in Theorem 4.1, and in particular in finite-dimensional spaces. For unrestricted sources in infinite dimensions, injective trace-class covariance is a sufficient choice. A regular kernel marginal alone is not sufficient in general, as the counterexample in Appendix C.1 shows.

Plan-level consistency. As smoothing vanishes, optimal maps of the regularized problems induce plans whose Wasserstein accumulation points solve the original singular Kantorovich problem (Theorem 4.2). This plan-level consistency does not require existence or uniqueness of a limiting Monge map.

Experimental results. Finally, we empirically validate the proposed framework through experiments on both synthetic singular transport tasks and real-world unpaired time-series imputation benchmarks (Table 2). Our results demonstrate that the annealed smoothing approach effectively mitigates spurious solutions in singular environments, allowing HiSNOT to achieve state-of-the-art or highly competitive performance across standard imputation tasks.

Our contributions are summarized as follows:

- We identify spurious solutions in Hilbert-space SNOT as a failure of inner-minimizer identifiability and prove that regular source measures restore unique recovery of the Monge map; see Theorem 3.2.
- We establish a necessary kernel condition for Gaussian smoothing, a converse under Cameron–Martin compatibility, and universal regularization by injective covariance; see Theorem 4.1.
- We prove plan-level consistency of annealed smoothing: every Wasserstein accumulation point of the smoothed optimal plans solves the original Kantorovich problem; see Theorem 4.2.
- We evaluate HiSNOT on finite-mode singular transports, unpaired time-series imputation, and Exchange generation, and report ablations of smoothing support and magnitude.

Relation to prior work. The spurious solution problem in SNOT has been studied in finite-dimensional Euclidean spaces [7, 8, 11]. Choi et al. [7] established that source measures assigning zero mass to sets of Hausdorff dimension at most d-1 ensure identifiability, and proposed Gaussian smoothing with annealing. For compact manifold-supported sources and regular targets, Chu et al. [8] derive finite-dimensional plan-stability rates and use them to select a terminal statistical noise scale. We study the corresponding identifiability and smoothing questions on separable Hilbert spaces using Gaussian-null regularity; we do not establish Hilbert-space analogues of the finite-dimensional convergence-rate guarantees. (For details, please refer to Appendix F.)

For functional data, Zhu et al. [42] also learn Hilbert–Schmidt linear maps between Hilbert spaces, balancing a squared 2-Wasserstein distribution mismatch against a squared Hilbert–Schmidt norm penalty. We study standard quadratic OT: when semi-dual inner minimization identifies the Monge map, and how Gaussian source smoothing restores identifiability under suitable conditions. Appendix F also compares weak-cost Kernel Neural OT and dynamic transport methods.

2 Background

Optimal Transport The Optimal Transport (OT) problem is a mathematical framework that investigates the transport problem between the source measure μ and the target measure ν [31, 34]. Formulated by Monge [25], the Monge OT problem defines the optimal transport map as a deterministic measurable function T that minimizes the total transport cost:

$$\mathcal{T}(\mu,\nu) := \inf_{T_{\#}\mu=\nu} \left[\int_{\mathcal{X}} c(x, T(x)) \mu(\mathrm{d}x) \right]. \tag{1}$$

The pushforward constraint $T_{\#}\mu = \nu$ ensures that the map T transforms the distribution μ into ν, *i.e.*, if $x \sim \mu$, then $T(x) \sim \nu$. Despite its intuitive nature, the Monge problem is inherently

non-convex, and the existence of an optimal transport map generally requires regularity assumptions on the measures μ and ν as well as on the cost function c.

To address these limitations, Kantorovich [15] proposed a convex relaxation of the Monge problem by allowing mass splitting via joint probability measures. The Kantorovich problem is defined as:

$$C(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left[\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y)\right], \tag{2}$$

We refer to the joint probability distribution $\pi \in \Pi(\mu,\nu)$ as the *transport plan* between μ and ν. Unlike the Monge OT problem, the optimal transport plan (OT Plan) $\pi^\star$ is guaranteed to exist under mild assumptions on $(\mathcal{X},\mu)$ and $(\mathcal{Y},\nu)$ and the cost function c [34]. When the optimal transport map $T^\star$ exists, the optimal transport plan $\pi^\star$ is given by the corresponding deterministic coupling, *i.e.*, $\pi^\star = (\mathrm{Id}, T^\star)_{\#}\mu$.

The Kantorovich problem admits the following semi-dual form under lower-semicontinuity and integrable lower-bound assumptions [34, Thm. 5.10]; Santambrogio [31, Prop. 1.11] treats the compact, continuous-cost case:

$$S(\mu,\nu) := \sup_{V\in S_c} \left[\int_{\mathcal{X}} V^c(x)\mu(\mathrm{d}x) + \int_{\mathcal{Y}} V(y)\nu(\mathrm{d}y)\right], \tag{3}$$

where S_c denotes the set of c-concave functions $V : \mathcal{Y} \to \mathbb{R}$ and V^c denotes the c-transform of V:

$$V^c(x) = \inf_{y\in\mathcal{Y}} \left[c(x,y) - V(y)\right]. \tag{4}$$

The semi-dual formulation optimizes out one of the two Kantorovich dual potentials through the c-transform, leaving a single potential V. When an admissible $V^\star$ attains the supremum, it is called a Kantorovich potential. We use globally defined dual representatives satisfying $V^c(x) + V(y) \le c(x,y)$, with $V \in L^1(\nu)$ and $V^c \in L^1(\mu)$; an optimal potential attains the common primal–dual value. All ideal recovery statements assume such attainment.

Neural Optimal Transport in Euclidean Spaces Neural OT learns transport maps in finite-dimensional Euclidean spaces. Balanced semi-dual methods use Eq. 3 [10, 23, 30]; related unbalanced variants relax the marginal constraints [3]. We parameterize the scalar-valued potential by $V_\phi : \mathcal{Y} \to \mathbb{R}$ and the transport map by $T_\theta : \mathcal{X} \to \mathcal{Y}$. The latter approximates a pointwise minimizer in the c-transform of V_ϕ. By leveraging a correspondence between the c-transform V_ϕ^c and the transport network T_θ, we define:

$$T_\theta(x) \in \arg\min_{y\in\mathcal{Y}} \left[c(x,y) - V_\phi(y)\right] \quad \Leftrightarrow \quad V_\phi^c(x) = c\left(x, T_\theta(x)\right) - V_\phi\left(T_\theta(x)\right). \tag{5}$$

This parameterization transforms the semi-dual formulation into a max-min optimization objective $\mathcal{L}(V_\phi, T_\theta)$:

$$\sup_{V_\phi\in S_c} \inf_{T_\theta:\mathcal{X}\to\mathcal{Y}} \mathcal{L}(V_\phi, T_\theta) \qquad \mathcal{L}(V,T) := \int_{\mathcal{X}} c\left(x, T(x)\right) - V\left(T(x)\right)\mu(\mathrm{d}x) + \int_{\mathcal{Y}} V(y)\nu(\mathrm{d}y). \tag{6}$$

Spurious Solutions in Finite Dimensions A fundamental challenge in the SNOT approach is the **spurious solution problem** [7, 11, 30]. This means that the max-min solution $(V^\star, T_{\mathrm{rec}})$ (Eq. 7) of the max-min objective function (Eq. 6) may not accurately recover the optimal transport map $T^\star$. Formally, we define a pair $(V^\star, T_{\mathrm{rec}})$ to be an *optimal max-min solution* to (Eq. 7) if:

$$V^\star \in \arg\max_{V\in S_c} \inf_{T:\mathcal{X}\to\mathcal{Y}} \mathcal{L}(V,T), \qquad T_{\mathrm{rec}} \in \arg\min_{T:\mathcal{X}\to\mathcal{Y}} \mathcal{L}(V^\star, T). \tag{7}$$

It is well-known that the true pair of Kantorovich potential and optimal transport map $(V^\star, T^\star)$ is an optimal max-min solution, but not every solution recovers the true optimal transport map [11, 30].

In finite-dimensional Euclidean spaces $\mathbb{R}^d$, Choi et al. [7] showed that source measures assigning zero mass to sets of Hausdorff dimension at most $d-1$ ensure identifiability. A standard sufficient condition is absolute continuity $\mu \ll \mathcal{L}^d$ with respect to Lebesgue measure. Under this regularity, the spurious solution problem can be resolved, and the neural optimal transport objective correctly recovers the true Monge map. Our work generalizes this sufficient condition to **infinite-dimensional Hilbert spaces**. This requires addressing new theoretical challenges, as there is no Lebesgue measure in Hilbert spaces.

3 Neural Optimal Transport on Hilbert Spaces

This section extends neural optimal transport to separable Hilbert spaces. We formulate the *Hilbert-space SNOT* (HiSNOT) objective (Section 2) and identify the spurious solution challenge for singular functional data. We then introduce a regular measure, which is the Hilbert-space analogue of absolute continuity, and prove that regularity restores identifiability (Section 3.2).

3.1 HiSNOT: Semi-Dual Neural OT on Hilbert Spaces

Let H be a separable Hilbert space. This setting includes the finite-dimensional Euclidean case $H = \mathbb{R}^d$ as a special case, but also covers infinite-dimensional function spaces such as L^2 spaces. Let $\mu, \nu \in \mathcal{P}_2(H)$ denote probability measures with finite second moments, and we consider the quadratic cost $c(x, y) := \frac{1}{2}\|x - y\|_H^2$. Then, the Kantorovich problem and its semi-dual formulation extend naturally to Hilbert spaces:

$$C(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \frac{1}{2} \int_{H \times H} \|x - y\|_H^2 \, \pi(\mathrm{d}x, \mathrm{d}y), \tag{8}$$

$$S(\mu, \nu) := \sup_{V \in S_c} \left\{ \int_H V^c(x) \, \mu(\mathrm{d}x) + \int_H V(y) \, \nu(\mathrm{d}y) \right\}, \tag{9}$$

where S_c consists of globally represented Borel c-concave potentials with $V \in L^1(\nu)$ and $V^c \in L^1(\mu)$, and the c-transform is $V^c(x) := \inf_{y \in H} [c(x, y) - V(y)]$.

Following the finite-dimensional framework (Section 2), our HiSNOT approach parameterizes the potential V by a neural functional V_ϕ and represents the pointwise minimizer in the c-transform by a transport network T_θ:

$$T_\theta(x) \in \arg\min_{y \in H} [c(x, y) - V_\phi(y)]. \tag{10}$$

This leads to the HiSNOT objective

$$\sup_{V_\phi \in S_c} \inf_{T_\theta : H \to H} \mathcal{L}(V_\phi, T_\theta), \quad \mathcal{L}(V, T) := \int_H c(x, T(x)) - V(T(x)) \mu(\mathrm{d}x) + \int_H V(y) \, \nu(\mathrm{d}y). \tag{11}$$

The HiSNOT formulation is a direct extension of finite-dimensional SNOT because the semi-dual structure and max-min formulation remain unchanged. However, as we show next, the spurious solution problem becomes critical for functional data.

Spurious solution in Hilbert spaces As discussed in Section 2, spurious solutions are not limited to finite-dimensional SNOT; the same obstruction appears in Hilbert spaces. Even for a Kantorovich potential $V^\star$, the pointwise minimization

$$\arg\min_{y \in H} [c(x, y) - V^\star(y)]$$

may contain multiple minimizers, allowing a recovered map T_{rec} to achieve the same objective value without coinciding with the Monge map $T^\star$; see Appendix E. We refer to the uniqueness condition

$$\arg\min_{y \in H} [c(x, y) - V^\star(y)] = \{T^\star(x)\} \qquad \mu\text{-a.e.}$$

as **inner-minimizer identifiability**. The next subsection proves that regular source measures restore this property.

3.2 Identifiability under Regular Source Measures

We now introduce a Hilbert-space regularity condition that rules out the spurious-solution obstruction described above. Our approach replaces absolute continuity by regularity with respect to Gaussian null sets. In finite-dimensional Euclidean spaces, Choi et al. [7] rule out spurious solutions under a Hausdorff-dimensional non-charging condition, and a standard sufficient condition is absolute continuity $\mu \ll \mathcal{L}^d$. In Hilbert spaces, there is no infinite-dimensional Lebesgue measure, so we replace absolute continuity by regularity with respect to Gaussian null sets.

Definition 3.1 (Regular measure)**.** Let H be a separable Hilbert space. A probability measure $\mu \in \mathcal{P}(H)$ is called *regular* if $\mu(N) = 0$ for every Gaussian null set $N \subset H$. We denote by $\mathcal{P}^r(H)$ the set of regular probability measures on H, and set $\mathcal{P}_2^r(H) := \mathcal{P}^r(H) \cap \mathcal{P}_2(H)$.

The formal definition of Gaussian null sets is given in Appendix A. When $H = \mathbb{R}^d$, Gaussian null sets coincide with Lebesgue-negligible sets [1]; hence regularity is equivalent to $\mu \ll \mathcal{L}^d$. Thus, regular measures provide the Hilbert-space analogue of absolutely continuous source measures. In particular, if $\mu \in \mathcal{P}_2^r(H)$ and $\nu \in \mathcal{P}_2(H)$, then the quadratic OT problem admits a unique Monge map; see Proposition B.1.

A source with finitely many modes and a joint coefficient density is regular in their span but non-regular in the ambient infinite-dimensional space. Laws on proper closed affine subspaces are also non-regular; nondegenerate Gaussians on H are regular. Appendix H.2 explains these examples and distinguishes empirical measures from population laws.

Theorem 3.2 establishes that regular source measures eliminate spurious solutions in HiSNOT:

Theorem 3.2 (Identifiability of HiSNOT under regularity)**.** *Let H be a separable Hilbert space. Assume that $\mu \in \mathcal{P}_2^r(H)$ and $\nu \in \mathcal{P}_2(H)$. Let $T^\star : H \to H$ be the unique Monge map from μ to ν, and let $V^\star$ be an admissible Kantorovich potential, with $V^\star \in L^1(\nu)$ and $(V^\star)^c \in L^1(\mu)$, represented globally as above. Then the following statements hold.*

(i) For μ-almost every $x \in H$,

$$\arg\min_{y \in H} [c(x, y) - V^\star(y)] = \{T^\star(x)\}. \tag{12}$$

(ii) Let $(V^\star, T_{\mathrm{rec}})$ be an optimal max-min solution of the HiSNOT objective (Eq. 11), with $V^\star$ a Kantorovich potential. Then every measurable inner minimizer coincides with the Monge map μ-almost everywhere; in particular, $T_{\mathrm{rec}}(x) = T^\star(x)$ for μ-a.e. $x \in H$.

Theorem 3.2 shows that regularity restores *inner-minimizer identifiability*: the Kantorovich potential uniquely determines the Monge map through the pointwise minimization problem Eq. 12. Equivalently, under a regular source measure, any optimal max–min solution whose potential is Kantorovich recovers the optimal transport map. Thus, for the population HiSNOT objective with inner minimization over all measurable maps for which the objective is defined, regular source measures rule out spurious inner minimizers. Theorem 3.2 gives pointwise identifiability, Theorem 4.1 specifies valid smoothing conditions, and Theorem 4.2 gives plan-level consistency. Their proofs appear in Appendices B, C, and D, respectively.

4 Method: HiSNOT with Annealed Smoothing

Challenge: singularity of functional data. Theorem 3.2 shows that regularity of the source measure rules out spurious inner minimizers in the ideal HiSNOT objective. However, functional data such as paths, time series, and fields often concentrate on low-dimensional or otherwise singular structures in Hilbert space [36]. Thus, directly applying HiSNOT to functional data can lead to inner-minimizer non-identifiability and spurious solutions, as illustrated in Appendix E. This section **addresses this challenge by introducing Gaussian smoothing that restores regularity** while remaining consistent with the original transport problem. Based on this, we introduce **HiSNOT with annealed smoothing**.

4.1 Theory: Regularization via Gaussian Smoothing

HiSNOT with annealed Gaussian smoothing. Let $Q : H \to H$ be a positive self-adjoint trace-class operator, and let $\mathcal{N}(0, Q)$ be the centered Gaussian measure on H with covariance Q. For each smoothing level $\varepsilon > 0$, we replace the source measure μ in the HiSNOT objective by $\mu_\varepsilon := \mu * \mathcal{N}(0, \varepsilon^2 Q)$. Equivalently, if $X \sim \mu$ and $G \sim \mathcal{N}(0, Q)$ is independent of X, then $X_\varepsilon := X + \varepsilon G \sim \mu_\varepsilon$. For this smoothed source measure, we optimize

$$\sup_{V_\phi \in S_c} \inf_{T_\theta : H \to H} \mathcal{L}_\varepsilon(V_\phi, T_\theta), \tag{13}$$

where

$$\mathcal{L}_\varepsilon(V, T) := \int_H (c(x, T(x)) - V(T(x)))\, \mu_\varepsilon(\mathrm{d}x) + \int_H V(y)\, \nu(\mathrm{d}y).$$

We then anneal the smoothing amplitude $\varepsilon \downarrow 0$.

This construction has two goals. First, for every positive smoothing level, we want $\mu_\varepsilon \in \mathcal{P}_2^r(H)$, so that the corresponding HiSNOT objective is identifiable by Theorem 3.2. Second, as $\varepsilon \downarrow 0$, we want the smoothed problems to remain consistent with the original singular OT problem: if $T_\varepsilon^\star$ is the Monge map from μ_ε to ν, then the induced plans $\pi_\varepsilon^\star := (\mathrm{Id}, T_\varepsilon^\star)_\# \mu_\varepsilon$ should have Wasserstein accumulation points that solve the original Kantorovich problem between μ and ν.

Conditions for successful Gaussian smoothing. The unsmoothed component must already be regular. The converse holds when the source coefficients in the noisy directions satisfy the weighted square-summability condition below.

Theorem 4.1 (Regularity under Gaussian smoothing)**.** *Let Q be a positive self-adjoint trace-class operator on a separable Hilbert space H, let $\gamma_Q = \mathcal{N}(0, Q)$, and let $\mu \in \mathcal{P}(H)$. Let $\{e_k\}$ be an orthonormal eigenbasis of Q, with $Qe_k = \lambda_k e_k$ and $\lambda_k \geq 0$, and let $\pi_{\mathrm{Ker}(Q)}$ denote the orthogonal projection onto* $\mathrm{Ker}(Q)$.

(i) If $\gamma_Q * \mu \in \mathcal{P}^r(H)$, then $(\pi_{\mathrm{Ker}(Q)})_\# \mu \in \mathcal{P}^r(\mathrm{Ker}(Q))$.

(ii) If

$$\sum_{k:\lambda_k>0} \frac{|\langle x, e_k\rangle_H|^2}{\lambda_k} < \infty \qquad \text{for } \mu\text{-almost every } x \in H, \tag{14}$$

then

$$\gamma_Q * \mu \in \mathcal{P}^r(H) \quad \Longleftrightarrow \quad (\pi_{\mathrm{Ker}(Q)})_\# \mu \in \mathcal{P}^r(\mathrm{Ker}(Q)). \tag{15}$$

The additional assumption is automatic when Q has finite rank.

(iii) If $\mathrm{Ker}(Q) = \{0\}$, then, without the additional assumption in part (ii), $\gamma_Q * \mu \in \mathcal{P}^r(H)$ for every $\mu \in \mathcal{P}(H)$.

The condition in Eq. 14 requires finite weighted source energy in positive-variance directions, with no restriction on zero-variance coefficients. It is sufficient for the converse, but not necessary for regularity. Appendix C proves the theorem and gives a counterexample to the unrestricted converse.

Under either sufficient condition above, Theorem 3.2 gives identifiability of exact inner minimizers at positive smoothing levels when an admissible Kantorovich potential is attained. In particular, any covariance with a summable, strictly positive spectrum is injective and hence regularizes every source. In practice, such a covariance may be augmented with a finite-rank term along FPCA directions, which allocates smoothing variance but plays no role in the regularity guarantee; see Appendix H.3.

Plan-level consistency as smoothing vanishes. The second goal is to ensure that the solutions of the smoothed HiSNOT problems remain consistent with the original singular OT problem as the smoothing amplitude vanishes. Since the limiting source measure may be singular, the correct convergence notion is at the level of transport plans rather than maps: the original problem may fail to admit a Monge map, or optimal plans may be non-unique.

Theorem 4.2 (Plan-level consistency of HiSNOT)**.** *Let $\{\mu_k\}_{k\in\mathbb{N}} \subset \mathcal{P}_2^r(H)$ and let $\mu, \nu \in \mathcal{P}_2(H)$ satisfy*

$$\lim_{k\to\infty} W_2(\mu_k, \mu) = 0.$$

For each k, let $T_k^\star$ be the unique optimal Monge map from μ_k to ν, and define the induced optimal plan $\pi_k^\star := (\mathrm{Id}, T_k^\star)_\# \mu_k$. Then the sequence $\{\pi_k^\star\}_{k\in\mathbb{N}}$ is relatively compact in $\mathcal{P}_2(H \times H)$ with respect to the W_2-topology, and every W_2-accumulation point of $\{\pi_k^\star\}_{k\in\mathbb{N}}$ is an optimal transport plan in $\Pi(\mu, \nu)$.

We apply Theorem 4.2 to an annealed sequence of smoothed source measures

$$\mu_{\varepsilon_k} = \mu * \mathcal{N}(0, \varepsilon_k^2 Q), \qquad \varepsilon_k \downarrow 0.$$

Since $W_2^2(\mu_{\varepsilon_k}, \mu) \leq \varepsilon_k^2 \,\mathrm{Tr}(Q) \to 0$, we have $\mu_{\varepsilon_k} \to \mu$ in $W_2(H)$. Under a sufficient condition of Theorem 4.1 (for example, injective Q), each μ_{ε_k} belongs to $\mathcal{P}_2^r(H)$. Hence each positive-smoothing HiSNOT problem admits a unique Monge map $T_{\varepsilon_k}^\star$, and the induced plans $\pi_{\varepsilon_k}^\star := (\mathrm{Id}, T_{\varepsilon_k}^\star)_\# \mu_{\varepsilon_k}$ are relatively compact in $\mathcal{P}_2(H \times H)$. Every Wasserstein accumulation point of these plans is an optimal

Kantorovich plan for the original problem between μ and ν. Thus, HiSNOT with annealed smoothing is consistent at the plan level. This is the natural form of convergence in singular settings, where a limiting Monge map may not exist, and optimal plans may not be unique.

4.2 Implementation Details

We describe the practical implementation of HiSNOT with annealed smoothing. The theoretical objective and consistency guarantees were introduced in Section 4.1; here we focus on the sample-based training procedure. At smoothing level $\varepsilon > 0$, we draw samples $x_i \sim \mu$, $y_i \sim \nu$, and independent perturbations $G_i \sim \mathcal{N}(0, Q)$, and form $\widetilde{x}_i = x_i + \varepsilon G_i \sim \mu_\varepsilon$. We train the potential network V_ϕ and the transport network T_θ using the empirical smoothed objective

$$\widehat{\mathcal{L}}_\varepsilon(\phi, \theta) = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{2}\|\widetilde{x}_i - T_\theta(\widetilde{x}_i)\|_H^2 - V_\phi(T_\theta(\widetilde{x}_i))\right] + \frac{1}{m}\sum_{i=1}^{m} V_\phi(y_i).$$

Training alternates between minimizing this objective with respect to θ and maximizing it with respect to ϕ. The potential V_ϕ is parameterized as a scalar functional on H, and T_θ is parameterized as an H-valued neural operator. In experiments, we use Fourier Neural Operator architectures [18, 20].

Practical Gaussian smoothing via spectral augmentation. For functional data, we generate Gaussian perturbations by adding noise to spectral coefficients. Let $\{e_k\}_{k\in\mathbb{N}}$ be an orthonormal basis of H, and choose summable eigenvalues $\{\lambda_k\}_{k\in\mathbb{N}} \subset [0, \infty)$ defining $Qe_k = \lambda_k e_k$. We sample

$$G = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\, \xi_k e_k, \qquad \xi_k \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

For $x \sim \mu$, where $\mu \in \mathcal{P}_2(H)$ and x is independent of G, the smoothed sample is $x_\varepsilon = x + \varepsilon G = \sum_k (\langle x, e_k\rangle_H + \varepsilon\sqrt{\lambda_k}\xi_k)e_k$, with $\mu_\varepsilon = \mathrm{Law}(x_\varepsilon)$. In practice, we retain M noise modes, using $G_M = \sum_{k=1}^{M}\sqrt{\lambda_k}\xi_k e_k$ and perturbing each sample as $x + \varepsilon G_M$. For fixed $\varepsilon > 0$ and Q, coupling this sample with x_ε using the same x and Gaussian coefficients gives

$$W_2^2(\mathrm{Law}(x + \varepsilon G_M), \mu_\varepsilon) \le \varepsilon^2 \sum_{k>M} \lambda_k \xrightarrow[M\to\infty]{} 0.$$

Because Q is trace class, the total variance $\sum_{k>M}\lambda_k$ of the omitted noise modes tends to zero. This justifies approximating the full-space smoothed source μ_ε by finite-mode perturbations; Theorem 4.1 gives conditions for regularity of μ_ε. Appendix G.1 gives the proof, the retained-coefficient density, and the additional error when the source is also projected onto the retained modes. The values λ_k determine the smoothing direction; directions with $\lambda_k = 0$ are not perturbed. (For details, please refer to Appendix G.)

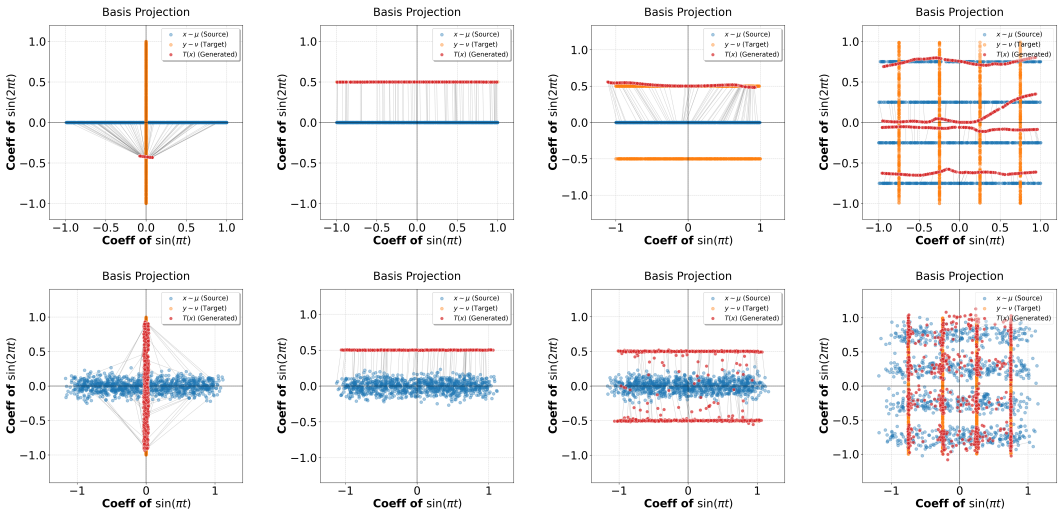
Annealing schedule. We train with a nonincreasing smoothing schedule $\varepsilon_0 \ge \varepsilon_1 \ge \cdots \ge \varepsilon_{N-1} > 0$. At each stage, the parameters learned at level ε_n are used to initialize training at level ε_{n+1}. Following Choi et al. [7], we use decreasing schedules with a positive terminal noise floor; the synthetic interpolation and the clipped imputation schedule are specified separately in Appendix G. Further implementation details and hyperparameters are provided in Appendix G. Appendix H.4 details the additional computational and memory costs of Gaussian smoothing and the assumptions underlying these estimates.

5 Experiments

We empirically validate our theoretical analysis on Semi-dual Neural OT on Hilbert spaces through the following experiments:

- In Sec 5.1, we evaluate the effectiveness of Gaussian smoothing in overcoming spurious solutions.
- In Sec 5.2, we empirically test our characterization theorem for smoothing operators (Theorem 4.1) by comparing theoretically appropriate operators and inappropriate operators.
- In Sec 5.3, we assess our SNOT model on time-series imputation benchmarks, which are real-world applications for high-dimensional function space datasets.

For implementation details of experiments, please refer to Appendix G.



(a) Perpendicular (b) Parallel (c) One-to-Many (d) Grid

Figure 1: **Comparison of transport maps** learned by vanilla HiSNOT (top row) and HiSNOT with annealed smoothing (bottom row). The vanilla model exhibits spurious solutions in singular cases (Perpendicular, One-to-Many, Grid), while smoothing improves agreement with the target distribution in these finite-mode experiments.

5.1 Effect of Gaussian Smoothing on Spurious Solutions

We compare vanilla HiSNOT with its annealed-smoothing variant on four synthetic singular transport examples: *Perpendicular*, *Parallel*, *One-to-Many*, and *Grid*. These examples illustrate the spurious minimizers discussed in Appendix E. Specifically, we compare the vanilla HiSNOT model (Eq. 11) and the regularized version utilizing Gaussian smoothing (Eq. 13). In these settings, the vanilla model is exposed to ill-posedness and spurious solution problems. In contrast, the regularized model introduces principled Gaussian smoothing with a noise level ε. This leads to a well-posed learning objective that eliminates spurious solutions. The evaluation is conducted on synthetic datasets, including Perpendicular (Example 1) and One-to-Many (Example 2) (See Appendix E.2 for details). Here, we utilize two metrics: the transport cost error $D_{cost} = \left|W_2^2(\mu,\nu) - \int_H \|T_\theta(x) - x\|_H^2 \mu(\mathrm{d}x)\right|$ and the target distribution error $D_{target} = W_2^2(T_{\theta\#}\mu, \nu)$. D_{cost} measures how closely the model approximates the theoretically optimal transport cost, while D_{target} measures how accurately the model generates the target distribution.

The results are provided in Table 1 and Figure 1. While the baseline HiSNOT exhibits spurious solutions and fails to accurately reconstruct the target measure, our regularized model substantially reduces target-distribution error in the singular examples. Quantitatively, our model is significantly more accurate in the Perpendicular, One-to-Many, and Grid tasks. Specifically, in the Perpendicular data, the introduction of Gaussian smoothing significantly reduces the transport cost error D_{cost} from 0.101 to 0.038 and the target distribution error D_{target} from 0.249 to 0.006. Interestingly, in the Parallel data, the unregularized model performs well. Thus, both models achieve results with competitive accuracy.

Table 1: **Quantitative comparison of HiSNOT performance with and without Gaussian smoothing** across synthetic scenarios. We report the transport cost error $D_{\text{cost}}(\downarrow)$ and target distribution error $D_{\text{target}}(\downarrow)$. Smoothing (✓) reduces the reported errors in three of the four tested scenarios.

Smoothing	Perpendicular		Parallel		One-to-Many		Grid	
	D_{cost}	D_{target}	D_{cost}	D_{target}	D_{cost}	D_{target}	D_{cost}	D_{target}
✗	0.101	0.249	**0.001**	**0.000**	0.051	0.288	0.007	0.031
✓	**0.038**	**0.006**	0.003	0.001	**0.001**	**0.004**	**0.007**	**0.008**

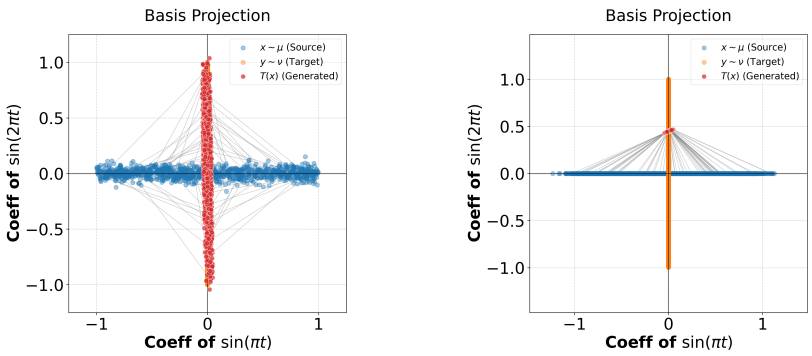


Figure 2: **Smoothing directions in the effective two-mode space. Left:** Smoothing the second mode makes the source absolutely continuous in H_2. **Right:** Smoothing only the first mode leaves a singular kernel marginal; the trained map exhibits a spurious-solution pattern.

5.2 Characterization of Smoothing Operator

Theorem 4.1 states that smoothing yields a regular source exactly when its projection onto the covariance kernel is regular. This provides a criterion for choosing the smoothing directions that make the source measure regular. We assess how this criterion affects HiSNOT in the effective two-mode space $H_2 = \mathrm{span}\{\sin(\pi t), \sin(2\pi t)\}$. The theorem also applies when these finite-rank covariances are extended by zero on $H_2^\perp$ to the full L^2 space; their kernels then additionally contain $H_2^\perp$. Appendix E.3 explains the resulting full-space condition and the joint-projection condition.

To empirically validate this, we compare (1) theoretically appropriate Gaussian smoothing against (2) inappropriate Gaussian smoothing using the Perpendicular dataset. In this setting, the source measure μ has a regular marginal along $\sin(\pi t)$ and is concentrated at zero along $\sin(2\pi t)$. We evaluate two Gaussian measures:

- **Appropriate smoothing** (γ_1) such that $\mathrm{Ker}(Q_1) = \mathrm{span}\{\sin(\pi t)\}$. This operator avoids smoothing only in the direction where μ is already regular, thereby satisfying the theorem.
- **Inappropriate smoothing** (γ_2) such that $\mathrm{Ker}(Q_2) = \mathrm{span}\{\sin(2\pi t)\}$. This operator fails to apply smoothing to a direction where μ is singular, thus failing to produce a regular measure.

Figure 2 presents the qualitative comparison of these two smoothings. With appropriate smoothing (left), the HiSNOT model successfully captures the target data distribution, as observed in the experiments with smoothing on all retained modes (Figure 1). However, under the inappropriate smoothing (right), the HiSNOT exhibits the spurious solution patterns. These results illustrate the finite-dimensional kernel criterion in Theorem 4.1. They do not test an unrestricted infinite-dimensional converse. For these diagonal two-mode examples, smoothing must include the unsmoothed singular direction.

5.3 Time Series Imputation Benchmarks

We evaluate HiSNOT with annealed smoothing on unpaired time series imputation benchmarks. Time-series data is one of the most widely used benchmarks for machine learning models operating on functional data [14, 29]. For time-series data, (possibly irregular) observations are regarded as discrete samples from underlying continuous processes defined in a Hilbert space. Here, we treat the imputation task as an optimal transport problem between the distribution of partially observed sequences (source) and fully observed paths (target).

Real-world time series are commonly modeled as lying near a low-dimensional manifold, and partial observation further restricts the data to a lower-dimensional structure. The source distribution of partially observed sequences is therefore expected to be singular, which makes standard SNOT prone to spurious solutions. HiSNOT with annealed smoothing is designed for exactly this regime. This experiment assesses its practical utility on high-dimensional, real-world functional data given as discretized observations; a rigorous discussion of population regularity for such data is deferred to Appendix H.2. We report mean squared error (MSE) and mean absolute error (MAE), following Wang et al. [35].

Table 2: **Unpaired time-series imputation performance (MSE and MAE)** across five datasets. Results are averaged over missing ratios of 0.5 and 0.7. The **bold** and underlined values denote the best and second-best results. All baseline results are taken from Wang et al. [35].

Datasets	**ETTh1**		**ETTh2**		**ETTm1**		**ETTm2**		**Exchange**	
Methods	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
Transformer [33]	0.338	0.410	0.282	0.383	0.085	0.196	0.056	0.161	0.320	0.264
DLinear [40]	0.187	0.307	0.138	0.262	0.131	0.250	0.107	0.227	0.288	0.238
TimesNet [38]	0.343	0.415	0.147	0.278	0.086	0.209	0.093	0.220	0.372	0.308
FreTS [39]	0.233	0.355	0.159	0.278	0.061	0.168	0.040	0.136	0.229	0.189
PatchTST [27]	0.216	0.338	0.144	0.272	**0.055**	0.153	0.033	0.122	0.227	0.188
SCINet [21]	0.195	0.321	0.172	0.288	0.088	0.204	0.085	0.211	0.326	0.269
iTransformer [22]	0.206	0.321	0.125	0.237	0.068	0.173	0.043	0.142	0.087	0.070
SAITS [9]	0.286	0.358	0.236	0.303	0.079	0.188	0.059	0.160	1.007	0.832
CSDI [32]	0.187	0.303	0.107	0.271	0.156	0.208	0.161	0.150	0.134	0.150
Sinkhorn [26]	0.936	0.685	0.877	0.644	0.965	0.717	0.927	0.690	0.783	0.648
TDM [41]	0.991	0.749	0.998	0.745	1.003	0.756	0.998	0.748	0.969	0.801
PSW-I [35]	**0.160**	**0.262**	0.053	0.155	0.057	**0.146**	0.025	0.103	0.036	**0.030**
HiSNOT (Ours)	0.227	0.296	**0.050**	**0.149**	0.067	0.152	**0.022**	**0.092**	**0.006**	0.048

Table 2 shows the lowest MSE and MAE on ETTh2 and ETTm2, and the lowest MSE on Exchange. On Exchange MAE, PSW-I performs better (0.030 versus 0.048). Performance on the other datasets depends on the metric. Exchange MSE is 0.006 compared with 0.036 for the next best baseline, PWS-I, a sixfold reduction in the reported values. Compared to classical OT baselines such as Sinkhorn and TDM, HiSNOT provides a significant improvement in accuracy, achieving better or comparable to previous state-of-the-art methods [35]. Detailed results for each missing ratio (0.5 and 0.7) are reported in Appendix G.5.

We additionally evaluate Gaussian-to-Exchange function-level generation using sliding windows and compare $1{,}000$ generated windows with $1{,}000$ target windows through W_2. HiSNOT has the lower reported central value in five of six window-size/stride settings, while SNOT has the lower value for window size 720 and stride 1. The complete results are in Appendix I. This finite-observation experiment does not establish resolution-independent numerical convergence.

Additional ablations in Appendix J test smoothing support and both schedule endpoints. In a controlled H_5 example, omitting a required smoothing direction gives $D_{\mathrm{target}} = 0.2063$, while all three tested supports containing both required directions give 0.0249–0.0300. The initial- and terminal-level ablations yield $D_{\mathrm{target}} < 0.01$ at every tested positive setting. These results support robustness within the tested ranges; the regularity comparison for the support ablation is relative to H_5.

6 Conclusion

We introduced HiSNOT, a Hilbert-space extension of semi-dual neural optimal transport for functional data. Regular source measures restore pointwise inner-minimizer identifiability for admissible optimal potentials. Gaussian smoothing obeys a necessary kernel condition, with a converse under Cameron–Martin compatibility; injective trace-class covariance regularizes arbitrary sources. The induced exact optimal plans are consistent as smoothing vanishes. Finite-mode synthetic experiments, imputation benchmarks, and Exchange generation show the empirical utility and setting-dependent benefits of the approach.

Limitations. The analysis concerns static quadratic OT on separable Hilbert spaces and exact population optimizers. It establishes plan-level accumulation-point consistency, not convergence to a unique limiting map. A kernel marginal alone is insufficient for arbitrary infinite-dimensional degenerate smoothing. Finite-mode discretization, finite-sample neural optimization, statistical estimation of smoothing directions, and optimal noise schedules require separate guarantees. We do not prove resolution-independent convergence, dynamic OT or gradient-flow results, or extensions to general Banach or metric spaces. Appendix H gives operational interpretations and scope.

Acknowledgments and Disclosure of Funding

J.-H. Choi is partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25488663). J. Yoon and J. W. Choi were partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00349646, RS-2026-25506577), National Institute for Mathematical Sciences (NIMS) grant funded by the Korea government (MSIT) (No. S26120000), and BK21 FOUR Project (X-AI Driven Immersive Learning Innovation Research and Education Group). D. Kwon is partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00252516, No. RS-2024-00408003, No. RS-2026-25488663, and No. RS-2026-25613008), the POSCO Science Fellowship of POSCO TJ Park Foundation, and the Korea Institute for Advanced Study. This research was supported by the Yonsei University Research Fund of 2026-22-0251. J. W. Choi and D. Kwon thank the Center for Advanced Computation in KIAS for providing computing resources.

Competing interests. Other than the funding and support disclosed above, the authors report no financial relationships that could be perceived as influencing this work.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser, Basel, first edition, 2005. doi: 10.1007/b137080. URL https://link.springer.com/book/10.1007/b137080.

[2] Jean-David Benamou and Yann Brenier. A Computational Fluid Mechanics Solution to the Monge–Kantorovich Mass Transfer Problem. *Numerische Mathematik*, 84(3):375–393, 2000. doi: 10.1007/s002110050002. URL https://link.springer.com/article/10.1007/s002110050002.

[3] Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative Modeling through the Semi-dual Formulation of Unbalanced Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1840. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/84706cdfc192cd0351daf48f379847e6-Abstract-Conference.html.

[4] Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Analyzing and Improving Optimal-Transport-based Adversarial Networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1d051fb631f104cb2a621451f37676b9-Abstract-Conference.html.

[5] Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable Wasserstein Gradient Flow for Generative Modeling through Unbalanced Optimal Transport. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8629–8650. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/choi24a.html.

[6] Jaemoo Choi, Yongxin Chen, and Jaewoong Choi. Improving Neural Optimal Transport via Displacement Interpolation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/040d3b6af368bf71f952c18da5713b48-Abstract-Conference.html.

[7] Jaemoo Choi, Jaewoong Choi, and Dohyun Kwon. Overcoming Spurious Solutions in Semi-Dual Neural Optimal Transport: A Smoothing Approach for Learning the Optimal Transport Plan. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10644–10665. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/choi25a.html.

[8] Raymond Chu, Jaewoong Choi, and Dohyun Kwon. Rate-Optimal Noise Annealing in Semi-Dual Neural Optimal Transport: Tangential Identifiability, Off-Manifold Ambiguity, and Guaranteed Recovery. *arXiv preprint arXiv:2602.04110*, 2026. doi: 10.48550/arXiv.2602.04110. URL https://arxiv.org/abs/2602.04110.

[9] Wenjie Du, David Cote, and Yan Liu. SAITS: Self-Attention-based Imputation for Time Series. *Expert Systems with Applications*, 219:119619, 2023. doi: 10.1016/j.eswa.2023.119619. URL https://doi.org/10.1016/j.eswa.2023.119619.

[10] Jiaojiao Fan, Shu Liu, Shaojun Ma, Yongxin Chen, and Hao-Min Zhou. Scalable Computation of Monge Maps with General Costs. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. URL https://openreview.net/forum?id=rEnGR3VdDW5.

[11] Jiaojiao Fan, Shu Liu, Shaojun Ma, Hao-Min Zhou, and Yongxin Chen. Neural Monge Map estimation and its applications. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=2mZSlQscj3. Featured Certification.

[12] Milena Gazdieva, Petr Mokrov, Litu Rout, Alexander Korotin, Andrey Kravchenko, Alexander Filippov, and Evgeny Burnaev. An Optimal Transport Perspective on Unpaired Image Super-Resolution. *Journal of Optimization Theory and Applications*, 207(2):40, 2025. doi: 10.1007/s10957-025-02781-7. URL https://link.springer.com/article/10.1007/s10957-025-02781-7.

[13] Martin Hairer. An Introduction to Stochastic PDEs. *arXiv preprint arXiv:0907.4178*, 2009. doi: 10.48550/arXiv.0907.4178. URL https://arxiv.org/abs/0907.4178v2. Version 2, 2023; see Theorem 4.44.

[14] Siyuan Jiang, Yihan Hu, Wenjie Li, and Pengcheng Zeng. DeepFRC: An End-to-End Deep Learning Model for Functional Registration and Classification. In *International Conference on Learning Representations*, volume 2026, pages 19002–19023, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/204810b414d4971c4bfccf7eb3422ecc-Abstract-Conference.html.

[15] Leonid V. Kantorovich. On a Problem of Monge. *Uspekhi Matematicheskikh Nauk*, 3(2(24)): 225–226, 1948. In Russian.

[16] Minjung Kim, Yusuke Hioka, and Michael Witbrock. Neural Fourier Modelling: A Highly Compact Approach to Time-Series Analysis. *arXiv preprint arXiv:2410.04703*, 2024. doi: 10.48550/arXiv.2410.04703. URL https://arxiv.org/abs/2410.04703.

[17] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Kernel Neural Optimal Transport. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zuc_MHtUma4.

[18] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://www.jmlr.org/papers/v24/21-1524.html.

[19] Donggyu Lee, Taekyung Lee, Jaewoong Choi, and Myungjoo Kang. ISUM: Inverse Problem Solver via Unbalanced Optimal Transport Map. In *NeurIPS 2025 Workshop on Structured Probabilistic Inference and Generative Modeling*, 2025. URL https://openreview.net/forum?id=xTgPer5mxg.

[20] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf/53c47f849d1cd4d21b865caf7d774e07a5c42aa4.pdf.

[21] Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. SCINet: Time Series Modeling and Forecasting with Sample Convolution and Interaction. In *Advances in Neural Information Processing Systems*, volume 35, pages 5816–5828. Curran Associates, Inc., 2022. doi: 10.52202/068431-0421. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/266983d0949aed78a16fa4782237dea7-Abstract-Conference.html.

[22] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *International Conference on Learning Representations*, volume 2024, pages 11116–11140, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

[23] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6672–6681. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/makkuva20a.html.

[24] Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M. Solomon, and Evgeny Burnaev. Large-Scale Wasserstein Gradient Flows. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper_files/paper/2021/hash/810dfbbebb17302018ae903e9cb7a483-Abstract.html.

[25] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mémoires de Mathématique et de Physique de l'Académie Royale des Sciences*, pages 666–704, 1781. Volume for 1781, printed in 1784.

[26] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing Data Imputation using Optimal Transport. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7130–7140. PMLR, 2020. URL https://proceedings.mlr.press/v119/muzellec20a.html.

[27] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vT0col.

[28] Md Ashiqur Rahman, Manuel A. Florez, Anima Anandkumar, Zachary E. Ross, and Kamyar Azizzadenesheli. Generative Adversarial Neural Operators. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=X1VzbBU6xZ.

[29] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005. doi: 10.1007/b98888. URL https://link.springer.com/book/10.1007/b98888.

[30] Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative Modeling with Optimal Transport Maps. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5JdLZg346Lw.

[31] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Cham, 2015. doi: 10.1007/978-3-319-20828-2. URL https://link.springer.com/book/10.1007/978-3-319-20828-2.

[32] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. In *Advances in Neural Information Processing Systems*, volume 34, pages 24804–24816. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[34] 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.

[35] Hao Wang, Zhengnan Li, Haoxuan Li, Xu Chen, Mingming Gong, Bin Chen, and Zhichao Chen. Optimal Transport for Time Series Imputation. In *International Conference on Learning Representations*, volume 2025, pages 828–852, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/033394ac055e3c47bfc3c967ffcfee74-Abstract-Conference.html.

[36] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016. doi: 10.1146/annurev-statistics-041715-033624. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-041715-033624.

[37] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

[38] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

[39] Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain MLPs are More Effective Learners in Time Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 36, pages 76656–76679. Curran Associates, Inc., 2023. doi: 10.52202/075280-3349. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f1d16af76939f476b5f040fd1398c0a3-Abstract-Conference.html.

[40] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9): 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317. URL https://ojs.aaai.org/index.php/AAAI/article/view/26317.

[41] He Zhao, Ke Sun, Amir Dezfouli, and Edwin V. Bonilla. Transformed Distribution Matching for Missing Value Imputation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42159–42186. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhao23h.html.

[42] Jiacheng Zhu, Aritra Guha, Dat Do, Mengdi Xu, XuanLong Nguyen, and Ding Zhao. Functional Optimal Transport: Regularized Map Estimation and Domain Adaptation for Functional Data. *Journal of Machine Learning Research*, 25(276):1–49, 2024. URL https://www.jmlr.org/papers/v25/22-0217.html.

A Gaussian Null Sets and Gâteaux Differentiability

We record the definitions used in Section 3.2.

Definition A.1 (Gaussian null set)**.** Let H be a separable Hilbert space, and let $\mathcal{P}(H)$ denote the set of Borel probability measures on H.

(i) A measure $\gamma \in \mathcal{P}(H)$ is called a *Gaussian measure* if, for every $h \in H$, the pushforward measure
$$(\langle h, \cdot \rangle_H)_{\#}\gamma$$
is a Gaussian measure on $\mathbb{R}$.

(ii) A Gaussian measure γ is called *nondegenerate* if, for every $h \in H \setminus \{0\}$, the one-dimensional Gaussian measure
$$(\langle h, \cdot \rangle_H)_{\#}\gamma$$
has strictly positive variance. Equivalently, for every $h \in H \setminus \{0\}$, there exist $m_h \in \mathbb{R}$ and $\sigma_h > 0$ such that, for every interval $(a, b) \subset \mathbb{R}$,
$$\gamma\big(\{x \in H : \, a < \langle h, x \rangle_H < b\}\big) = \frac{1}{\sqrt{2\pi\sigma_h^2}} \int_a^b \exp\left(-\frac{(t - m_h)^2}{2\sigma_h^2}\right) \mathrm{d}t.$$

(iii) A Borel set $N \in \mathcal{B}(H)$ is called a *Gaussian null set* if
$$\gamma(N) = 0$$
for every nondegenerate Gaussian measure γ on H.

A probability measure $\mu \in \mathcal{P}(H)$ is called regular if it assigns zero mass to every Gaussian null set; see Definition 3.1.

Because standard calculus is insufficient for infinite-dimensional spaces, we employ the *Gâteaux differential*. This generalizes the directional derivative to Banach spaces, providing the necessary mathematical tools to rigorously capture functional variations.

Definition A.2 (Gâteaux differentiability)**.** Let H be a Hilbert space, and let $J : H \to \mathbb{R}$ be a real-valued functional. For $x, y \in H$, the directional derivative of J at x in the direction y is defined by
$$\partial_y J(x) := \lim_{\varepsilon \to 0} \frac{J(x + \varepsilon y) - J(x)}{\varepsilon},$$
provided the limit exists.

We say that J is *Gâteaux differentiable* at x if there exists a continuous linear functional $DJ(x) \in H^*$ such that
$$DJ(x)[y] = \partial_y J(x) \qquad \text{for every } y \in H.$$
Since H is a Hilbert space, we identify $DJ(x)$ with its Riesz representative, denoted by $\nabla J(x) \in H$, so that
$$DJ(x)[y] = \langle \nabla J(x), y \rangle_H \qquad \text{for every } y \in H.$$

B Proof of Theorem 3.2

Admissible dual potentials. Throughout this proof, a dual potential is represented by a Borel c-concave function $V : H \to [-\infty, \infty)$ with $V \in L^1(\nu)$ and $u := V^c \in L^1(\mu)$, where
$$u(x) = \inf_{y \in H} \{c(x, y) - V(y)\}, \qquad c(x, y) = \tfrac{1}{2}\|x - y\|_H^2.$$
The transform is defined using this representative on all of H. A Kantorovich potential is an admissible potential attaining $\int_H u \, \mathrm{d}\mu + \int_H V \, \mathrm{d}\nu = C(\mu, \nu)$. The objective $\mathcal{L}(V, T)$ is interpreted with $c(x, T(x)) - V(T(x))$ as a single integrand. Its negative part is integrable because it is bounded below by $u \in L^1(\mu)$; thus $\mathcal{L}(V, T) \in \mathbb{R} \cup \{+\infty\}$ is well defined for every measurable map $T : H \to H$.

Proposition B.1 (Optimal maps and dual attainment)**.** *Let H be a separable Hilbert space, $\mu \in \mathcal{P}_2^r(H)$, and $\nu \in \mathcal{P}_2(H)$. For the quadratic cost, there is a unique optimal plan,*

$$\pi^\star = (\mathrm{Id}, T^\star)_{\#}\mu,$$

where $T^\star \in L^2(H, \mu; H)$ is unique μ-almost everywhere. An admissible Kantorovich potential exists. For every such potential $V^\star$, with $u = (V^\star)^c$, one has

$$u(x) + V^\star(T^\star(x)) = c(x, T^\star(x)) \quad \textit{for } \mu\textit{-almost every } x. \tag{16}$$

Proof. Existence and uniqueness of the optimal map follow from Ambrosio et al. [1, Theorem 6.2.10]. Dual attainment follows from Ambrosio et al. [1, Theorem 6.1.5], whose integrability conditions hold because $c(x, y) \le \|x\|_H^2 + \|y\|_H^2$ and both marginals have finite second moments. For any dual maximizer, the nonnegative slack $c(x, y) - u(x) - V^\star(y)$ has integral zero under $\pi^\star$. Consequently it vanishes $\pi^\star$-almost everywhere, which gives the stated identity. In particular,

$$T^\star(x) \in \arg\min_{y \in H} \{c(x, y) - V^\star(y)\} \quad \text{for } \mu\text{-almost every } x. \tag{17}$$

□

B.1 Uniqueness of the pointwise minimizer

Proof of Theorem 3.2(i). Write $q(x) = \frac{1}{2}\|x\|_H^2$, $u = (V^\star)^c$, and

$$\Phi(x) := q(x) - u(x) = \sup_{y:\, V^\star(y) > -\infty} \{\langle x, y\rangle_H - q(y) + V^\star(y)\}.$$

Choose $x_0, y_0 \in H$ such that $u(x_0)$ and $V^\star(y_0)$ are finite; such points exist by admissibility. For every integer $n > \|y_0\|_H$, define the truncated convex function

$$\Phi_n(x) := \sup_{\substack{\|y\|_H \le n \\ V^\star(y) > -\infty}} \{\langle x, y\rangle_H - q(y) + V^\star(y)\}.$$

The term corresponding to y_0 is a finite affine lower bound. For every candidate y, the affine term is bounded above by $\Phi(x_0) + n\|x - x_0\|_H$. Hence Φ_n is finite everywhere. Moreover, all its affine slopes have norm at most n, so

$$|\Phi_n(x) - \Phi_n(z)| \le n\|x - z\|_H \qquad (x, z \in H).$$

By Ambrosio et al. [1, Theorem 6.2.3], Φ_n is Gâteaux differentiable outside a Gaussian null set N_n. The countable union $N := \bigcup_{n > \|y_0\|_H} N_n$ is Gaussian null, and therefore $\mu(N) = 0$.

Fix $x \notin N$ for which $u(x)$ is finite and Eq. 16 holds. Suppose that y_1, y_2 minimize $c(x, y) - V^\star(y)$. Choose an integer $n > \|y_0\|_H$ containing both y_1 and y_2 in its closed ball. Each of their affine terms equals $\Phi(x)$, so $\Phi_n(x) = \Phi(x)$. Thus, for $i = 1, 2$,

$$\Phi_n(z) \ge \Phi_n(x) + \langle z - x, y_i\rangle_H \qquad (z \in H).$$

Differentiating along positive and negative multiples of each direction shows that $y_i = \nabla\Phi_n(x)$. Therefore $y_1 = y_2$. Since $T^\star(x)$ is a minimizer by Eq. 17, the argmin is exactly $\{T^\star(x)\}$. This establishes the claim without requiring bounded target support or differentiability of the untruncated potential u. □

B.2 Uniqueness of the integrated inner minimizer

Proof of Theorem 3.2(ii). For every measurable $T : H \to H$, define the nonnegative slack

$$d_T(x) := c(x, T(x)) - V^\star(T(x)) - (V^\star)^c(x)$$

on the full-measure set where $(V^\star)^c$ is finite. Admissibility and dual optimality give

$$\mathcal{L}(V^\star, T) - C(\mu, \nu) = \int_H d_T(x)\, \mu(\mathrm{d}x) \ge 0. \tag{18}$$

Equation 16 implies that $T^\star$ attains this lower bound. If T_{rec} is an inner minimizer, then $\int_H d_{T_{\mathrm{rec}}}\, \mathrm{d}\mu = 0$. Hence $d_{T_{\mathrm{rec}}} = 0$ μ-almost everywhere, so $T_{\mathrm{rec}}(x)$ is a pointwise minimizer there. Part (i) yields $T_{\mathrm{rec}} = T^\star$ μ-almost everywhere. □

Scope of the ideal objective. The argument optimizes over the admissible dual class and measurable transport maps containing $T^\star$. It is an identifiability result for the distributional objective; it does not assert exact recovery by a fixed neural architecture or an empirical optimization algorithm. For completeness, $(V^\star, T^\star)$ also satisfies the saddle inequalities in this ideal class: the preceding proof gives $\mathcal{L}(V^\star, T) \geq C(\mu, \nu)$ for every T, whereas $T^\star_\# \mu = \nu$ implies $\mathcal{L}(V, T^\star) = C(\mu, \nu)$ for every admissible V.

C Proof of Theorem 4.1

Use the orthonormal eigenbasis (e_k) of H from Theorem 4.1, with $Qe_k = \lambda_k e_k$, $\lambda_k \geq 0$, and $\sum_k \lambda_k < \infty$. All sums run over the available eigenbasis indices, a finite set when H is finite-dimensional. For this appendix section only, abbreviate $K = \mathrm{Ker}(Q)$, $S = K^\perp$, $\pi_K = \pi_{\mathrm{Ker}(Q)}$, and $\mu_K = (\pi_K)_\# \mu$; π_S denotes the orthogonal projection onto S. Then $H = S \oplus K$, and the Gaussian $\gamma_Q = \mathcal{N}(0, Q)$ is concentrated on S and nondegenerate when considered as a measure on S. We identify H with $S \times K$ through the isometry $(h, k) \mapsto h + k$. Its Cameron–Martin space is

$$H_Q = \mathrm{Ran}(Q^{1/2}) = \left\{ h \in S : \sum_{k:\lambda_k > 0} \frac{|\langle h, e_k \rangle_H|^2}{\lambda_k} < \infty \right\}, \qquad \|h\|^2_{H_Q} = \sum_{k:\lambda_k > 0} \frac{|\langle h, e_k \rangle_H|^2}{\lambda_k}.$$

This is a Borel subspace of H and a Hilbert space in its own norm. Its closure in the ambient norm is S; this closure must not be substituted for H_Q in a translation argument. For zero-dimensional factors, the unique probability measure is understood to be Gaussian and nondegenerate, consistently with the definition using nonzero continuous linear functionals.

Lemma C.1 (Cameron–Martin translations)**.** *Let $(\gamma_Q)_h(A) = \gamma_Q(A - h)$ for $h \in H$. Then $(\gamma_Q)_h$ and γ_Q are equivalent if $h \in H_Q$, and are mutually singular if $h \notin H_Q$. In particular, $(\gamma_Q)_h \ll \gamma_Q$ if and only if $h \in H_Q$.*

Proof. This is the Cameron–Martin theorem; see Hairer [13, Theorem 4.44]. For completeness, its spectral argument also shows why the ambient closure is insufficient. If $h \notin S$, the measures are supported on the disjoint affine spaces $S + h$ and S, so they are singular. For $h \in H_Q$, let $h_k = \langle h, e_k \rangle_H$ and consider, for $x \sim \gamma_Q$,

$$M_h(x) = \lim_{n \to \infty} \sum_{\substack{k \leq n \\ \lambda_k > 0}} \frac{h_k \langle x, e_k \rangle_H}{\lambda_k}.$$

The sum converges in $L^2(\gamma_Q)$ and almost surely, since its independent centered Gaussian terms have summable variances. Its distribution is $\mathcal{N}(0, \|h\|^2_{H_Q})$. The positive density

$$R_h(x) = \exp\big(M_h(x) - \tfrac{1}{2}\|h\|^2_{H_Q}\big)$$

has integral one. Computing its integrals against finite-coordinate characteristic functions shows that $R_h \gamma_Q$ has the same coordinate laws as $(\gamma_Q)_h$. Those coordinates generate the Borel sigma-algebra on S, so $R_h \gamma_Q = (\gamma_Q)_h$. Positivity gives equivalence.

Finally suppose $h \in S \setminus H_Q$. Then $a_n = \sum_{\substack{k \leq n \\ \lambda_k > 0}} h_k^2 / \lambda_k \to \infty$. For $a_n > 0$, the functional

$$\ell_n(x) = a_n^{-1/2} \sum_{\substack{k \leq n \\ \lambda_k > 0}} \frac{h_k \langle x, e_k \rangle_H}{\lambda_k}$$

has distribution $\mathcal{N}(0, 1)$ under γ_Q and $\mathcal{N}(\sqrt{a_n}, 1)$ under $(\gamma_Q)_h$. For $B_n = \{x : \ell_n(x) > \sqrt{a_n}/2\}$, therefore, $\gamma_Q(B_n) \to 0$ and $(\gamma_Q)_h(B_n) \to 1$. Choose a subsequence along which both error probabilities are summable. The Borel–Cantelli lemma implies that the limit inferior of these events has measure zero under γ_Q and one under $(\gamma_Q)_h$. This proves singularity. □

Lemma C.2 (Gaussian sections and a regular factor)**.** *Let γ be a nondegenerate Gaussian probability measure on S and $\eta \in \mathcal{P}^r(K)$. Then $\gamma \otimes \eta$ is regular on $S \oplus K$. More precisely, for any Gaussian-null Borel set $A \subset S \oplus K$, the set*

$$N_A = \{k \in K : \gamma(A_k) > 0\}, \qquad A_k = \{h \in S : (h, k) \in A\},$$

is Gaussian null in K.

Proof. The function $k \mapsto \gamma(A_k)$ is Borel measurable. For every nondegenerate Gaussian Γ_K on K, $\gamma \otimes \Gamma_K$ is a nondegenerate Gaussian on $S \oplus K$. Hence, by the definition of Gaussian nullity and Fubini's theorem,

$$0 = (\gamma \otimes \Gamma_K)(A) = \int_K \gamma(A_k)\, \Gamma_K(\mathrm{d}k).$$

Nonnegativity gives $\Gamma_K(N_A) = 0$. Since Γ_K was arbitrary, N_A is Gaussian null in K. Regularity of η then yields $\eta(N_A) = 0$, and another application of Fubini gives $(\gamma \otimes \eta)(A) = 0$. □

Lemma C.3 (Disintegration after Gaussian convolution)**.** *Let $\mu \in \mathcal{P}(S \oplus K)$ and $\mu_K = (\pi_K)_\# \mu$. There is a Borel probability kernel $k \mapsto \mu_k \in \mathcal{P}(S)$ such that $\mu(\mathrm{d}h, \mathrm{d}k) = \mu_k(\mathrm{d}h)\mu_K(\mathrm{d}k)$. With γ_Q regarded as a probability measure on S,*

$$(\gamma_Q * \mu)(\mathrm{d}h, \mathrm{d}k) = (\gamma_Q * \mu_k)(\mathrm{d}h)\mu_K(\mathrm{d}k).$$

Proof. Existence of the disintegration follows because S and K are Polish spaces; see Ambrosio et al. [1, Theorem 5.3.1]. Let $X = (X_S, X_K) \sim \mu$ and $G \sim \gamma_Q$ be independent. Since G takes values in S, adding it leaves the K-coordinate unchanged:

$$\pi_K(X + G) = X_K \quad \text{almost surely,} \qquad (\pi_K)_\#(\gamma_Q * \mu) = \mu_K.$$

Thus the smoothed measure has the same kernel marginal as μ.

For a bounded Borel function f on $S \times K$, independence and disintegration give

$$\mathbb{E} f(X_S + G, X_K) = \int_K \int_S \int_S f(h + g, k)\, \gamma_Q(\mathrm{d}g)\mu_k(\mathrm{d}h)\mu_K(\mathrm{d}k).$$

For each k, the two inner integrals are exactly integration against the convolution $\gamma_Q * \mu_k$:

$$\int_S \int_S f(h + g, k)\, \gamma_Q(\mathrm{d}g)\mu_k(\mathrm{d}h) = \int_S f(h, k)\, (\gamma_Q * \mu_k)(\mathrm{d}h).$$

Substituting this equality above shows that the conditional law of the smoothed S-coordinate given $X_K = k$ is $\gamma_Q * \mu_k$ for μ_K-almost every k. Since the identity holds for every bounded Borel f, it proves the claimed disintegration with the unchanged marginal μ_K. □

Lemma C.4 (Nondegenerate smoothing)**.** *If $\mathrm{Ker}(Q) = \{0\}$, then $\gamma_Q * \mu \in \mathcal{P}^r(H)$ for every $\mu \in \mathcal{P}(H)$.*

Proof. For each $x \in H$, the translated measure $\mathcal{N}(x, Q)$ is a nondegenerate Gaussian on the whole space H. If A is Gaussian null, then $\mathcal{N}(x, Q)(A) = 0$ for every x, and

$$(\gamma_Q * \mu)(A) = \int_H \mathcal{N}(x, Q)(A)\, \mu(\mathrm{d}x) = 0.$$

This argument requires no equivalence between translated Gaussian measures. □

Proof of Theorem 4.1. **Part (i): necessity without an additional assumption.** Suppose $\gamma_Q * \mu$ is regular. If $B \subset K$ is Gaussian null, then $A = \pi_K^{-1}(B)$ is Gaussian null in H: the projection of every nondegenerate Gaussian on H is nondegenerate on K. Since the added noise has zero K-coordinate,

$$\mu_K(B) = (\gamma_Q * \mu)(\pi_K^{-1}(B)) = 0.$$

Thus μ_K is regular, proving necessity.

Part (ii): equivalence under Cameron–Martin compatibility. Assume the weighted coefficient condition in equation (14). For every $x \in H$, the projection $\pi_S x$ lies in S and has the same coefficients as x along every e_k with $\lambda_k > 0$. The spectral description of H_Q therefore gives the exact equivalence

$$\pi_S x \in H_Q \quad \Longleftrightarrow \quad \sum_{k:\lambda_k > 0} \frac{|\langle x, e_k \rangle_H|^2}{\lambda_k} < \infty.$$

Thus the assumption is precisely $\mu(\pi_S^{-1}(H_Q)) = 1$. It imposes no restriction on the source coefficients with $\lambda_k = 0$ and requires only almost-sure finiteness, not a finite expectation of the sum. The forward

implication in equation (15) is part (i). For the reverse implication, suppose $\mu_K \in \mathcal{P}^r(K)$. The spectral description of H_Q above shows that it is Borel. Disintegration therefore gives

$$1 = \mu(\pi_S^{-1}(H_Q)) = \int_K \mu_k(H_Q)\, \mu_K(\mathrm{d}k),$$

so $\mu_k(H_Q) = 1$ for μ_K-almost every k. Let $A \subset H$ be Gaussian null. Lemma C.2, applied to γ_Q on S, gives $\gamma_Q(A_k) = 0$ for μ_K-almost every k. Thus the Borel set

$$K_0 = \{k \in K : \mu_k(H_Q) = 1 \text{ and } \gamma_Q(A_k) = 0\}$$

has $\mu_K(K_0) = 1$. For each $k \in K_0$, Lemma C.1 gives $\gamma_Q(A_k - h) = 0$ for every $h \in H_Q$. Consequently Lemma C.3 yields

$$\begin{aligned}(\gamma_Q * \mu)(A) &= \int_K \int_S \gamma_Q(A_k - h)\, \mu_k(\mathrm{d}h)\mu_K(\mathrm{d}k) \\ &= 0.\end{aligned}$$

Indeed, for each $k \in K_0$ the inner integrand vanishes for every $h \in H_Q$, and $\mu_k(H_Q) = 1$; the complement of K_0 is μ_K-null. Hence $\gamma_Q * \mu$ is regular, proving the reverse implication and therefore the equivalence. The conditional source laws are used only on a μ_K-full set; no statement about their versions under an unrelated Gaussian measure is required. If Q has finite rank, S is finite-dimensional and $H_Q = S$, so the additional source condition is automatic.

Part (iii): injective covariance. For $K = \{0\}$, Lemma C.4 gives regularity for every μ by integrating the nondegenerate Gaussian translates $\mathcal{N}(x, Q)$. This does not invoke the additional assumption in part (ii). □

The Cameron–Martin compatibility assumption in part (ii) is a sufficient hypothesis for the converse, not a necessary condition for regularity of the convolution.

Corollary C.5 (Finite-rank covariance)**.** *If Q has finite rank, then for every $\mu \in \mathcal{P}(H)$,*

$$\gamma_Q * \mu \in \mathcal{P}^r(H) \quad \Longleftrightarrow \quad (\pi_{\mathrm{Ker}(Q)})_{\#}\mu \in \mathcal{P}^r(\mathrm{Ker}(Q)).$$

In particular, this equivalence holds for any covariance on a finite-dimensional ambient space.

Proof. Apply part (ii) of the theorem with $\mathrm{Ran}(Q^{1/2}) = S$. □

Positive smoothing amplitudes. For every $\varepsilon > 0$, $\mathrm{Ker}(\varepsilon^2 Q) = K$ and $\mathrm{Ran}((\varepsilon^2 Q)^{1/2}) = H_Q$. Thus the same sufficient assumptions apply to $\mu * \mathcal{N}(0, \varepsilon^2 Q)$ at every positive amplitude. When $\mu \in \mathcal{P}_2(H)$, finite trace also ensures that the smoothed measure has finite second moment.

C.1 Why kernel marginal regularity alone is insufficient

The additional assumption in the degenerate-noise converse cannot be omitted for general infinite-rank covariance, even when the kernel is one-dimensional and the source has finite moments of every order.

Let $H = \ell^2(\mathbb{N}) \oplus \mathbb{R}$, write points as (x, u), and set

$$v = (1/n)_{n\geq 1}, \qquad Q(x, u) = ((x_n/n^2)_{n\geq 1}, 0).$$

This is a positive self-adjoint trace-class covariance with $\mathrm{Tr}(Q) = \sum_n n^{-2}$ and $K = \{0\} \oplus \mathbb{R}$. For $U \sim \mathcal{N}(0, 1)$, let $X = (U^2 v, U)$ and $\mu = \mathrm{Law}(X)$. Then $\mu_K = \mathcal{N}(0, 1)$ on K is regular, and

$$\mathbb{E}\|X\|_H^2 = 3\sum_n n^{-2} + 1 < \infty.$$

Let (ξ_n) be independent standard Gaussian variables, independent of U. The random element $G = ((\xi_n/n)_{n\geq 1}, 0)$ is H-valued almost surely because its expected squared norm is finite, and has law γ_Q. Hence $\rho = \gamma_Q * \mu$ is the law of

$$X + G = \big(((U^2 + \xi_n)/n)_{n\geq 1}, U\big).$$

For $N \geq 1$ define the continuous linear functional $L_N(x,u) = N^{-1}\sum_{n=1}^{N} nx_n$ and the Borel set

$$A = \{(x,u) : \lim_{N\to\infty} L_N(x,u) = u^2\}.$$

The strong law of large numbers gives $\rho(A) = 1$. We show that A is Gaussian null, which proves that ρ is not regular.

Let Γ be any nondegenerate Gaussian on H, with arbitrary mean and covariance, and let $(Z,V) \sim \Gamma$. The last coordinate V is a nondegenerate real Gaussian. Set $Y_N = L_N(Z,V)$, so that each (Y_N, V) is jointly Gaussian. Suppose, toward a contradiction, that $\Gamma(A) > 0$. On A the sequence Y_N converges to V^2 and is eventually bounded. Taking a countable union of possible bounds and initial indices shows that there are $M > 0$, $\delta > 0$ and N_0 such that

$$\Gamma\big(A \cap \{\sup_{N\geq N_0} |Y_N| \leq M\}\big) \geq \delta.$$

In particular, $\Gamma(|Y_N| \leq M) \geq \delta$ for every $N \geq N_0$. Write $m_N = \mathbb{E}_\Gamma Y_N$ and $\sigma_N^2 = \mathrm{Var}_\Gamma(Y_N)$. When $\sigma_N > 0$, the Gaussian density bound gives

$$\delta \leq \frac{2M}{\sqrt{2\pi}\sigma_N}, \qquad \sigma_N \leq C := \frac{2M}{\delta\sqrt{2\pi}}.$$

For $|m_N| > M$, Chebyshev's inequality further gives $\delta \leq \sigma_N^2/(|m_N| - M)^2$. The same resulting bounds hold directly for a degenerate Y_N. Thus $|m_N| \leq M + C/\sqrt{\delta}$, and the tail sequence (Y_N) is bounded in the Hilbert space $L^2(\Gamma)$.

By the Hilbert-space Banach–Saks property, a subsequence (Y_{N_j}) has arithmetic means $B_J = J^{-1}\sum_{j=1}^{J} Y_{N_j}$ converging in $L^2(\Gamma)$ to a random variable Y. Each pair (B_J, V) is jointly Gaussian. Convergence in L^2 implies convergence of its mean vector and covariance matrix and of its joint distribution, so (Y,V) is jointly Gaussian too, possibly degenerate. On A, Cesaro convergence gives $B_J \to V^2$ pointwise. An almost surely convergent subsequence of the L^2 convergence therefore implies $Y = V^2$ on A up to a Γ-null set, and hence $\Gamma(Y = V^2) > 0$.

Gaussian regression writes $Y = aV + b + R$, where R is a centered Gaussian independent of V. If $\mathrm{Var}(R) > 0$, conditioning on V gives $\Gamma(Y = V^2) = 0$ because the conditional law of Y is continuous. If $\mathrm{Var}(R) = 0$, equality requires $V^2 - aV - b = 0$. This polynomial has at most two roots, to which the nondegenerate Gaussian V assigns probability zero. Both cases are contradictory. Thus $\Gamma(A) = 0$ for every nondegenerate Gaussian Γ on H, proving the claimed Gaussian nullity.

In this example,

$$H_Q = \{(h,0) : h \in \ell^2,\ \sum_n n^2 h_n^2 < \infty\}, \qquad (v,0) \notin H_Q, \quad (v,0) \in \overline{H_Q} = S.$$

The projected source $\pi_S X = (U^2 v, 0)$ therefore lies outside the Cameron–Martin space almost surely. Equivalently, the weighted sum in equation (14) is $\sum_n n^2(U^2/n)^2 = \infty$ whenever $U \neq 0$. Although the kernel marginal is regular, a nonlinear relation with that coordinate survives the smoothing in the infinitely many noisy coordinates.

D Plan-Level Consistency: Proof of Theorem 4.2

We use narrow convergence with respect to the norm topology of H and the product norm on $H \times H$ throughout this section.

Lemma D.1 (Convergence of Gaussian smoothing)**.** *Let $\mu \in \mathcal{P}_2(H)$, and let Q be a positive self-adjoint trace-class covariance operator. Then*

$$W_2^2\big(\mu * \mathcal{N}(0, \varepsilon^2 Q), \mu\big) \leq \varepsilon^2 \,\mathrm{Tr}(Q) \longrightarrow 0 \qquad (\varepsilon \downarrow 0).$$

Proof. For independent $X \sim \mu$ and $G \sim \mathcal{N}(0,Q)$, the pair $(X + \varepsilon G, X)$ is an admissible coupling. Its expected squared displacement is $\varepsilon^2 \mathbb{E}\|G\|_H^2 = \varepsilon^2 \,\mathrm{Tr}(Q)$. □

Lemma D.2 (Compactness and stability of couplings)**.** *Suppose that $\mu_k \to \mu$ in $W_2(H)$ and $\mu_k, \mu, \nu \in \mathcal{P}_2(H)$. Then $\bigcup_{k\in\mathbb{N}} \Pi(\mu_k, \nu)$ is relatively compact in $\mathcal{P}_2(H \times H)$ for the W_2 topology. If $\pi_k \in \Pi(\mu_k, \nu)$ converges narrowly to π, then $\pi \in \Pi(\mu, \nu)$ and $\pi_k \to \pi$ in $W_2(H \times H)$.*

Proof. The W_2 convergence of μ_k implies that the family $\{\mu_k\}_{k\in\mathbb{N}}$ is tight and has uniformly integrable second moments; see Ambrosio et al. [1, Proposition 7.1.5]. The same properties hold for the singleton $\{\nu\}$. Given compact sets $K, L \subset H$, every $\pi \in \Pi(\mu_k, \nu)$ satisfies

$$\pi\big((K \times L)^c\big) \le \mu_k(K^c) + \nu(L^c).$$

Thus the entire union of coupling sets is tight.

For $a, b \ge 0$ and $R > 0$,

$$(a+b)\mathbf{1}_{\{a+b>R^2\}} \le 2a\mathbf{1}_{\{a>R^2/2\}} + 2b\mathbf{1}_{\{b>R^2/2\}}.$$

Indeed, if just one threshold is exceeded, that variable is the larger one; if both are exceeded, the inequality is immediate. Applying this with $a = \|x\|_H^2$ and $b = \|y\|_H^2$ gives, uniformly over $\pi \in \Pi(\mu_k, \nu)$,

$$\begin{aligned}&\int_{\{\|(x,y)\|_{H\times H}>R\}} \|(x,y)\|_{H\times H}^2 \,\pi(\mathrm{d}x, \mathrm{d}y)\\ &\quad\le 2\int_{\{\|x\|_H>R/\sqrt{2}\}} \|x\|_H^2 \,\mu_k(\mathrm{d}x) + 2\int_{\{\|y\|_H>R/\sqrt{2}\}} \|y\|_H^2 \,\nu(\mathrm{d}y).\end{aligned}$$

Taking the supremum over k and letting $R \to \infty$ proves uniform integrability of the product second moments. Tightness and this bound imply relative W_2 compactness by the same proposition.

If π_k converges narrowly to π, testing against bounded continuous functions of either coordinate gives the marginals μ and ν. The uniform moment bound just established and narrow convergence give $W_2(\pi_k, \pi) \to 0$. □

Proof of Theorem 4.2. The preceding lemma gives relative W_2 compactness of the optimal plans $\pi_k^\star$. Let a subsequence converge in W_2 to π. Its marginals are μ and ν. The continuous cost $c(x, y) = \frac{1}{2}\|x - y\|_H^2$ has at most quadratic growth, so convergence in $W_2(H \times H)$ implies convergence of its integrals. Moreover, the triangle inequality gives

$$|W_2(\mu_k, \nu) - W_2(\mu, \nu)| \le W_2(\mu_k, \mu) \longrightarrow 0.$$

Along the chosen subsequence we therefore obtain

$$\int_{H\times H} c(x,y)\,\pi(\mathrm{d}x,\mathrm{d}y) = \lim_k \int_{H\times H} c(x,y)\,\pi_k^\star(\mathrm{d}x,\mathrm{d}y) = \lim_k \tfrac{1}{2}W_2^2(\mu_k,\nu) = \tfrac{1}{2}W_2^2(\mu,\nu).$$

Thus π is optimal. This argument applies to every W_2 accumulation point. □

Application to smoothing. If $\mu_{\varepsilon_k} := \mu * \mathcal{N}(0, \varepsilon_k^2 Q)$ is regular for every k and $\varepsilon_k \downarrow 0$, then Lemma D.1 verifies the required marginal convergence. Proposition B.1 supplies the unique optimal maps for these regularized sources, and the theorem applies to their induced plans. Regularity is used for this map representation; the compactness and optimal-plan stability argument itself requires only the stated W_2 convergence of the marginals.

E Examples of Spurious Solutions in Non-Regular Settings

These examples illustrate how the ideal semi-dual objective can have spurious inner minimizers when the source is non-regular. Regularity is a sufficient condition for identifiability, not a necessary condition for every individual transport problem. We use globally defined admissible potentials, with the inner minimization taken over all of H. Smoothing restores the stated guarantees under a sufficient condition of Theorem 4.1.

E.1 Setup and two singular transport problems

Let $H = L^2((-1, 1))$ and choose distinct positive integers k_1, k_2. Set $e_1(t) = \sin(k_1\pi t)$ and $e_2(t) = \sin(k_2\pi t)$, which are orthonormal in H. Write $U_i = \operatorname{span}\{e_i\}$, $I = [-1, 1]$, and let m be the uniform probability measure on I. The source is $\mu = \operatorname{Law}(Ae_1)$ with $A \sim m$. It is non-regular in H because it is concentrated on the proper closed subspace U_1, a Gaussian null set.

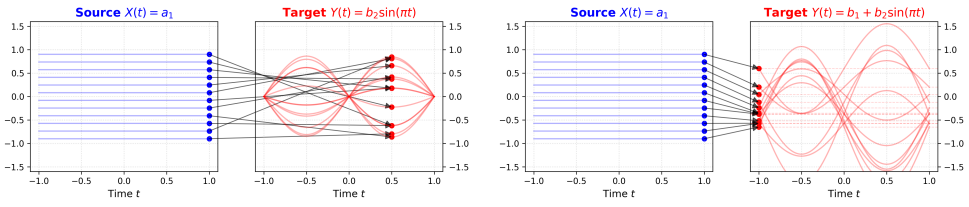


Figure 3: Illustrations of non-unique optimal transport maps in the perpendicular example.

In the perpendicular example, the target is $\nu_1 = \mathrm{Law}(Be_2)$ with $B \sim m$. In the one-to-many example, it is $\nu_2 = \mathrm{Law}(Be_1 + Re_2)$, where $B \sim m$ and the independent sign R is uniform on $\{-1, 1\}$. The first problem has many optimal Monge maps. The second has an optimal Kantorovich plan but no optimal deterministic transport map. Valid deterministic transports do exist in the second problem, and their costs can approach the Kantorovich optimum.

E.2 Details on Examples 1 and 2

E.2.1 Example 1: Perpendicular transport

For $x = ae_1$ and $y = be_2$, orthogonality gives $c(x, y) = \frac{1}{2}(a^2 + b^2)$. Consequently,

$$C(\mu, \nu_1) = \tfrac{1}{2} \int_I a^2 \, m(\mathrm{d}a) + \tfrac{1}{2} \int_I b^2 \, m(\mathrm{d}b) = \tfrac{1}{3}, \tag{19}$$

and every coupling is optimal. Each measure-preserving map $\psi : I \to I$ gives an optimal Monge map $T_\psi(ae_1) = \psi(a)e_2$, so the optimal map is not unique.

To exhibit spurious inner minimizers using the full-space objective, write $q(y) = \frac{1}{2}\|y\|_H^2$, let $\pi_{U_2^\perp}$ denote the orthogonal projection onto $U_2^\perp$, and choose

$$V_1(y) := q(y) - 2\|\pi_{U_2^\perp} y\|_H \qquad (y \in H).$$

This is c-concave: if $D = \{z \in U_2^\perp : \|z\|_H \le 2\}$ and $w(z) = q(z)$ on D, $w(z) = -\infty$ elsewhere, then

$$w^c(y) = q(y) - \sup_{z \in D} \langle z, y \rangle_H = q(y) - 2\|\pi_{U_2^\perp} y\|_H = V_1(y).$$

For $x = ae_1$, $a \in I$, and $p = \pi_{U_2^\perp} y$,

$$c(x, y) - V_1(y) = q(x) - \langle x, p \rangle_H + 2\|p\|_H \ge q(x) + (2 - |a|)\|p\|_H.$$

Equality at the minimum holds exactly when $p = 0$. Thus

$$V_1^c(x) = q(x), \qquad \arg\min_{y \in H} \{c(x, y) - V_1(y)\} = U_2 \qquad (x \in Ie_1).$$

Since $V_1 = q$ on the target line, both dual terms are integrable and their sum is $1/3$. Hence V_1 is an admissible Kantorovich potential. Every measurable map with $T(x) \in U_2$ for μ-almost every x minimizes

$$\inf_{T : H \to H} \mathcal{L}(V_1, T), \tag{20}$$

with value $C(\mu, \nu_1)$. Many of these maps fail the marginal constraint; for example, $T(x) = 0$ is an inner minimizer but $T_\# \mu = \delta_0 \ne \nu_1$. The flat family therefore consists of line-valued maps, rather than arbitrary H-valued maps.

E.2.2 Example 2: One-to-many transport

For any coupling between $x = ae_1$ and $y = be_1 + re_2$,

$$c(x, y) = \tfrac{1}{2}(a - b)^2 + \tfrac{1}{2}.$$

The plan which keeps $b = a$ and independently chooses $r = \pm 1$ with equal probabilities has cost $1/2$, the global lower bound. Thus $C(\mu, \nu_2) = 1/2$.

An optimal deterministic transport would have to preserve the first coordinate, hence have the form $T(ae_1) = ae_1 + s(a)e_2$ with $s(a) \in \{-1, 1\}$. Let $E = \{a : s(a) = 1\}$. The target marginal would require, for every Borel $B \subset I$,

$$m(E \cap B) = \tfrac{1}{2} m(B).$$

Taking $B = I$ yields $m(E) = 1/2$, while taking $B = E$ yields $m(E) = \frac{1}{2} m(E)$, a contradiction. No optimal deterministic transport exists.

This is a failure of attainment, not a positive gap between the Monge and Kantorovich infima. To see this, partition I into n equal cells. On the first half of each cell, map the first coordinate affinely onto the whole cell and choose sign $+1$; on the second half, do the same and choose sign -1. This defines a valid transport T_n to ν_2. Its first-coordinate displacement is at most $1/n$, so

$$\tfrac{1}{2} < C(T_n) := \int_H c(x, T_n(x))\, \mu(\mathrm{d}x) \leq \tfrac{1}{2} + \frac{1}{2n^2}.$$

Thus valid maps can approximate the optimal cost even though no one of them attains it.

For the semi-dual problem, choose the global potential

$$V_2(y) := \tfrac{1}{2} \|\pi_{U_2} y\|_H^2 \qquad (y \in H).$$

It is c-concave: $V_2 = w^c$ for $w = 0$ on $U_2^\perp$ and $w = -\infty$ elsewhere. For $x \in U_1$,

$$c(x, y) - V_2(y) = \tfrac{1}{2} \|x - \pi_{U_2^\perp} y\|_H^2, \qquad V_2^c(x) = 0.$$

On the target support $V_2 = 1/2$, so it is an admissible Kantorovich potential with dual value $1/2$. Its full-space argmin is

$$\operatorname*{arg\,min}_{y \in H} \{c(x, y) - V_2(y)\} = x + U_2.$$

Every measurable selector from these lines is an inner minimizer. In particular, selectors $T(ae_1) = ae_1 + s(a)e_2$ with $s(a) \in \{-1, 1\}$ lie on the target support and attain the optimal semi-dual value, yet fail its marginal constraint as shown above. This is a failure of exact inner-minimizer identifiability; it does not rule out approximation by other deterministic maps or learning a conditional plan with an explicitly stochastic model.

E.3 Revisiting the examples with smoothing

Let Q be an injective nonnegative self-adjoint trace-class covariance on H. For every $\varepsilon > 0$, the smoothed source $\mu_\varepsilon = \mu * \mathcal{N}(0, \varepsilon^2 Q)$ is regular by Theorem 4.1. Proposition B.1 and Theorem 3.2 therefore give a unique optimal map and a unique inner minimizer for each admissible optimal potential. Theorem 4.2 concerns the induced plans as $\varepsilon \downarrow 0$. In the one-to-many example, these plans can approach the splitting optimal plan without the maps converging to an optimal deterministic map for the original source.

The finite-mode experiments must be interpreted in their retained space. In $H_2 = \operatorname{span}\{e_1, e_2\}$, noise along e_2 leaves a regular source marginal on U_1 and restores regularity in H_2. Noise only along e_1 leaves the unsmoothed e_2 marginal atomic and does not restore regularity there. A perturbation supported on H_2 alone is not a regularizer of this source in the entire infinite-dimensional H.

E.4 Comparison with finite-dimensional results

In finite dimensions, the source non-charging condition studied by Choi et al. [7] is more general than absolute continuity. Our Gaussian-null regularity assumption supplies a sufficient Hilbert-space condition; the examples above do not establish its necessity for all transport problems. Extensions to weaker geometric source assumptions are beyond the present analysis.

F Related Works

Neural Optimal Transport (Neural OT) learns the optimal transport maps between the source and target distributions using neural networks. A prominent family of approaches, called Semi-dual Neural OT (SNOT), builds on the semi-dual formulation of the OT problem [3, 11, 23, 30]. These

methods differ in their transport map parameterization [11, 30], the strict convexity constraints on the potential via ICNNs [23], and the OT formulation [3]. SNOT has been successfully applied to various machine learning tasks in finite-dimensional Euclidean spaces, including generative modeling [3, 5, 11, 30] and inverse problems [12, 19].

Despite these successes, SNOT suffers from a fundamental limitation known as the **spurious solution problem**: not all solutions to the max-min objective $\mathcal{L}(V, T)$ (Eq. 11) recover the correct optimal transport map. Although the ground-truth Kantorovich potential and transport map $(V^\star, T^\star)$ are always the max-min solution, the converse does not hold in general [11, 30]. In other words, even when the max-min objective is solved exactly, the recovered map T_{rec} may fail to push the source measure μ to the target ν, leading to incorrect transport maps.

Fan et al. [11] showed that, when the finite-dimensional source measure μ is atomless, the saddle point solution (V_{sad}, T_{sad}) of $\mathcal{L}_{V,T}$ satisfying

$$V_{sad} \in \arg\max_V \mathcal{L}(V, T_{sad}), \qquad T_{sad} \in \arg\min_T \mathcal{L}(V_{sad}, T). \tag{21}$$

is equivalent to the Kantorovich optimal potential and optimal transport map. Note that the saddle point condition is a stronger assumption than the max-min solution, and this recovery is not guaranteed for the generic max-min solution of $\mathcal{L}_{V,T}$.

Choi et al. [7] established a sufficient condition for the max-min solution to recover optimal transport map in **finite-dimensional Euclidean spaces**. The sufficient condition is that the source measure μ must not assign positive mass to measurable sets of Hausdorff dimension at most $d-1$, where d denotes the ambient dimension. To enforce this condition in practice, they proposed Gaussian smoothing of the source measure with an annealing scheme, and proved weak convergence of the annealed optimal transport plan $\pi^\star_\varepsilon$. Chu et al. [8] strengthened this weak convergence analysis by deriving quantitative convergence rates. Based on this, they proposed a convergence-rate-optimal noise annealing schedule. They also analyzed necessary smoothing directions in finite dimensions; our Hilbert-space analysis distinguishes the necessary kernel condition from the additional hypotheses needed for a converse.

We study inner-minimizer identifiability for regular Hilbert-space sources (Theorem 3.2), W_2 accumulation-point consistency of exact optimal plans (Theorem 4.2), and necessary and sufficient smoothing conditions with their precise scope (Theorem 4.1). Finite-dimensional directional and rate analyses are complementary [8].

Zhu et al.Zhu et al. [42] estimate Hilbert–Schmidt linear maps using the squared 2-Wasserstein distance between the pushforward source and target distributions and a squared Hilbert–Schmidt norm penalty. Their finite-basis empirical formulation jointly optimizes the map coefficients and a coupling between source and target samples [42, Eq. (12)]. Our focus is the identifiability of the quadratic semi-dual inner minimizer and regularization of singular Hilbert-valued sources. These are population statements; the numerical models use discretized functional observations.

Kernel Neural Optimal Transport [17] studies weak costs $C(x, \pi(\cdot \mid x))$ and represents conditional laws with a stochastic map $T(x, z)$. A weak quadratic objective can fail to identify a target-preserving optimal conditional plan; a continuous characteristic kernel, under the paper's hypotheses, restores conditional-law identifiability. Its Hilbert feature space distinguishes distributions and need not be the data space. In contrast, we retain the standard pairwise cost $\frac{1}{2}\|x-y\|_H^2$ on a Hilbert data space and analyze the deterministic pointwise selector $T(x)$. We regularize the source rather than replacing the transport cost. The weak-OT formulation is stated on general spaces, with Euclidean constructions and image experiments; it should not be described as excluding all non-Euclidean spaces.

Dynamic OT [2] and Wasserstein gradient flows [1, 24] concern evolving measures. The present theory solves a static endpoint problem and does not analyze that evolution; see Appendix H.5.

G Algorithm and Implementation Details

This appendix collects the implementation details omitted from the main text. We begin with a rigorous construction of the Hilbert-valued Gaussian perturbation used for spectral augmentation. We then give examples of orthonormal bases for functional data, state the full HiSNOT training algorithm, and describe the experimental setups for the synthetic transport examples and time-series imputation benchmarks.

G.1 Hilbert-Valued Gaussian Perturbations

We provide the formal statement and proof for the Hilbert-valued Gaussian perturbation used in the spectral augmentation step.

Proposition G.1 (Hilbert-valued Gaussian perturbation)**.** *Let $\{e_k\}_{k\in\mathbb{N}}$ be an orthonormal basis of H, and let $\{\lambda_k\}_{k\in\mathbb{N}} \subset [0,\infty)$ satisfy*

$$\sum_{k=1}^{\infty} \lambda_k < \infty.$$

Let $\{\xi_k\}_{k\in\mathbb{N}}$ be i.i.d. standard Gaussian random variables. Then

$$G := \sum_{k=1}^{\infty} \sqrt{\lambda_k}\, \xi_k e_k \tag{22}$$

defines an H-valued centered Gaussian random variable. Its covariance is the trace-class operator defined by $Qe_k = \lambda_k e_k$. In particular, $G \sim \mathcal{N}(0, Q)$.

Proof. Let $G_N := \sum_{k=1}^{N} \sqrt{\lambda_k}\xi_k e_k$ be the partial sum. We first show that $\{G_N\}_{N\in\mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega; H)$. For $M > N$, we have:

$$\begin{aligned}
\mathbb{E}\left[\|G_M - G_N\|_H^2\right] &= \mathbb{E}\left[\left\|\sum_{k=N+1}^{M} \sqrt{\lambda_k}\xi_k e_k\right\|_H^2\right] \\
&= \sum_{k=N+1}^{M} \lambda_k \mathbb{E}[\xi_k^2]\|e_k\|_H^2 \\
&= \sum_{k=N+1}^{M} \lambda_k.
\end{aligned}$$

Since $\{\lambda_k\}_{k\in\mathbb{N}}$ is summable (trace-class condition), the tail sum converges to 0 as $N, M \to \infty$. Thus, G_N converges to a limit G in $L^2(\Omega; H)$. Hence, there exists $G \in L^2(\Omega; H)$ such that $G_N \to G$ in $L^2(\Omega; H)$.

To verify that G is Gaussian, consider an arbitrary linear functional $\ell(h) = \langle h, u\rangle_H$ for $u \in H$. Then

$$\langle G, u\rangle_H = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k \langle e_k, u\rangle_H.$$

For each N, $\langle G_N, u\rangle_H$ is a centered Gaussian random variable, and it converges in $L^2(\Omega)$ to $\langle G, u\rangle_H$. Since the class of Gaussian laws is closed under convergence in distribution, $\langle G, u\rangle_H$ is Gaussian. Since the variance

$$\sum_{k=1}^{\infty} \lambda_k \langle e_k, u\rangle^2 \le \left(\max_{k\in\mathbb{N}} \lambda_k\right) \|u\|_H^2 < \infty,$$

the limit is a univariate Gaussian random variable. Thus, G is a Gaussian random vector.

Finally, we identify the covariance operator Q. For any $u, v \in H$,

$$\begin{aligned}
\mathbb{E}[\langle G, u\rangle\langle G, v\rangle] &= \mathbb{E}\left[\left(\sum_{j=1}^{\infty} \sqrt{\lambda_j}\xi_j\langle e_j, u\rangle\right)\left(\sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k\langle e_k, v\rangle\right)\right] \\
&= \sum_{j,k=1}^{\infty} \sqrt{\lambda_j\lambda_k}\delta_{jk}\langle e_j, u\rangle\langle e_k, v\rangle \\
&= \sum_{k=1}^{\infty} \lambda_k \langle e_k, u\rangle\langle e_k, v\rangle \\
&= \langle Qu, v\rangle,
\end{aligned}$$

where Q is the operator defined by $Qe_k = \lambda_k e_k$. This completes the proof. □

Finite-mode coefficients and source approximation Fix $\varepsilon > 0$, $\mu \in \mathcal{P}_2(H)$, and the trace-class covariance Q and orthonormal basis of Proposition G.1. Let $x \sim \mu$ be independent of the Gaussian coefficients, set $x_\varepsilon = x + \varepsilon G$, and write $\mu_\varepsilon = \mathrm{Law}(x_\varepsilon)$. With $G_M = \sum_{k=1}^{M} \sqrt{\lambda_k}\xi_k e_k$, the noise truncation used in Algorithm 1 keeps the source signal and forms $x + \varepsilon G_M$. Coupling this sample with $x + \varepsilon G$ gives

$$\begin{aligned} W_2^2(\mathrm{Law}(x + \varepsilon G_M), \mu_\varepsilon) &\le \mathbb{E}\|\varepsilon(G - G_M)\|_H^2 \\ &= \varepsilon^2 \sum_{k>M} \lambda_k \longrightarrow 0. \end{aligned}$$

The equality follows from orthonormality and the unit variances of the Gaussian coefficients; the tail vanishes because Q is trace class. Thus increasing M approximates the smoothed source law for the same fixed Q and ε. Under the sufficient conditions of Theorem 4.1, including injective Q, this limiting law is regular on H.

For a representation that also projects the source onto $H_M = \mathrm{span}\{e_1, \ldots, e_M\}$, the retained coefficient vector and reconstructed function are

$$\begin{aligned} \mathbf{c}_{\varepsilon,M} &:= \left(\langle x, e_k\rangle_H + \varepsilon\sqrt{\lambda_k}\xi_k\right)_{k=1}^{M} \in \mathbb{R}^M, \\ x_{\varepsilon,M} &:= \sum_{k=1}^{M} (\mathbf{c}_{\varepsilon,M})_k e_k = \sum_{k=1}^{M} \langle x_\varepsilon, e_k\rangle_H e_k \in H_M. \end{aligned}$$

These constructions agree when $x \in H_M$ almost surely. If $\lambda_k > 0$ for all $k \le M$, the added coefficient vector is an independent Gaussian with positive-definite covariance $\varepsilon^2 \,\mathrm{diag}(\lambda_1, \ldots, \lambda_M)$. Convolution therefore gives a joint density $f_{\varepsilon,M}$ for $\mathbf{c}_{\varepsilon,M}$ with respect to Lebesgue measure on $\mathbb{R}^M$:

$$\begin{aligned} \mathbb{P}(\mathbf{c}_{\varepsilon,M} \in A) &= \int_A f_{\varepsilon,M}(z)\,\mathrm{d}z, \qquad A \subseteq \mathbb{R}^M \text{ Borel}, \\ \mathrm{Law}(x_{\varepsilon,M}) &= \Big(z \mapsto \sum_{k=1}^{M} z_k e_k\Big)_{\#}\big(f_{\varepsilon,M}(z)\,\mathrm{d}z\big) \in \mathcal{P}_2^r(H_M). \end{aligned}$$

In infinite-dimensional H, this law gives unit mass to the Gaussian-null subspace H_M and hence is non-regular as a law on H.

The reconstruction still approximates the full-space smoothed source in Wasserstein distance. Using the same x and Gaussian coefficients couples $x_{\varepsilon,M}$ and x_ε. Parseval's identity and independence, which makes the source–noise cross terms vanish, yield

$$\begin{aligned} W_2^2(\mathrm{Law}(x_{\varepsilon,M}), \mu_\varepsilon) &\le \mathbb{E}\|x_\varepsilon - x_{\varepsilon,M}\|_H^2 \\ &= \sum_{k>M} \mathbb{E}|\langle x, e_k\rangle_H|^2 + \varepsilon^2 \sum_{k>M} \lambda_k \longrightarrow 0. \end{aligned}$$

The source tail vanishes since $\mathbb{E}\|x\|_H^2 < \infty$, and the noise tail vanishes by the trace-class assumption. This estimate holds also for zero eigenvalues; positivity of the retained eigenvalues is used for the density statement above. These estimates quantify source-law truncation using exact spectral coefficients. Reconstruction of coefficients from discrete observations and convergence of learned transport maps require additional numerical and optimization analysis.

G.2 Examples of Orthonormal Bases

As described in the **Practical Smoothing via Spectral Augmentation** section, our method projects data onto an orthonormal basis $\{e_k\}$ to inject noise. In our experiments, we specifically utilized the Fourier basis. We detail this basis below, along with other illustrative examples for $H = L^2(-1, 1)$ that are compatible with our framework.

1. Fourier Basis. On the interval $(-1, 1)$, the orthonormal basis is defined as:

$$e_k(t) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k = 0, \\ \cos(m\pi t) & \text{if } k = 2m,\ m \ge 1, \\ \sin(m\pi t) & \text{if } k = 2m-1,\ m \ge 1. \end{cases}$$

2. Legendre Polynomials. Let $P_n(t)$ be the Legendre polynomials satisfying the recurrence relation $(n+1)P_{n+1}(t) = (2n+1)tP_n(t) - nP_{n-1}(t)$ with $P_0(t) = 1, P_1(t) = t$. To ensure orthonormality on $L^2(-1,1)$, we apply the normalization factor $\sqrt{\frac{2k+1}{2}}$:

$$e_k(t) = \sqrt{\frac{2k+1}{2}} P_k(t), \quad k \in \mathbb{N}_0.$$

This basis is particularly useful for capturing global trends without enforcing periodicity at the boundaries.

3. Haar Wavelets. Let $h_0(t) = \frac{1}{\sqrt{2}} \mathbf{1}_{(-1,1)}(t)$. For scale $j \in \mathbb{N}_0$ and shift $k = 0, \ldots, 2^j - 1$, define:

$$h_{j,k}(t) := 2^{j/2} \frac{1}{\sqrt{2}} \left(\mathbf{1}_{\left[-1+\frac{2k}{2^j}, -1+\frac{2(k+0.5)}{2^j}\right)} - \mathbf{1}_{\left[-1+\frac{2(k+0.5)}{2^j}, -1+\frac{2(k+1)}{2^j}\right)} \right).$$

The collection $\{h_0\} \cup \{h_{j,k}\}_{j,k}$ forms a complete orthonormal system for $L^2(-1,1)$.

G.3 Algorithm

Algorithm 1 Training algorithm of HiSNOT with Gaussian Annealing

Require: Source measure μ, target measure ν; Orthonormal basis $\{e_k\}$ and summable sequence $\{\lambda_k\}$; Transport operator T_θ and potential V_ϕ; batch size m; outer iterations N, retained modes M, inner-loop steps K_T; annealing schedule $\{\varepsilon_k\}_{k=0}^{N-1}$.

for $k = 0, 1, \ldots, N-1$ **do**
  Sample batches $\{x_i\}_{i=1}^m \sim \mu$ and $\{y_i\}_{i=1}^m \sim \nu$.
  For each sample i, generate $G_i = \sum_{\ell=1}^M \sqrt{\lambda_\ell}\xi_{i\ell}e_\ell$, with independent $\xi_{i\ell} \sim \mathcal{N}(0,1)$.
  Apply Spectral Smoothing
  $\tilde{x}_i \leftarrow x_i + \varepsilon_k G_i$ for each i.
  Update ϕ to maximize:
    $\mathcal{L}_\phi = \frac{1}{m}\sum_{i=1}^m \left[-V_\phi\left(T_\theta(\tilde{x}_i)\right) + V_\phi(y_i)\right]$.
  for $j = 1, \ldots, K_T$ **do**
    Sample a batch $\{x_i\}_{i=1}^m \sim \mu$ and generate fresh independent noises $\{G_i\}_{i=1}^m$ as above.
    $\tilde{x}_i \leftarrow x_i + \varepsilon_k G_i$ for each i.
    Update θ to minimize:
      $\mathcal{L}_\theta = \frac{1}{m}\sum_{i=1}^m \left[\frac{1}{2}\|\tilde{x}_i - T_\theta(\tilde{x}_i)\|_H^2 - V_\phi\left(T_\theta(\tilde{x}_i)\right)\right]$.
  end for
end for

G.4 Synthetic Data Experiments

In this section, we describe the experimental setup and the implementation details for the synthetic data experiments

Dataset Description Throughout this paragraph, let $e_1(t) := \sin(\pi t)$ and $e_2(t) := \sin(2\pi t)$ be the basis functions for $t \in [0,1]$. We represent each signal as a linear combination of these bases, where $\mathbf{c} = (c_1, c_2) \in \mathbb{R}^2$ and $\mathbf{d} = (d_1, d_2) \in \mathbb{R}^2$ denote the coefficient vectors for the source signal $x(t) = \sum_{i=1}^2 c_i e_i(t)$ and the target signal $y(t) = \sum_{i=1}^2 d_i e_i(t)$, respectively. Here c_i, d_i are expansion coefficients in the displayed sine functions; these functions are not normalized in $L^2(0,1)$, so their coefficients need not equal the inner products used for an orthonormal basis. We consider the following four scenarios:

- **Perpendicular**: We generate the source coefficients $\mathbf{c} \sim \mu$ as follows: $c_1 \sim U([-1,1])$ and $c_2 \equiv 0$. Similarly, we sample the target coefficients $\mathbf{d} \sim \nu$ by $d_1 \equiv 0$ and $d_2 \sim U([-1,1])$.
- **Parallel**: We generate $\mathbf{c} \sim \mu$ as follows: $c_1 \sim U([-1,1])$, and $c_2 \equiv 0$. Similarly, we sample $\mathbf{d} \sim \nu$ by $d_1 \sim U([-1,1])$ and $d_2 \equiv 0.5$.
- **One-to-Many**: We generate $\mathbf{c} \sim \mu$ as follows: $c_1 \sim U([-1,1])$, and $c_2 \equiv 0$. Similarly, we sample $\mathbf{d} \sim \nu$ by $d_1 \sim U([-1,1])$ and $d_2 \sim \text{Cat}(\{-0.5, 0.5\}, (0.5, 0.5))$.
- **Grid**: Let $\mathbb{P} := \text{Cat}(\{-0.75, -0.25, 0.25, 0.75\}, (0.25, 0.25, 0.25, 0.25))$. We generate $\mathbf{c} \sim \mu$ by sampling $c_1 \sim U([-1,1])$, and $c_2 \sim \mathbb{P}$. Similarly, we sample $\mathbf{d} \sim \nu$ by $d_1 \sim \mathbb{P}$ and $d_2 \sim U([-1,1])$.

Training Details Across all experiments, we use the same network architecture and training hyperparameters unless otherwise stated. Both the potential function V_ϕ and the transport map T_θ are parameterized by neural networks composed of two Fourier layers [20]. In particular, the potential network V_ϕ additionally includes a final linear integral functional layer [28]. We train the models with a batch size of 256 for 5000 epochs. We set the OT cost parameter to $\tau = 0.2$ for the Multi-Perpendicular case, and $\tau = 0.001$ for all other cases. To improve the optimization of the inner-loop in the dual formulation of equation 11, we perform $K_T = 5$ updates of T_θ for every single update of V_ϕ. Regarding the regularization, we set $\lambda = 0$ for the Orthogonal case and $\lambda = 0.01$ otherwise. Here λ is a regularization weight, distinct from the covariance eigenvalues λ_k.

For our experiments, we generate perturbed inputs $\widetilde{x}(t)$ as follows: $\widetilde{x}(t) = x(t) + \sigma G(t)$, where $x(t)$ is a clean source signal and G is a finite Fourier-basis noise. The experimental schedules below use σ

for the smoothing amplitude denoted by ε in the main text. Using e_k for the Fourier basis functions, the noise expansion is

$$G(t) = \sum_{k=1}^{M} \sqrt{\lambda_k}\, \xi_k\, e_k(t), \qquad \xi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad \lambda_k = \frac{1}{k^2}, \tag{23}$$

where we use $M = 16$ Fourier modes in all experiments.

We employ a linear noise schedule from the initial noise $\sigma_{\max} = 0.5$ to the terminal noise $\sigma_{\min} = 0.06$. The noise level is updated over the first 80% of training epochs and kept fixed afterwards. At epoch e, the noise level is

$$\sigma = \max(\sigma_{\min},\ (1 - t_e)\sigma_{\max} + t_e \sigma_{\min})\,, \qquad t_e = \frac{e}{0.8E},\ \ \sigma_{\max} = 0.5,\ \ \sigma_{\min} = 0.06, \tag{24}$$

where E is the total number of epochs.

G.5 Time Series Imputation

In this section, we describe the experimental setup and implementation details for the time series imputation task. We conduct experiments on ETTm1, ETTm2, ETTh1, ETTh2, and Exchange datasets [37]. For each dataset, we consider missing ratios of 0.5 and 0.7, and generate missing patterns following the same setup as Wang et al. [35], and split the data into training, validation, and test sets with a ratio of 7:1:2.

Both the potential function V_ϕ and the transport map T_θ are parameterized using the Neural Fourier Model (NFM) [16]. Unless otherwise specified, we follow Kim et al. [16]. We additionally adopt channel-independent modeling as in NFM and use a batch size of 1792. We train the model for 40 epochs with learning rate 2×10^{-4}, dropout 0.1, regularization weight $\lambda = 0.01$, and $\tau = 50$.

We follow the same input perturbation scheme as in the synthetic time-series experiments G.4. We employ a linear noise schedule from the initial noise $\sigma_{\max} = 0.2$ to the terminal noise $\sigma_{\min} = 0.01$. The stated clipped linear schedule is

$$\sigma_i = \max\left(\sigma_{\min},\ \sigma_{\max}\left(1 - \frac{i}{0.8I}\right)\right), \qquad \sigma_{\max} = 0.2\ \ , \sigma_{\min} = 0.01. \tag{25}$$

where i is the iteration index and I denotes the maximum schedule iterations. For the displayed parameters, this formula reaches $\sigma_{\min}$ at $i = 0.76I$ and remains there. This describes the displayed formula; the precise implemented schedule requires the training record.

H Interpretation, Computation, and Scope

H.1 What Hilbert geometry supplies

The Hilbert assumption supports the arguments used here, beyond providing an ambient space for functional observations. For $c(x, y) = \frac{1}{2}\|x - y\|_H^2$, the Riesz identification $H^* \simeq H$ yields $D_x c(x, y) = x - y$. At differentiability points of a source-side truncated transform, this identity determines its contacting target point; Appendix B gives the argument without assuming that an untruncated potential is locally Lipschitz on its whole domain. A Hilbert-valued Gaussian also has a positive self-adjoint trace-class covariance, an orthonormal eigenbasis, and canonical projections onto $\ker Q$ and its orthogonal complement. However, its Cameron–Martin space $\mathrm{Ran}(Q^{1/2})$ need not be closed in H. This distinction is essential to the qualified smoothing result in Theorem 4.1.

Gaussian laws exist on Banach spaces too, but their covariance naturally maps the dual space into the original space; the same orthogonal decomposition and gradient identification are not generally available. Our results apply to structured objects represented in a separable Hilbert space, including paths, fields, signals on a fixed graph, and Hilbert–Schmidt operators. They do not establish an intrinsic theory for arbitrary manifolds, varying graphs, or general Banach, Fréchet, or metric spaces.

H.2 Interpreting source regularity

Regularity is a population property. In finite dimensions it is equivalent to absolute continuity with respect to Lebesgue measure. A nondegenerate Gaussian on H is regular by definition, as is its

Table 3: Comparison of Time-Series Imputation Performance for Missing Ratios 0.5 and 0.7. Results are reported as mean±standard deviation, computed over 5 random seeds. All baseline results (except Ours) are taken from Wang et al. [35].

Datasets		**ETTh1**		**ETTh2**		**ETTm1**		**ETTm2**		**Exchange**	
Methods	**Ratio**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
Transformer	0.5	0.249 ± 0.007	0.359 ± 0.005	0.250 ± 0.007	0.360 ± 0.008	0.069 ± 0.009	0.173 ± 0.005	0.032 ± 0.008	0.122 ± 0.008	0.261 ± 0.009	0.215 ± 0.003
	0.7	0.426 ± 0.005	0.461 ± 0.002	0.313 ± 0.004	0.405 ± 0.0	0.100 ± 0.007	0.219 ± 0.008	0.080 ± 0.002	0.199 ± 0.007	0.378 ± 0.006	0.362 ± 0.0
DLinear	0.5	0.148 ± 0.008	0.272 ± 0.003	0.117 ± 0.004	0.240 ± 0.007	0.110 ± 0.01	0.229 ± 0.001	0.089 ± 0.005	0.206 ± 0.002	0.252 ± 0.01	0.208 ± 0.003
	0.7	0.225 ± 0.009	0.341 ± 0.001	0.159 ± 0.002	0.283 ± 0.008	0.152 ± 0.007	0.270 ± 0.007	0.125 ± 0.007	0.248 ± 0.003	0.323 ± 0.007	0.267 ± 0.006
TimesNet	0.5	0.163 ± 0.008	0.297 ± 0.005	0.126 ± 0.005	0.260 ± 0.0	0.061 ± 0.002	0.179 ± 0.0	0.072 ± 0.001	0.197 ± 0.01	0.343 ± 0.008	0.283 ± 0.007
	0.7	0.523 ± 0.007	0.532 ± 0.01	0.168 ± 0.002	0.295 ± 0.005	0.110 ± 0.005	0.238 ± 0.002	0.114 ± 0.002	0.243 ± 0.005	0.401 ± 0.009	0.332 ± 0.004
FreTS	0.5	0.229 ± 0.008	0.358 ± 0.003	0.141 ± 0.004	0.261 ± 0.001	0.048 ± 0.001	0.149 ± 0.0	0.037 ± 0.007	0.134 ± 0.005	0.256 ± 0.002	0.211 ± 0.002
	0.7	0.236 ± 0.003	0.351 ± 0.006	0.176 ± 0.006	0.294 ± 0.009	0.074 ± 0.008	0.187 ± 0.007	0.042 ± 0.008	0.137 ± 0.002	0.201 ± 0.002	0.166 ± 0.009
PatchTST	0.5	0.153 ± 0.003	0.284 ± 0.008	0.137 ± 0.003	0.269 ± 0.003	**0.046** ± 0.006	0.139 ± 0.004	0.029 ± 0.006	0.116 ± 0.001	0.204 ± 0.009	0.169 ± 0.007
	0.7	0.278 ± 0.001	0.392 ± 0.004	0.150 ± 0.005	0.275 ± 0.008	**0.064** ± 0.009	0.166 ± 0.001	0.036 ± 0.005	0.127 ± 0.002	0.250 ± 0.005	0.206 ± 0.009
SCINet	0.5	0.169 ± 0.001	0.300 ± 0.005	0.176 ± 0.004	0.285 ± 0.008	0.062 ± 0.001	0.173 ± 0.004	0.077 ± 0.001	0.206 ± 0.0	0.280 ± 0.001	0.231 ± 0.003
	0.7	0.220 ± 0.001	0.341 ± 0.006	0.168 ± 0.001	0.290 ± 0.007	0.113 ± 0.005	0.234 ± 0.009	0.093 ± 0.007	0.216 ± 0.004	0.371 ± 0.007	0.307 ± 0.006
iTransformer	0.5	0.168 ± 0.007	0.289 ± 0.009	0.095 ± 0.004	0.210 ± 0.004	0.054 ± 0.003	0.154 ± 0.002	0.029 ± 0.009	0.117 ± 0.01	0.058 ± 0.002	0.045 ± 0.007
	0.7	0.244 ± 0.004	0.353 ± 0.007	0.155 ± 0.007	0.263 ± 0.003	0.082 ± 0.003	0.191 ± 0.008	0.057 ± 0.001	0.167 ± 0.004	0.116 ± 0.008	0.095 ± 0.005
SAITS	0.5	0.229 ± 0.009	0.321 ± 0.008	0.184 ± 0.004	0.264 ± 0.01	0.064 ± 0.001	0.169 ± 0.008	0.057 ± 0.001	0.156 ± 0.008	1.007 ± 0.004	0.831 ± 0.004
	0.7	0.343 ± 0.005	0.395 ± 0.004	0.287 ± 0.001	0.341 ± 0.004	0.094 ± 0.002	0.207 ± 0.002	0.061 ± 0.009	0.164 ± 0.009	1.006 ± 0.01	0.833 ± 0.0
CSDI	0.5	0.161 ± 0.005	0.293 ± 0.001	0.096 ± 0.001	0.269 ± 0.0	0.059 ± 0.009	0.139 ± 0.009	0.186 ± 0.002	0.126 ± 0.008	0.125 ± 0.0	0.139 ± 0.008
	0.7	0.213 ± 0.002	0.312 ± 0.01	0.117 ± 0.004	0.273 ± 0.002	0.252 ± 0.004	0.273 ± 0.006	0.136 ± 0.002	0.174 ± 0.005	0.143 ± 0.005	0.161 ± 0.01
Sinkhorn	0.5	0.910 ± 0.008	0.669 ± 0.007	0.820 ± 0.002	0.614 ± 0.01	0.950 ± 0.005	0.708 ± 0.003	0.905 ± 0.001	0.678 ± 0.008	0.740 ± 0.005	0.611 ± 0.001
	0.7	0.962 ± 0.009	0.701 ± 0.004	0.934 ± 0.008	0.673 ± 0.001	0.979 ± 0.005	0.725 ± 0.01	0.948 ± 0.002	0.725 ± 0.003	0.826 ± 0.002	0.684 ± 0.005
TDM	0.5	0.993 ± 0.004	0.749 ± 0.008	0.978 ± 0.007	0.738 ± 0.009	1.002 ± 0.003	0.754 ± 0.005	0.995 ± 0.0	0.746 ± 0.004	0.957 ± 0.005	0.790 ± 0.006
	0.7	0.989 ± 0.008	0.749 ± 0.008	1.017 ± 0.007	0.751 ± 0.007	1.004 ± 0.003	0.757 ± 0.006	1.001 ± 0.008	0.749 ± 0.0	0.981 ± 0.001	0.812 ± 0.009
PWS-I	0.5	**0.125** ± 0.003	**0.234** ± 0.009	0.047 ± 0.006	0.145 ± 0.003	0.048 ± 0.0	**0.133** ± 0.008	0.021 ± 0.008	0.096 ± 0.007	0.032 ± 0.005	**0.026** ± 0.005
	0.7	**0.194** ± 0.004	**0.289** ± 0.006	0.059 ± 0.001	0.165 ± 0.003	0.066 ± 0.003	**0.158** ± 0.009	0.029 ± 0.006	0.110 ± 0.002	0.039 ± 0.009	**0.033** ± 0.004
Ours	0.5	0.166 ± 0.004	0.260 ± 0.006	**0.045** ± 0.001	**0.141** ± 0.002	0.057 ± 0.002	0.141 ± 0.004	**0.019** ± 0.0	**0.086** ± 0.001	**0.005** ± 0.0	0.048 ± 0.003
	0.7	0.288 ± 0.009	0.333 ± 0.006	**0.054** ± 0.001	**0.158** ± 0.001	0.076 ± 0.001	0.163 ± 0.003	**0.025** ± 0.0	**0.098** ± 0.001	**0.006** ± 0.001	0.049 ± 0.004

convolution with any probability measure. By contrast, for a finite-mode model $X = \sum_{j=1}^{r} A_j \phi_j$ in an infinite-dimensional H, the source lies in a proper finite-dimensional subspace even if $(A_1, \ldots, A_r)$ has a density. Every proper closed affine subspace is Gaussian null: it is contained in a level set of a nonzero continuous linear functional, whose distribution under any nondegenerate Gaussian has no atoms. Thus this source is non-regular in the ambient H, although it may be regular in its intrinsic finite-dimensional space.

Similarly, if $A : H \to Y$ is a nonzero bounded linear operator into a normed space and $Au = f$ has a solution u_0, the solution set $u_0 + \ker A$ is a proper closed affine subspace. Any law concentrated there is non-regular. This covers bounded linear observation constraints and appropriately posed linear physical constraints; an unbounded PDE operator requires a suitable domain and topology before this reasoning applies. Exact low-rank models provide a rigorous example of singularity. A general nonlinear manifold hypothesis alone is not a proof that a particular observed population is non-regular.

A finite empirical measure is atomic and hence non-regular on a nontrivial H, regardless of the unknown population law. Finite observations do not certify population regularity. Smoothing therefore offers a constructive regularization strategy rather than a statistical test for that property.

H.3 Choosing a covariance

One distribution-independent choice is an injective trace-class covariance

$$Q_0 = \sum_{j \geq 1} \tau_j (e_j \otimes e_j), \qquad \tau_j > 0, \qquad \sum_{j \geq 1} \tau_j < \infty.$$

By Theorem 4.1, it regularizes every source. To allocate more variance to target-relevant directions, finite empirical principal subspaces $\widehat{U}_\mu, \widehat{U}_\nu$ can guide an optional refinement:

$$\widehat{W} = (\widehat{U}_\mu + \widehat{U}_\nu) \cap \widehat{U}_\mu^{\perp}, \qquad Q = \sigma_{\text{adapt}}^2 \pi_{\widehat{W}} + \sigma_{\text{floor}}^2 Q_0, \quad \sigma_{\text{floor}} > 0.$$

Here $\pi_{\widehat{W}}$ is the orthogonal projection onto $\widehat{W}$. The finite-rank first term emphasizes selected directions; the injective floor supplies the population guarantee. FPCA neither certifies regularity nor consistently recovers the exact singular geometry under the present analysis. The source determines validity, while the target guides allocation. This is a practical design suggestion, not an additional evaluated algorithm. A finite-mode isotropic floor is nondegenerate only in the retained space H_M; it is not an injective covariance on the entire infinite-dimensional H.

H.4 Computational overhead and perturbation size

Let m denote batch size, M the number of retained spectral modes, and K_T the number of transport updates per potential update, as in Algorithm 1. Sampling and adding one perturbation takes $O(mM)$ coefficient operations and $O(mM)$ temporary storage. With freshly sampled perturbations for one potential update and K_T transport updates, the additional cost per outer iteration is $O((K_T+1)mM)$. This count assumes independent spectral coefficients in the chosen basis. Explicit synthesis on an observation grid adds the cost of the selected basis transform. No additional network, trainable parameter, or optimization loop is introduced, and the number of neural forward/backward passes is unchanged. These are operation counts; no separate wall-clock overhead measurement is reported.

The coupling $(X + \varepsilon G, X)$ gives

$$W_2^2(\mu * \mathcal{N}(0, \varepsilon^2 Q), \mu) \leq \varepsilon^2 \operatorname{Tr}(Q), \qquad \mu \in \mathcal{P}_2(H).$$

Larger noise can aid regularization while increasing finite-ε source perturbation. This bound concerns the source distribution, not learned-map error or downstream prediction accuracy. The ablations in Appendix J assess only the tested smoothing supports and endpoint values. We do not derive a finite-sample optimal covariance, magnitude, or annealing schedule.

H.5 Static transport, dynamic transport, and generation

Dynamic OT describes paths of measures and velocity fields subject to a continuity equation [2]. Wasserstein gradient flows describe energy-driven evolution, often approximated with successive

JKO steps [1, 24]. Neural methods can approximate those steps [24], but the present analysis concerns one static problem between fixed endpoint laws. It proves neither continuity-equation estimates nor energy dissipation or convergence of a JKO scheme. Using HiSNOT as a component of such a method would require additional analysis along the evolving marginals.

For functional generation, a nondegenerate Hilbert-valued Gaussian reference γ is regular, and a learned operator aims to satisfy $T_{\#}\gamma \approx \nu$; samples are obtained as $T(X)$ for $X \sim \gamma$. Identifiability applies to the ideal problem with an admissible optimal potential. Appendix I reports the additional Exchange experiment; it does not establish arbitrary-resolution generation or finite-sample optimization guarantees.

I Additional Function-Level Generation Experiment

We evaluate function-level generation on the Exchange dataset using a Gaussian source distribution and Exchange time-series windows as the target distribution. Training samples are constructed by moving a fixed-length window over the target series with the specified stride. The window size determines the temporal length of a sample, while the stride determines the displacement between successive windows. At a fixed window size, a larger stride reduces the number of training windows and their overlap.

For evaluation, we generate $1{,}000$ windows from each trained model and randomly sample $1{,}000$ windows from the target dataset. We report the 2-Wasserstein distance W_2 between the generated and target samples; lower values indicate a closer match under this evaluation. Table 4 compares HiSNOT with the vector SNOT baseline for six window-size/stride settings.

Table 4: **Function-level generation on Exchange.** Evaluation uses $1{,}000$ generated and $1{,}000$ target windows. The $\pm$ terms are reported dispersion values; their definition and replication count are unavailable in the provided experiment record. Bold marks the lower reported central W_2 value within each row.

Window size	Stride	HiSNOT ($W_2 \downarrow$)	SNOT ($W_2 \downarrow$)
720	1	9.632 ± 0.201	$\mathbf{9.341} \pm 0.341$
360	1	$\mathbf{5.009} \pm 0.161$	5.133 ± 0.120
240	1	$\mathbf{3.543} \pm 0.031$	3.817 ± 0.010
240	2	$\mathbf{3.687} \pm 0.232$	4.426 ± 0.080
240	3	$\mathbf{3.681} \pm 0.143$	4.783 ± 0.120
120	4	$\mathbf{2.055} \pm 0.126$	2.383 ± 0.085

HiSNOT has a lower reported central W_2 value in five of the six settings; SNOT has the lower value for window size 720 and stride 1. At fixed window size 240, the difference in the reported central values favors HiSNOT at all three tested strides. These results provide an additional finite-observation functional generation example. They do not establish statistical significance, an advantage in every setting, or resolution-independent convergence of the numerical method.

J Additional Smoothing Ablations

We report three controlled ablations of smoothing support and schedule endpoints. The target-distribution error D_{target} is the metric defined in Section 5.1; lower values are better.

J.1 Smoothing Support in a Finite-Dimensional Subspace

Consider

$$H_5 = \mathrm{span}\{e_1, \ldots, e_5\}, \qquad e_k(t) = \sin(k\pi t).$$

The source has the form $X = A_3 e_3 + A_4 e_4 + A_5 e_5$, where (A_3, A_4, A_5) has a density on $\mathbb{R}^3$, and the target is supported on $\mathrm{span}\{e_1, e_2\}$. For the diagonal covariances tested here, smoothing must include both e_1 and e_2 to make this source absolutely continuous within H_5. This is equivalent to absolute continuity of the source projection onto the unsmoothed kernel in that kernel's finite-dimensional space. All regularity labels in Table 5 refer to H_5, not to a larger infinite-dimensional ambient space.

Table 5: **Smoothing-support ablation in H_5.** The source occupies modes e_3, e_4, e_5 with a joint coefficient density. The condition column records whether its unsmoothed projection is absolutely continuous in the finite-dimensional covariance kernel. Replication counts and dispersion estimates are unavailable in the provided experiment record.

Active smoothing modes	Kernel condition in H_5	$D_{\mathrm{target}} \downarrow$
e_1	Violated	0.2063
e_1, e_2	Satisfied	0.0276
e_1, e_2, e_3	Satisfied	0.0300
e_1, e_2, e_3, e_4	Satisfied	0.0249

Omitting e_2 gives an error of 0.2063, whereas the three tested supports containing both required directions give errors between 0.0249 and 0.0300. Thus, in this controlled finite-dimensional example, the exact minimal support does not have to be recovered to obtain comparable errors. Additional directions can nevertheless increase $\mathrm{Tr}(Q)$ and the source perturbation at positive smoothing amplitude; these results do not imply that arbitrary extra smoothing is harmless.

J.2 Initial Smoothing Level

In the One-to-Many experiment, we vary the initial smoothing level $\varepsilon_{\max}$ while keeping the remaining annealing protocol fixed. The setting $\varepsilon_{\max} = 0$ denotes the unsmoothed baseline. Table 6 reports all seven tested values.

Table 6: **Initial smoothing-level ablation on One-to-Many.** The remaining annealing protocol is fixed. Replication counts and dispersion estimates are unavailable in the provided experiment record.

$\varepsilon_{\max}$	$D_{\mathrm{target}} \downarrow$
0.0	0.288
0.2	0.004
0.5	0.004
0.6	0.003
0.8	0.004
1.0	0.004
2.0	0.009

All tested positive initial levels yield $D_{\mathrm{target}} < 0.01$, compared with 0.288 without smoothing. This supports robustness to the initial amplitude over the tested range in this experiment.

J.3 Terminal Smoothing Level

We next fix $\varepsilon_{\max} = 0.5$ and vary the terminal level $\varepsilon_{\min}$ at which the annealing schedule ends. Table 7 reports all six tested values.

Table 7: **Terminal smoothing-level ablation.** The initial level is fixed at $\varepsilon_{\max} = 0.5$. Replication counts and dispersion estimates are unavailable in the provided experiment record.

$\varepsilon_{\min}$	$D_{\mathrm{target}} \downarrow$
0.02	0.008
0.06	0.003
0.10	0.004
0.14	0.004
0.18	0.004
0.20	0.004

All tested terminal levels yield target-distribution error below 0.01. The endpoint ablations support robustness within the tested ranges; they neither identify an optimal finite-sample schedule nor eliminate tuning on other datasets. The population bound in Appendix H.4 controls source perturbation, not learned-map or downstream prediction error.

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 precisely state the theoretical contributions that correspond to Theorems 3.2, 4.1, and 4.2 and experimental claims align with the results in Sections 5.

   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: The conclusion explicitly notes that consistency is established only at the plan level (not map level) and that learning the smoothing covariance operator Q from finite samples remains open.

   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: All theorems (3.2, 4.1, 4.2) and supporting propositions are stated with explicit assumptions by complete proofs in Appendix B, C, and D.

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: Appendix G provides full implementation details including dataset, network architectures, batch sizes, learning rates, noise schedules, and hyperparameters.

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: Anonymized code is provided as supplementary material.

   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: Training details, optimizer settings, annealing schedules ($\sigma_{\max}$, $\sigma_{\min}$), data splits (7:1:2), and missing ratios (0.5, 0.7) are fully specified in Appendix G.

   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: Yes. As reported in Table 3, all results are reported as mean$\pm$standard deviation over 5 random seeds.

   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: All experiments were conducted on a single NVIDIA A100 80GB GPU, as noted in Appendix G.

   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: The research conforms with the NeurIPS Code of Ethics. It involves no human subjects, uses only publicly available time-series benchmarks (ETT, Exchange) [37].

   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 work is foundational theoretical and methodological research on optimal transport plan in Hilbert spaces with no direct path to specific harmful applications.
    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: The paper does not release pretrained models, generative systems, or scraped datasets that would pose misuse risks.

    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 (ETT,Exchange [37]) and baseline results from Wang et al. [35] are properly cited. The neural network architectures used in our experiments are also properly cited in Appendix G. These include Fourier Neural Operator [20], the integral functional layer from [28], and Neural Fourier Model [16].

    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: An anonymized implementation of HiSNOT is provided as supplementary material, including a README with dependencies, training commands.

    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 human subjects research.

    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 human subjects research and therefore does not require IRB approval.

    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 not used as a component of the core methodology, theoretical analysis, or experimental design.

    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.