PROXIMAL TERM MATTERS: RETHINKING THE DIFFUSION MODELS FOR TIME-SERIES DATA IMPUTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models (DMs) have shown promise for Time-Series Data Imputation (TSDI). However, their performance remains inconsistent in complex scenarios. We attribute this to two primary obstacles: 1). non-stationary temporal dynamics, which can bias the inference trajectory and lead to outlier-sensitive imputations; and 2). objective inconsistency, since imputation favors accurate pointwise recovery, whereas DMs are inherently trained to generate diverse samples. To better understand these issues, we analyze the DM-based TSDI process from a proximal-operator perspective and uncover two inherent components: Wasserstein-distance-based proximal regularizer, which hinders the model's ability to counteract non-stationarity, and a dissipative regularizer, which amplifies diversity at the expense of fidelity. Building on this insight, we propose a novel framework called SPIRIT (Semi-Proximal Transport Regularized time-series Imputation). Specifically, we introduce an entropy-induced Bregman divergence to relax the mass-preserving constraint in the Wasserstein distance, formulate the semi-proximal transport (SPT) discrepancy, and theoretically establish the robustness of SPT to non-stationary pattern. Subsequently, we remove the dissipative structure and derive the complete SPIRIT workflow, with SPT serving as the proximal operator. Extensive experiments demonstrate the efficacy of the proposed SPIRIT approach.

1 INTRODUCTION

Data completeness in time series is paramount across many domains (Wang et al., 2025c). For example, in healthcare (Prosperi et al., 2020), clinicians rely on wearable and ambient sensors to continuously monitor patients' physiological signals, yet the collected records are often incomplete due to sensor disconnections. Similarly, in industrial manufacturing (Liu et al., 2024a; Wang et al., 2025b), sensor networks are deployed to support process monitoring; however, harsh operating conditions and mechanical vibrations can cause sensor malfunctions and missing readings. Such incompleteness undermines data integrity, which is essential for accurate analysis (Qiu et al., 2024; Liu et al., 2025), thereby underscoring the need for effective time-series data imputation (TSDI).

Recently, diffusion models (DMs) have been widely adopted for the TSDI task due to their excellent performance on data-generation tasks (Yang et al., 2023). Given partially observed time-series data, DMs first define a forward diffusion process, also known as the noising process, which gradually perturbs the target data with Gaussian noise and produces a sequence of increasingly noisy variables. Imputation is then performed by reversing this process conditional on the observed entries and progressively recovering the missing values from an initial noise sample (Johnson & Zhang, 2021; Dong et al., 2022). Since the reverse dynamics depend on the time-dependent conditional score of the noisy marginal distributions, a neural network is trained to estimate this score. Accordingly, TSDI can be formulated as a conditional distribution-sampling problem and solved by integrating a score-guided stochastic differential equation (SDE). Following Tashiro et al. (2021), subsequent DM-based TSDI methods have mainly focused on refining model architectures (Liu et al., 2023a; 2024a), redesigning the forward noising process (Chen et al., 2023), and improving training objectives (Yu et al., 2025).

Despite the success of DMs, we argue that directly applying DMs to TSDI can lead to suboptimal performance due to two latent limitations. First, *non-stationarity*: time-series data often exhibit non-stationary fluctuations (Wang et al., 2025c; Liu et al., 2022b), which are not explicitly considered in existing SDE-based diffusion formulations to the best of our knowledge. Second, the *objective*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

inconsistency: DMs are designed for generative task and thus favor *diverse* samples, whereas imputation prioritizes *accurate* recovery of the missing values (Liu et al., 2024c; Chen et al., 2024).

To address these issues, we first *analyze* DM-based TSDI by formulating it as a recursion induced by a proximal operator, hereafter referred to as *proximal recursion* (Caluya & Halder, 2019; 2020). This formulation reveals two key terms underlying the aforementioned problems: the Wasserstein distance and a dissipative regularizer. Building on this analysis, we mitigate these issues by relaxing the Wasserstein distance with a generalized Bregman divergence induced by the entropy functional and removing the dissipative regularizer. We then re-derive an *improved* framework, termed Semi-Proximal Transport Regularized time-series Imputation (SPIRIT), for the TSDI task.

Contributions: The main contributions of this manuscript can be articulated as follows:

1. We cast DM-based TSDI as an optimization problem with a proximal term and pinpoint two key bottlenecks: 1). limited robustness to non-stationarity induced by the Wasserstein distance, and 2). imputation inaccuracy caused by the dissipative regularizer.
2. We relax the Wasserstein distance via the generalized Bregman divergence, propose a semi-proximal transport discrepancy, and provide theoretical analysis on robustness to non-stationarity.
3. By removing the dissipative regularizer, we re-derive the TSDI procedure within the proximal recursion framework using the proposed semi-proximal transport discrepancy, yielding our proposed method: Semi-Proximal Transport Regularized Time-Series Imputation (SPIRIT).

2 Preliminaries

2.1 Notations

This manuscript focuses on TSDI *per se*, namely, estimating the most probable values of missing entries. We do not consider imputation as a means of generating inputs for downstream tasks (Jarrett et al., 2022), such as label prediction (Ma et al., 2022; Zhao et al., 2023) or pseudo-label-based unbiased learning (Li et al., 2024), which may require joint training under task-specific objectives (Wang et al., 2025b). We primarily consider the missing completely at random (MCAR) setting to facilitate theoretical analysis. Due to space limitations, additional preliminaries are provided in Section B.

2.2 Problem Formulation

Let $\mathbf{X}^{\text{ideal}} \in \mathbb{R}^{N\times T\times D}$ denote a fully observed time-series dataset of N samples, each containing T time steps and D features. A binary mask $\mathbf{M} \in \{0,1\}^{N\times T\times D}$ indicates missingness, where $\mathbf{M}_{n,t,d} = 1$ denotes a missing entry and 0 otherwise. The observed data are $\mathbf{X}^{\text{obs}} = \mathbf{X}^{\text{ideal}} \odot (1 - \mathbf{M}) + \texttt{NaN} \odot \mathbf{M}$, where $\odot$ denotes the Hadamard product. Time-series imputation aims to construct $\mathbf{X}^{\text{imp}} \in \mathbb{R}^{N\times T\times D}$ such that $\mathbf{X}^{\text{imp}} \odot \mathbf{M} + \mathbf{X}^{\text{ideal}} \odot (1 - \mathbf{M}) \approx \mathbf{X}^{\text{ideal}}$, while preserving the observed entries. We denote the data distribution by $p(\mathbf{X}^{\text{ideal}})$ and the empirical distribution of imputed samples by $q(\mathbf{X}^{\text{imp}}) = \frac{1}{N}\sum_{n=1}^{N} \delta_{\mathbf{x}_n^{\text{imp}}}$, where $\delta_{\mathbf{x}_n^{\text{imp}}}$ is the Dirac measure concentrated at the n-th imputed sample. To distinguish dataset time from diffusion time, we use t and T for the time index and sequence length of the dataset, and τ and $\mathscr{T}$ for the diffusion time and its terminal value.

2.3 Proximal Operator and Proximal Recursion

Given a proper, lower semi-continuous convex function $g : \mathbb{R}^D \to \mathbb{R}\cup\{+\infty\}$, the proximal operator $\text{prox}_{\varepsilon g}(\cdot)$ with step size ε is given by:

$$\text{prox}_{\varepsilon g}(\mathbf{x}) = \arg\min_{\mathbf{y}} g(\mathbf{y}) + \frac{1}{2\varepsilon}\|\mathbf{y} - \mathbf{x}\|_2^2. \tag{1}$$

Crucially, Equation (1) balances minimizing g against a "proximal regularization term", $\frac{1}{2\varepsilon}\|\mathbf{y} - \mathbf{x}\|_2^2$, which forces the solution $\mathbf{y}$ to remain in the neighborhood of $\mathbf{x}$ with a strength controlled by the coefficient $\frac{1}{2\varepsilon}$ (Parikh & Boyd, 2014). The process of iteratively applying Equation (1) is called proximal recursion (Caluya & Halder, 2020; 2021).

2.4 Wasserstein Distance

Let $\mathcal{P}_2(\mathbb{R}^D)$ be the set of probability measures on $\mathbb{R}^D$ with finite second moment. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^D)$, the 2-Wasserstein distance (Villani, 2009) is defined by the "optimal transport" (OT) problem:

$$\mathbb{W}_2^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int \|\mathbf{x} - \mathbf{y}\|^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}), \tag{2}$$

where $\Pi(\mu, \nu)$ denotes the joint distributions on $\mathbb{R}^D \times \mathbb{R}^D$ with marginals μ and ν.

2.5 Dissipative Structure in Stochastic Dynamics

In continuous-time stochastic dynamics (Särkkä & Solin, 2019), the evolution of a state $\mathbf{x}_\tau$ is modeled by the following stochastic differential equation:

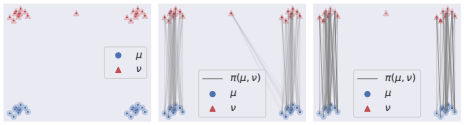
$$\mathrm{d}\mathbf{x}_\tau = b(\mathbf{x}_\tau, \tau)\mathrm{d}\tau + \sigma(\mathbf{x}_\tau, \tau)\mathrm{d}W_\tau, \tag{3}$$

where $b(\mathbf{x}_\tau, \tau)$ is the drift term, $\sigma(\mathbf{x}_\tau, \tau)$ is the diffusion coefficient, and $\mathrm{d}W_\tau$ denotes the Wiener process. From the perspective of non-equilibrium stochastic dynamics, both the deterministic drift $b(\mathbf{x}_\tau, \tau) = -\mathbf{x}_\tau$ and the stochastic diffusion $\mathrm{d}W_\tau$ contribute to the dissipative behavior of the system (Otto, 2001): $-\mathbf{x}_\tau$ pulls the state back toward the origin, causing energy dissipation, while the Wiener-driven diffusion $\mathrm{d}W_\tau$ spreads the state distribution to prevent collapse.

3 Methodology

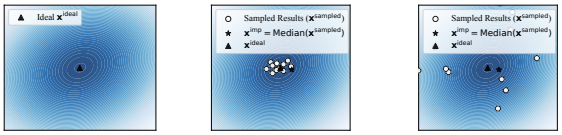
3.1 Motivation Analysis

DMs implicitly introduce Wasserstein distance regularization and a dissipative structure during inference (see our further derivations). However, in the Time-Series Data Imputation (TSDI) task, these properties encounter two major challenges arising from the characteristics of the data and the evaluation protocol. First, the data are highly *non-stationary*: for example, in the Electricity dataset, consumption patterns differ substantially across weekdays and holidays, while in the Weather dataset, climatological patterns vary markedly across seasons. Second, the task demands stringent pointwise *accuracy*, as exemplified by mean squared error (MSE) and mean absolute error (MAE). To illustrate how these challenges interact with the diffusion framework, we present the following studies.



(a) Data. (b) OT plan. (c) SPT plan.

Figure 1: Transport plans for OT and SPT.



(a) Ideal $\mathbf{x}^{\text{ideal}}$. (b) With $\mathrm{d}W_\tau$. (c) With $-\mathbf{x}^{\text{imp}}$.

Figure 2: Imputation under different dissipative terms.

Wasserstein Distance Meets Non-Stationarity. We illustrate the non-stationarity in Figure 1(a), where the data exhibit distinct multimodal structures accompanied by transient outliers. However, directly applying the canonical Wasserstein distance leads to severe misalignment, as demonstrated in Figure 1(b). Due to the rigorous mass-conservation constraint, the transport plan generates erroneous couplings: the right mode of the source μ is incorrectly matched with the distant left mode of the target ν, while significant mass is also forcibly diverted to the outliers in the upper-middle region.

Dissipative Structure Meets Optimization. In this part, we investigate how the dissipative structure affects imputation accuracy. Since the inference process of diffusion models is the time reversal of the predefined SDE in Equation (3), we first consider the optimal imputed value and its target distribution, as illustrated in Figure 2(a). We then examine the effects of the dissipative terms, namely the linear drift $-\mathbf{x}^{\text{imp}}$ and the stochastic diffusion $\mathrm{d}W_\tau$, in Figures 2(b) and 2(c). From these figures, we observe that the imputed value $\mathbf{x}^{\text{imp}}$ (the star, computed as the median of the inferred samples (Tashiro et al., 2021), which are shown as white dots) can deviate significantly from the ideal value (the triangle). This indicates that, although the dissipative structure promotes diversity in the inferred samples, it can also lead to dispersed imputations and thus reduced accuracy.

3.2 Analysis of DMs in the Context of TSDI

In the context of DM-based TSDI, suppose that we have learned the conditional score function $\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})$. The imputation procedure is then carried out by simulating the following SDE from $\tau = \mathscr{T}$ to $\tau = 0$ (Tashiro et al., 2021; Chen et al., 2023):

$$\mathrm{d}\mathbf{x}^{\text{imp}} = [f(\mathbf{x}^{\text{imp}}, \tau) - g^2 \nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})]\mathrm{d}\tau + g\mathrm{d}W_\tau, \tag{4}$$

where $f(\mathbf{x}^{\text{imp}}, \tau)$ and g are the predefined drift and diffusion coefficients, respectively. In particular, we update only entries for which $\mathbf{M}_{n,t,d} = 1$.

Directly solving Equation (4) may be suboptimal for TSDI, as illustrated by the following proposition:

Proposition 3.1. *The imputation process for DMs can be formulated as an iterative solution of the upper bound of the following optimization problem in a proximal operator form:*

$$\inf_{q'} \; -\mathbb{E}_{q'}[\log p(\mathbf{x}^{imp}|\mathbf{x}^{obs})] + \frac{1}{2\eta}\mathbb{W}_2^2(q', q) + \phi(\mathbf{x}^{imp}), \tag{5}$$

where we abbreviate the candidate distribution (distribution for current iteration) $q'(\mathbf{x}^{imp})$ and the base distribution (distribution for previous iteration) $q(\mathbf{x}^{imp})$ as q' and q, respectively. The term $\frac{1}{\eta}$ is a predefined positive term determined by the noise schedule of DMs, and the term $\phi(\mathbf{x}^{imp})$ denotes the dissipative structure-related term, which depends on the underlying SDE and is specified as follows:

- *Variance Preserving SDE :* $\phi(\mathbf{x}^{imp}) = \frac{1}{2}\mathbb{E}_{q'}[\log q'(\mathbf{x}^{imp})] - \frac{1}{4}\mathbb{E}_{q'}[\|\mathbf{x}^{imp}\|_2^2]$, *and* $\eta = \beta(\tau)$.
- *Variance Exploding SDE:* $\phi(\mathbf{x}^{imp}) = \frac{1}{2}\mathbb{E}_{q'}[\log q'(\mathbf{x}^{imp})]$, *and* $\eta = \frac{1}{2}\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}$.

From the proposition above, it is evident that directly deploying DMs for TSDI poses two critical challenges. First, the proximal term induced by the squared 2-Wasserstein distance $\mathbb{W}_2^2(q', q)$ restricts the model's flexibility, resulting in a lack of robustness to non-stationary dynamics. Second, the proximal term induced by the dissipative regularizer $\phi(\mathbf{x}^{\text{imp}})$ injects entropy that encourages diverse outputs, potentially degrading the deterministic accuracy required for TSDI.

On this basis, the key to addressing the first issue is to devise a novel discrepancy metric that mitigates the impact of non-stationarity. For the second issue, we aim to remove the dissipative structure and re-derive an alternative inference procedure that ensures accurate imputation.

3.3 Semi-Proximal Transport Framework

To address the sensitivity of the vanilla OT framework to outliers due to strict mass conservation in non-stationary environments, we propose the SPT discrepancy. Unlike standard OT, which forces a complete match between distributions, SPT relaxes the hard constraint on the target marginal by introducing a divergence penalty.

Formally, we define the SPT discrepancy as follows:

$$\mathbb{S}(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu)} \int \|\mathbf{x} - \mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + D_\psi(\pi_{\mathbf{y}}, \nu), \tag{6}$$

where $\Pi(\mu)$ denotes the set of couplings with the first marginal fixed to μ, and $\pi_{\mathbf{y}}$ denotes the second marginal of π. The term $D_\psi(\cdot, \cdot)$ is the generalized Bregman divergence defined as follows (Blondel et al., 2022):

$$D_\psi(\rho, \nu) \coloneqq \psi(\rho) - \psi(\nu) - \langle \delta_\nu \psi(\nu), \rho - \nu \rangle, \tag{7}$$

where $\delta_\nu \psi(\nu)$ is the first variation of $\psi(\nu)$ with respect to ν, and $\psi : \mathcal{P}_2(\mathbb{R}^D) \to \mathbb{R} \cup \{\infty\}$ is a strictly convex functional termed the Bregman potential (Blair, 1985; Bonet et al., 2024). While the framework admits general potentials, we specifically adopt the entropic potential $\psi(\mu) \coloneqq \int \mu(\mathbf{x})[\log \mu(\mathbf{x}) - 1]\mathrm{d}\mathbf{x}$ in this work. We select this specific potential for its geometric properties. Specifically, its gradient map acts as a mirror map ensuring that iterations naturally remain within the space of positive measures (i.e., $\rho > 0$) throughout the optimization (Hsieh et al., 2018; Shi et al., 2022; Sharrock et al., 2023), thereby guaranteeing a well-defined transport plan.

By replacing the strict constraint $\pi_{\mathbf{y}} = \nu$ with a soft penalty $D_\psi(\pi_{\mathbf{y}}, \nu)$, the SPT framework transforms vanilla OT into a selective matching process. When the cost of transporting mass to an

outlier is excessively high, the optimization prefers to incur the penalty rather than distort the transport plan, thereby ensuring robustness in non-stationary scenarios. Based on Lemma 1 in reference (Fatras et al., 2021) and Theorem C.2 in reference (Wang et al., 2025c), we state the following lemma to demonstrate the outlier robustness of our proposed SPT discrepancy:

Lemma 3.2. *Let μ and ν be probability measures on $\mathbb{R}^D$, and consider a contaminated target distribution $\tilde{\nu} = (1-\zeta)\nu + \zeta\delta_{\mathbf{z}}, \zeta \in (0,1)$, where $\delta_{\mathbf{z}}$ denotes a Dirac mass at the outlier location $\mathbf{z} \in \mathbb{R}^D$. The Wasserstein distance has the following lower bound:*

$$\mathbb{W}_2^2(\mu, \tilde{\nu}) \geq \zeta\mathbb{W}_2^2(\mu, \nu) + (1-\zeta)[\|\mathbf{y}^* - \mathbf{z}\|_2^2 - g(\mathbf{y}^*) + \int h(\mathbf{x})\mu(\mathbf{x})\mathrm{d}\mathbf{x}], \tag{8}$$

for some $\mathbf{y}^*$ belonging to the support of ν, where f and g are optimal dual potentials for $\mathbb{W}_2^2(\mu, \nu)$. Meanwhile, the SPT discrepancy with Bregman potential $\psi(\nu) \coloneqq \int \nu(\mathbf{y})[\log \nu(\mathbf{y}) - 1]\mathrm{d}\mathbf{y}$ under target contamination admits the following bound:

$$\mathbb{S}(\mu, \tilde{\nu}) \leq (1-\zeta)\mathbb{S}(\mu, \nu) + \zeta(1 - e^{-D(\mathbf{z})}) + C(\zeta), \tag{9}$$

where $D(\mathbf{z}) \coloneqq \int \|\mathbf{z} - \mathbf{x}\|_2^2 \mu(\mathbf{x})\mathrm{d}\mathbf{x}$ is the average distance between $\mathbf{z}$ and samples from ν, and $C(\zeta)$ is a constant defined as $C(\zeta) \coloneqq (1-\zeta)\log\frac{1}{1-\zeta} - \zeta\log\zeta$.

This theoretical advantage is visually illustrated in the case study shown in Figure 1. As demonstrated in Figure 1(c), while standard OT incorrectly pairs distinct modes due to forced matching, the SPT framework successfully ignores transient noise and correctly aligns the stable modes, effectively handling non-stationarity in time-series data.

3.4 Diversity-Eliminated Functional for Imputation

Based on the theoretical results in Sections 3.2 and 3.3, we design the following objective functional to *improve* the DM-based TSDI task:

$$\inf_{q'} \; -\mathbb{E}_{q'}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})] + \frac{1}{2\eta}\mathbb{S}(q', q) + \frac{1}{2}\mathbb{E}_{q'}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2], \tag{10}$$

where we eliminate the dissipative structure-related term $\phi(\mathbf{x}^{\text{imp}})$, which encourages diversity in the imputation results. Based on this, we add gradient-norm regularization of the corresponding $\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})$, namely $\mathbb{E}_{q'}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2]$, which seeks to find a saddle point. As such, we formulate the following proposition to demonstrate a sufficient condition for reducing the functional given by Equation (10):

Proposition 3.3. *Assume $p(\mathbf{x}^{imp}|\mathbf{x}^{obs})$ is once continuously differentiable and $\nabla \log p(\mathbf{x}^{imp}|\mathbf{x}^{obs})$ is square-integrable under the measures considered. Let $\psi(\rho) = \int \rho(\mathbf{x})[\log \rho(\mathbf{x}) - 1]\mathrm{d}\mathbf{x}$. Represent q' by an empirical measure $q' = \sum_{i=1}^N w_i \delta_{\mathbf{x}_i}$ with $w_i \geq 0$ and $\sum_{i=1}^N w_i = 1$. Under the alternating splitting scheme detailed in Section C.3, the following coordinate-wise directions provide first-order updates for the resulting location and weight surrogate subproblems, respectively:*

- ***Location direction*** *(score-direction-guided), where $\{\mathbf{x}_i\}_{i=1}^N$ are updated by:*

$$\boldsymbol{T}_{\mathbf{x}^{imp}}(\mathbf{x}^{imp}) = \nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}}). \tag{11}$$

- ***Weight direction*** *(score-norm-guided), where logarithmic weights $\{\log w_i\}_{i=1}^N$ are updated by:*

$$\boldsymbol{T}_w(\mathbf{x}^{imp}) = -2\|\nabla \log p(\mathbf{x}^{imp}|\mathbf{x}^{obs})\|_2^2 + 2\mathbb{E}[\|\nabla \log p(\mathbf{x}^{imp}|\mathbf{x}^{obs})\|_2^2]. \tag{12}$$

Although Proposition 3.3 provides an update direction for the log-weights $\{\log w_i\}_{i=1}^N$, directly applying this direction does not guarantee that the resulting weights satisfy the simplex constraints, i.e., $w_i \geq 0$ and $\sum_{i=1}^N w_i = 1$. To address this feasibility issue, we introduce the following proposition to ensure a well-defined iteration process for $\{w_i\}_{i=1}^N$ using the proximal recursion framework:

Proposition 3.4. *Let $\eta > 0$ and $\boldsymbol{T}_w \in \mathbb{R}^N \to \mathbb{R}$. Define the intermediate log-weights as $\log \widehat{w}_i^{(k+1)} \coloneqq \log w_i^{(k)} + \eta \boldsymbol{T}_w(\mathbf{x}^{imp})$. The normalized weights are given by $w_i^{(k+1)} = \frac{\widehat{w}_i^{(k+1)}}{\sum_{j=1}^N \widehat{w}_j^{(k+1)}}$.*

Because the centering term in $\boldsymbol{T}_w$ is shared by all particles and cancels after normalization, the update is equivalently proportional to $w_i^{(k)} \exp\{-2\eta\|\nabla \log p(\mathbf{x}_i^{\text{imp}} \mid \mathbf{x}^{\text{obs}})\|_2^2\}$. Thus, the score norm acts as a continuous reweighting signal: particles farther from stationarity are down-weighted instead of being rolled back. In addition, the constant factor 2 is absorbed into the effective step size of the weight block. So far, within the proximal recursion framework, we have addressed robustness to non-stationarity and the inaccuracy induced by the two proximal terms, namely the Wasserstein distance and the dissipative regularizer, in DM-based TSDI.

3.5 Overall Workflow

Although Section 3.4 proposes a novel imputation process under the proximal recursion framework, the overall SPIRIT workflow is not yet explicitly summarized. To complete the workflow, we still need a principled way to learn the conditional score $\nabla \log p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}})$.

However, unlike previous DM-based TSDI approaches, in which the score function is obtained by bridging missing and observed data using a predefined SDE and can be learned via score matching, our method modifies the objective functional and thus cannot directly adopt the same learning strategy. Therefore, we introduce the following proposition to enable learning of $\nabla \log p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}})$:

Proposition 3.5. *Let $\mathscr{X} = \mathbf{x}^{imp}$ and $\mathscr{O} = \mathbf{x}^{obs}$, and define $\mathscr{S}(\mathscr{X}, \mathscr{O}) := \nabla_{\mathscr{X}} \log p(\mathscr{X}|\mathscr{O})$, $m(\mathscr{X}) := \nabla_{\mathscr{X}} \log p(\mathscr{X})$. Assume that differentiation and marginalization can be interchanged and that the scores are square-integrable. Since $m(\mathscr{X}) = \mathbb{E}[\mathscr{S}(\mathscr{X}, \mathscr{O}) \mid \mathscr{X}]$, we have:*

$$\mathbb{E}_{\mathscr{X},\mathscr{O}}[\|s_\theta(\mathscr{X}) - \mathscr{S}(\mathscr{X}, \mathscr{O})\|_2^2] = \mathbb{E}_{\mathscr{X}}[\|s_\theta(\mathscr{X}) - m(\mathscr{X})\|_2^2] + \mathbb{E}_{\mathscr{X},\mathscr{O}}[\|\mathscr{S}(\mathscr{X}, \mathscr{O}) - m(\mathscr{X})\|_2^2]. \tag{13}$$

Hence, the two risks have the same minimizers over predictors depending only on $\mathscr{X}$.

Proposition 3.5 shows that, among predictors depending only on $\mathscr{X}$, the conditional-score and marginal-score risks have the same population minimizers. We therefore use the marginal score as a practical approximation to the directions in Proposition 3.3. Given an initial imputation $\mathbf{x}^{\text{imp}}$, we train the score network by denoising score matching (Vincent, 2011):

$$\arg\min_{s_\theta} \mathbb{E}_{q_\sigma(\widehat{\mathbf{x}}^{\text{imp}}|\mathbf{x}^{\text{imp}})}[\|s_\theta(\widehat{\mathbf{x}}^{\text{imp}}) - \nabla \log q_\sigma(\widehat{\mathbf{x}}^{\text{imp}}|\mathbf{x}^{\text{imp}})\|_2^2], \tag{14}$$

where σ is the variance scale, $\widehat{\mathbf{x}}^{\text{imp}}$ is obtained as $\widehat{\mathbf{x}}^{\text{imp}} = \mathbf{x}^{\text{imp}} + \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, and $\nabla \log q_\sigma(\widehat{\mathbf{x}}^{\text{imp}}|\mathbf{x}^{\text{imp}}) = -\frac{\widehat{\mathbf{x}}^{\text{imp}} - \mathbf{x}^{\text{imp}}}{\sigma^2}$.

Finally, the overall workflow of SPIRIT is summarized in Figure 3. It consists of two stages: "Score Learning" and "Recursive Imputation". In the "Score Learning" stage, we train the score network s_θ using the DSM objective in Equation (14). In the "Recursive Imputation" stage, we implement approximate updates inspired by Proposition 3.3, replacing its conditional-score directions with the learned score, and normalize the weights as in Proposition 3.4. Alternating between these two stages yields the final imputation $\mathbf{X}^{\text{imp}}$. In addition, Figure 3 introduces an operator, `ApplyGrad`, which takes (i) a gradient direction, (ii) a learning rate, and (iii) the current variable to be updated, and applies a gradient-based update. This operator can be implemented using standard optimizers in deep learning backends (Paszke et al., 2019). Due to space limit, we provide the complete pseudocode in Algorithm 1 of Section C.6.

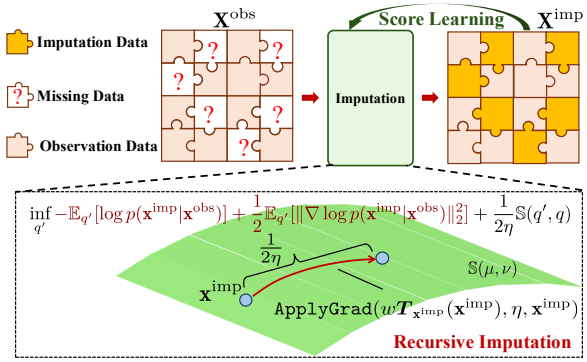


Figure 3: Overview of the SPIRIT's two stages.

Notably, SPIRIT consists of two alternating stages, "Score Learning" and "Recursive Imputation", whose convergence properties we analyze separately. For "Score Learning", convergence to a stationary point follows from standard optimization results under mild regularity conditions (Bottou et al., 2018). For "Recursive Imputation", define the energy $\mathcal{J}[q'] = -\mathbb{E}_{q'}[\log p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}})] + \frac{1}{2}\mathbb{E}_{q'}[\|\nabla \log p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}})\|_2^2]$. Under suitable smoothness and step-size conditions, the location and weight updates provide sufficient decrease for their respective frozen-coordinate surrogate blocks. In the idealized continuous-time limit of the split transport–reaction dynamics, the energy is non-increasing when both directions are generated by the same energy; see Section C.6. The finite-step updates illustrated in Figure 3 are interpreted as a discretization of this limiting flow as $\eta \to 0$.

4 Main Experimental Results

4.1 Setup

Datasets. We evaluate our method using standard public benchmarks for time-series imputation following Wu et al. (2023), namely the ETT dataset (four subsets), Exchange, Illness, and Traffic.

Baseline Models. We compare SPIRIT with following baselines: Crossformer (Zhang & Yan, 2023), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), Autoformer (Wu et al., 2021), ETSformer (Woo et al., 2022), FiLM (Zhou et al., 2022), DLinear (Zeng et al., 2023), HELIX (Zhang et al., 2026), GP-VAE (Fortuin et al., 2020), CSDI (Tashiro et al., 2021), MIRI (Yu et al., 2026), Glocal (Yang et al., 2025b), Sinkhorn (Muzellec et al., 2020), TDM (Zhao et al., 2023), and PSW-I (Wang et al., 2025c). We group them into three categories: discriminative methods (Crossformer, TimesNet, PatchTST, Autoformer, ETSformer, FiLM, DLinear, Glocal, and HELIX), which learn predictive models for imputation; generative methods (GP-VAE, CSDI, and MIRI); and alignment-based methods (Sinkhorn, TDM, and PSW-I), which impute by aligning distributions.

Implementation. Based on our preliminary notes outlined in Section 2, we simulate the missingness mechanism under the protocol described in Section D.4. The missing ratio p_{Miss} is simulated within $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. The patch length in our experiment is set as 24. Further details are provided in Sections D.2 to D.4.

4.2 Overall Performance

Point-Estimation Metrics (MAE and MSE). Table 1 reports the overall performance of our method and the related baselines, which makes the following observations:

- **Efficacy of existing TSDI approaches:** Existing TSDI methods exhibit diverse performance across datasets. Among the discriminative approaches, PatchTST achieves the best or second-best result in 3 out of 14 cases, whereas TimesNet does not rank among the top two in the updated results. Methods designed specifically for imputation also remain competitive: CSDI attains the lowest MAE on ETT-m1, while PSW-I is ranked second in 8 cases. These results highlight the strengths of existing approaches, but also show that their effectiveness varies substantially across datasets.
- **Mass relaxation matters:** The balanced OT-based baselines, Sinkhorn, TDM, CSDI, and MIRI, are consistently outperformed by the relaxed transportation-based methods PSW-I and SPIRIT across all reported metrics. In particular, SPIRIT achieves the best result in 10 out of 14 cases. This contrast is consistent with the view that mass relaxation helps accommodate non-stationary.
- **Role of dissipative-structure elimination:** CSDI achieves the best MAE on ETT-m1, but its errors increase on Illness and Traffic, indicating that its performance can be dataset-dependent. MIRI and GP-VAE are also outperformed by SPIRIT on every reported metric. These observations are consistent with the benefit of eliminating the dissipative-structure term for imputation accuracy.
- **Efficacy of the proposed SPIRIT method:** SPIRIT achieves the best result in 10 out of 14 cases and the second-best result in another 2, ranking among the top two on 12 of the 14 reported metrics. It performs particularly well on ETT-h2, ETT-m2, Exchange, and Illness, and obtains the lowest MAE on Traffic. Its main exceptions are the two metrics on ETT-h1, where HELIX and PatchTST lead. Overall, the results demonstrate strong imputation accuracy across the evaluated datasets.

Uncertainty-Oriented Metric (CRPS). We further evaluate uncertainty quantification using the CRPS, comparing SPIRIT with the generative baselines GP-VAE, CSDI, and MIRI. For SPIRIT and MIRI, we add white noise to the initial imputation to obtain diverse outputs across multiple runs and

Table 1: Imputation performance comparison in terms of MAE and MSE.

Dataset	ETT-h1		ETT-h2		ETT-m1		ETT-m2		Exchange		Illness		Traffic	
Metric	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Crossformer	0.265	0.191	0.483	0.591	0.703*	0.966*	0.740*	1.086*	0.318	0.291	0.281*	0.291	0.303	0.320
TimesNet	0.246*	0.116	0.214*	0.087*	0.147*	0.045*	0.163*	0.070*	0.240*	0.127*	0.289*	0.246*	0.301	0.288
PatchTST	0.204	**0.093**	0.171	0.062	0.126*	0.038	0.107*	0.025*	0.170*	0.058*	0.300*	0.298	0.403	0.520
Autoformer	0.628*	0.915*	0.684*	1.239*	0.650*	1.121*	0.671*	1.351*	0.712*	1.193*	0.682*	1.022*	0.471	0.574
ETSformer	0.279*	0.179*	0.202*	0.087*	0.201*	0.097*	0.134*	0.038*	0.178*	0.057*	0.321*	0.302*	0.291	**0.218**
FiLM	0.434*	0.436*	0.271*	0.143*	0.248*	0.127*	0.215*	0.095*	0.204*	0.077*	0.491*	0.608*	0.851	1.341
DLinear	0.251*	0.135*	0.216*	0.097*	0.211*	0.096*	0.191*	0.076*	0.199*	0.073*	0.254*	0.174*	0.364	0.367
Glocal	0.240*	0.121*	0.194*	0.076*	0.140*	0.044*	0.113*	0.026*	0.163*	0.051*	0.348*	0.314*	0.381	0.452
HELIX	**0.203**	0.100	0.139	0.047*	0.126*	0.043*	0.092*	0.023*	0.046*	0.005*	0.192*	0.127*	0.452	0.509
GP-VAE	0.316*	0.192*	0.271*	0.150	0.246*	0.121*	0.258*	0.158*	0.337*	0.211*	0.557*	0.752*	0.726	0.959
CSDI	0.248	0.319	0.259	1.437	**0.117**	0.041	0.085	0.076	0.087*	0.030*	8.568*	400.6*	23.58	1488
MIRI	0.736*	1.082*	0.597*	0.863*	0.674*	0.980*	0.529*	0.723*	0.443*	0.417*	0.358*	0.370*	0.303	0.297
Sinkhorn	0.752*	0.994*	0.752*	1.003*	0.754*	1.000*	0.749*	0.998*	0.826*	0.996*	0.715*	1.012*	0.780	1.041
TDM	0.750*	0.992*	0.750*	1.000*	0.754*	1.000*	0.748*	0.998*	0.822*	0.991*	0.705*	0.997*	0.773	1.033
PSW-I	0.219*	0.112*	0.137*	0.042	0.122*	0.041*	0.083	0.019*	0.024	0.002*	0.111	0.054*	0.292	0.329
SPIRIT (Ours)	0.209	0.108	**0.131**	**0.038**	0.118	**0.037**	**0.082**	**0.016**	**0.023**	**0.002**	**0.107**	**0.045**	**0.235**	0.228
Win Counts	13	13	15	15	14	15	15	15	15	15	15	15	15	14

Kindly Note: Each entry represents the average results at $p_{\text{miss}} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. Best results are in **bold**; second best are in wavy underline. "*" marks the results that SPIRIT significantly outperform with p-value< 0.05 over paired samples t-test.

construct an imputation distribution. As shown in Table 2, SPIRIT achieves the lowest CRPS on all four ETT datasets at every evaluated missing ratio, yielding the best result in all 24 comparisons. MIRI is the strongest baseline in 17 of the 24 settings, but SPIRIT consistently outperforms it. As the missing ratio increases from 0.1 to 0.6, SPIRIT's CRPS rises more slowly than MIRI's, widening their gap on all four datasets. Meanwhile, CSDI deteriorates sharply on ETT-h1 and ETT-h2, whereas GP-VAE shows more variable performance across datasets and missing ratios. These results suggest that SPIRIT maintains strong probabilistic imputation performance even under severe missingness.

Table 2: CRPS results on the ETT datasets.

Missing Ratio	0.1				0.2				0.3				Average			
Model	GP-VAE	CSDI	MIRI	SPIRIT	GP-VAE	CSDI	MIRI	SPIRIT	GP-VAE	CSDI	MIRI	SPIRIT	GP-VAE	CSDI	MIRI	SPIRIT
ETT-h1	0.368	0.327	0.253	**0.239**	0.391	0.375	0.287	**0.248**	0.412	0.443	0.313	**0.262**	0.390	0.382	0.284	**0.250**
ETT-h2	0.319	0.448	0.197	**0.153**	0.314	0.334	0.213	**0.159**	0.319	0.308	0.243	**0.163**	0.317	0.363	0.218	**0.158**
ETT-m1	0.275	0.184	0.201	**0.134**	0.306	0.205	0.215	**0.137**	0.386	0.236	0.252	**0.142**	0.322	0.208	0.222	**0.138**
ETT-m2	0.585	0.152	0.145	**0.093**	0.299	0.152	0.161	**0.095**	0.465	0.151	0.192	**0.098**	0.450	0.152	0.166	**0.095**
Missing Ratio	0.4				0.5				0.6				Average			
Model	GP-VAE	CSDI	MIRI	SPIRIT	GP-VAE	CSDI	MIRI	SPIRIT	GP-VAE	CSDI	MIRI	SPIRIT	GP-VAE	CSDI	MIRI	SPIRIT
ETT-h1	0.452	0.542	0.351	**0.278**	0.463	0.655	0.380	**0.301**	0.505	0.918	0.422	**0.337**	0.473	0.705	0.384	**0.305**
ETT-h2	0.380	0.565	0.262	**0.171**	0.375	0.611	0.305	**0.181**	0.390	1.296	0.336	**0.194**	0.382	0.824	0.301	**0.182**
ETT-m1	0.324	0.264	0.232	**0.150**	0.377	0.332	0.298	**0.159**	0.375	0.386	0.319	**0.171**	0.359	0.327	0.283	**0.160**
ETT-m2	0.312	0.198	0.219	**0.102**	0.326	0.221	0.243	**0.109**	0.479	0.308	0.293	**0.116**	0.372	0.242	0.252	**0.109**

Kindly Note: The best results are in **bold**, and the second best are in wavy underline.

4.3 Ablation Studies

In this subsection, we present the ablation study. Our SPIRIT approach has two main contributions: (1) introducing the SPT discrepancy as a proximal term, and (2) removing terms associated with the dissipative structure. Notably, the dissipative structure induced by the Variance Preserving SDE is not appropriate in our setting because the conditional distribution defined by the DDPM underlying SDE is incompatible with the SPIRIT framework. We therefore focus on using $\mathbb{E}_{q'}[\log q']$ to represent the dissipative structure. The corresponding results are reported in Table 3.

From Table 3, we observe consistent performance drops when ablating either the SPT proximal term or the $\mathbb{E}_{q'}[\log q']$ component. This indicates that both introducing SPT and eliminating the dissipative-structure term are important for SPIRIT's overall performance. Moreover, the relative impact differs across datasets. On ETT-h1 and ETT-h2, removing SPT leads to a smaller degradation than removing $\mathbb{E}_{q'}[\log q']$, whereas on ETT-m1 and ETT-m2 the performance drop from removing SPT is larger. This suggests that on larger-scale datasets such as ETT-m1 and ETT-m2, the SPT component contributes

more substantially to accuracy improvements, potentially because larger datasets are more likely to contain outliers. This finding highlights the importance of introducing SPT. Moreover, it provides empirical support for Lemma 3.2 and further motivates the use of the entropy-induced generalized Bregman divergence. Finally, when both components are removed simultaneously, SPIRIT exhibits the most severe degradation, further confirming that SPT and dissipative-structure elimination are complementary and jointly underpin SPIRIT's superior performance.

Table 3: Ablation study results in terms of MAE and MSE.

SPT	w/o $\mathbb{E}_{q'}[\log q']$	ETT-h1				ETT-h2				ETT-m1				ETT-m2			
		MAE		MSE		MAE		MSE		MAE		MSE		MAE		MSE	
		Value	$\Delta(\uparrow)$	Value	$\Delta(\uparrow)$	Value	$\Delta(\uparrow)$	Value	$\Delta(\uparrow)$	Value	$\Delta(\uparrow)$	Value	$\Delta(\uparrow)$	Value	$\Delta(\uparrow)$	Value	$\Delta(\uparrow)$
✓	✗	0.237*	13.1%	0.137*	27.0%	0.153*	16.7%	0.051*	33.5%	0.137*	16.2%	0.047*	25.3%	0.105*	28.1%	0.023*	41.4%
✗	✓	0.227*	8.42%	0.133*	23.3%	0.142*	8.36%	0.046*	20.1%	0.162*	37.0%	0.053*	42.0%	0.128*	55.8%	0.033*	103%
✗	✗	0.573*	173%	0.534*	394%	0.532*	305%	0.449*	1066%	0.529*	347%	0.443*	1091%	0.516*	528%	0.420*	2457%
✓	✓	0.209	-	0.108	-	0.131	-	0.038	-	0.118	-	0.037	-	0.082	-	0.016	-

Kindly Note: Each entry is averaged over $p_{\text{miss}} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. $\Delta(\uparrow)$ denotes the percentage performance degradation relative to SPIRIT. "*" indicates that SPIRIT significantly outperforms the baseline under a paired t-test ($p < 0.05$).

5 Related Works

DMs for Imputation Tasks. DMs have been widely adapted to TSDI because of their strong generative capabilities (Wang et al., 2025a; Cao et al., 2024; Wang et al., 2025c). Existing efforts primarily modify the learning or inference procedures of diffusion models: For instance, Schrödinger-bridge formulations have been introduced to accelerate imputation (Chen et al., 2023), frequency-domain diffusion methods have been introduced to capture periodic patterns (Yang et al., 2024), and mutual-information-based objectives (Liu et al., 2024b; Yu et al., 2025) have been used to redesign the training loss and improve performance, as exemplified by MIRI (Yu et al., 2025). However, it has been observed that the diversity-seeking nature of diffusion sampling can conflict with the accuracy-oriented objective of imputation (Xu et al., 2023b). To address this issue, Chen et al. (2024) studied this issue through a global functional via Wasserstein gradient flow, primarily for tabular data. In contrast, we characterize the local functional (see Equation (5)) induced by DM-based TSDI to improve robustness to non-stationarity and pointwise accuracy. As such, we introduce the SPT discrepancy (Section 3.3), re-derive the imputation procedure (Section 3.4), and develop SPIRIT (Section 3.5), which together form our main theoretical contributions.

Proximal Recursion for Differential Equation-based Algorithms. Jordan et al. (1998) connected the Fokker–Planck equation with Wasserstein proximal recursion. Subsequent work along this line can be broadly grouped into three directions: 1). *Synthesis*, which develops recursive optimization schemes on specific metric spaces (e.g., Wasserstein space) to design new algorithms for sampling (Neklyudov et al., 2023), generation (Xu et al., 2023a), and density estimation (Mokrov et al., 2021); 2). *Constraint handling*, which incorporates constraints via proximal regularization to solve constrained optimization problems and to enable sampling on constrained domains (Shi et al., 2022; Hsieh et al., 2018; Sharrock et al., 2023); and 3). *Theoretical analysis*, which uses this framework to study existing methods, including convergence and stability (Cheng et al., 2024; Fang et al., 2025) as well as population-dynamics modeling (Bunne et al., 2022; Chen, 2023). Our work is most closely related to the *Synthesis* and *Theoretical analysis* lines: we adopt a proximal recursion viewpoint to examine DMs for TSDI, and develop a novel formulation tailored to DM-based TSDI.

6 Conclusions

In this manuscript, we analyzed and improved DM-based TSDI from the perspective of proximal recursion. We reformulated the imputation procedure as a proximal recursion process and identified two performance-limiting components: the Wasserstein-distance term, which reduced robustness to non-stationary features, and the dissipative term, which encouraged diverse imputations. To address these limitations, we relaxed the vanilla Wasserstein distance using a generalized Bregman divergence induced by an entropy functional and removed the dissipative term. These modifications led to SPIRIT, a new DM-based TSDI approach whose effectiveness was validated through extensive experiments. The experimental results demonstrated that SPIRIT achieved competitive imputation accuracy across diverse time-series benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT-5.6 Sol and Claude Opus 5, to assist with editing the manuscript for grammar, clarity, and readability; organizing and formatting the experimental results; and checking the clarity and logical consistency of proofs drafted by the authors. The tools were not used to generate research ideas, formulate hypotheses, design the research methodology or experiments, or implement the proposed method. The authors independently evaluated any AI-generated comments on the proofs, verified the final mathematical arguments, and reviewed and revised all AI-assisted edits. We take full responsibility for the final content of this work, including its claims, proofs, experimental results, and conclusions.

ETHICS STATEMENT

This work studies time-series data imputation using publicly available benchmark datasets and does not involve new data collection, human-subject intervention, or the identification of individuals. Nevertheless, imputed values are model-generated estimates and may be inaccurate or poorly calibrated, particularly under distribution shifts, high missingness, or missingness mechanisms that differ from the experimental settings. Consequently, SPIRIT should not be used as a substitute for observed measurements or expert judgment in high-stakes clinical, financial, industrial, or public-safety applications without application-specific validation, uncertainty assessment, and human oversight. The method may also inherit biases or artifacts present in the training data. Our experiments retain challenging regimes, including outliers and continuous missing blocks, rather than selectively excluding unfavorable cases.

REPRODUCIBILITY STATEMENT

The supplementary material documents the mathematical background in Section B and provides the assumptions and proofs for all main theoretical results in Section C. The complete SPIRIT pseudocode and convergence analysis are given in Section C.6. The benchmark datasets, model hyperparameters, evaluation metrics, and missingness protocol are specified in Section D, with further details in Sections D.1 to D.4. Additional uncertainty-quantification, ablation, sensitivity, outlier-preprocessing, block-missing, convergence, and computational-time results are reported in Section E. Tables identify the evaluated missing ratios or block lengths and state the aggregation and significance-testing conventions where applicable. These materials are intended to support independent reimplementation and verification of the reported claims; results outside the documented datasets, missingness mechanisms, and experimental configurations are not claimed. The codebase for this manuscript is available in the following anonymous github link: https://anonymous.4open.science/r/SPIRITCodeBase-2481.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, Berlin, Germany, 2005.

Luigi Ambrosio, Elia Brué, Daniele Semola, et al. *Lectures on optimal transport*, volume 130. Springer, Berlin, Germany, 2021.

Francis Bach. *Learning theory from first principles*. MIT press, Cambridge, USA, 2024.

Charles Blair. Problem complexity and method efficiency in optimization (as Nemirovsky and db Yudin). *SIAM Review*, 27(2):264, 1985.

Mathieu Blondel, Felipe Llinares-López, Robert Dadashi, Léonard Hussenot, and Matthieu Geist. Learning energy networks with generalized Fenchel-Young losses. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 35, pp. 12516–12528, 2022.

Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and preconditioned gradient descent in Wasserstein space. In *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1–64, 2024.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Paul Brunzema, Mikkel Jordahn, John Willes, Sebastian Trimpe, Jasper Snoek, and James Harrison. Bayesian optimization via continual Variational last layer training. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–30, 2025.

Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal optimal transport modeling of population dynamics. In *Int. Conf. Artif. Intell. Stat.*, pp. 6511–6528. PMLR, 2022.

Kenneth F. Caluya and Abhishek Halder. Proximal recursion for solving the Fokker-Planck equation. In *Proc. Am. Control Conf.*, pp. 4098–4103, 2019.

Kenneth F. Caluya and Abhishek Halder. Gradient flow algorithms for density propagation in stochastic systems. *IEEE Trans. Autom. Control.*, 65(10):3991–4004, 2020.

Kenneth F Caluya and Abhishek Halder. Wasserstein proximal algorithms for the Schrödinger bridge problem: Density control with nonlinear drift. *IEEE Trans. Autom. Control.*, 67(3):1163–1178, 2021.

Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. A survey on generative diffusion models. *IEEE Trans. Knowl. Data Eng.*, 36(7):2814–2830, 2024.

Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood training of Schrödinger bridge using forward-backward SDEs theory. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–27, 2022.

Yongxin Chen. Density control of interacting agent systems. *IEEE Trans. Autom. Control.*, 69(1): 246–260, 2023.

Yu Chen, Wei Deng, Shikai Fang, Fengpei Li, Nicole Tianjiao Yang, Yikai Zhang, Kashif Rasul, Shandian Zhe, Anderson Schneider, and Yuriy Nevmyvaka. Provably convergent Schrödinger bridge with applications to probabilistic time series imputation. In *Proc. Int. Conf. Mach. Learn.*, pp. 4485–4513, 2023.

Zhichao Chen, Haoxuan Li, Fangyikang Wang, Odin Zhang, Hu Xu, Xiaoyu Jiang, Zhihuan Song, and Hao Wang. Rethinking the diffusion models for missing data imputation: A gradient flow perspective. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 37, pp. 112050–112103, 2024.

Xiuyuan Cheng, Jianfeng Lu, Yixin Tan, and Yao Xie. Convergence of flow-based generative models via proximal gradient descent in Wasserstein space. *IEEE Trans. Inf. Theory*, 70(11):8087–8106, 2024.

Jonathan Crabbé, Nicolas Huynh, Jan Pawel Stanczuk, and Mihaela van der Schaar. Time series diffusion in the frequency domain. In *Proc. Int. Conf. Mach. Learn.*, pp. 9407–9438, 2024.

Hanze Dong, Xi Wang, LIN Yong, and Tong Zhang. Particle-based variational inference with preconditioned functional gradient flow. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–26, 2022.

Wenjie Du, Jun Wang, Linglong Qian, Yiyuan Yang, Fanxing Liu, Zepu Wang, Zina Ibrahim, Haoxin Liu, Zhiyuan Zhao, Yingjie Zhou, Wenjia Wang, Kaize Ding, Yuxuan Liang, B. Aditya Prakash, and Qingsong Wen. TSI-Bench: Benchmarking time series imputation. *arXiv preprint arXiv:2406.12747*, pp. 1–46, 2024. URL https://arxiv.org/abs/2406.12747.

Wenjie Du, Yiyuan Yang, Tianxiang Zhan, and Qingsong Wen. End-to-end learning for partially-observed time series with PyPOTS. In *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 13283–13284, 2026.

Zhenghan Fang, Mateo Díaz, Sam Buchanan, and Jeremias Sulam. Beyond scores: Proximal diffusion models. In *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1–50, 2025.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *Proc. Int. Conf. Mach. Learn.*, pp. 3186–3197. PMLR, 2021.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. GP-VAE: Deep probabilistic time series imputation. In *Int. Conf. Artif. Intell. Stat.*, pp. 1651–1661. PMLR, 2020.

Chengyue Gong, Lemeng Wu, and Qiang Liu. How to fill the optimum set? Population gradient descent with harmless diversity. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–15, 2022.

James Harrison, John Willes, and Jasper Snoek. Variational Bayesian last layers. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–30, 2024.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin dynamics. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 31, pp. 1–10, 2018.

Daniel Jarrett, Bogdan C Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. Hyperimpute: Generalized iterative imputation with automatic model selection. In *Proc. Int. Conf. Mach. Learn.*, pp. 9916–9937, 2022.

Rie Johnson and Tong Zhang. A framework of composite functional gradient methods for generative adversarial models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):17–32, 2021.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

Trent Kyono, Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. MIRACLE: causally-aware imputation via learning missing data mechanisms. *Proc. Adv. Neural Inf. Process. Syst.*, pp. 23806–23817, 2021.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 33, pp. 8847–8860, 2020.

Haoxuan Li, Chunyuan Zheng, Shuyi Wang, Kunhan Wu, Eric Wang, Peng Wu, Zhi Geng, Xu Chen, and Xiao-Hua Zhou. Relaxing the accurate imputation assumption in doubly robust learning for debiased collaborative filtering. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–13, 2024.

Zhouchen Lin, Huan Li, and Cong Fang. *Alternating direction method of multipliers for machine learning*. Springer, Berlin, Germany, 2022.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–28, 2023.

Diju Liu, Yalin Wang, Chenliang Liu, Xiaofeng Yuan, Kai Wang, and Chunhua Yang. Scope-free global multi-condition-aware industrial missing data imputation framework via diffusion Transformer. *IEEE Trans. Knowl. Data Eng.*, pp. 1–12, 2024a.

Jiashuo Liu, Jiayun Wu, Bo Li, and Peng Cui. Distributionally robust optimization with data geometry. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 35, pp. 33689–33701, 2022a.

Mingzhe Liu, Han Huang, Hao Feng, Leilei Sun, Bowen Du, and Yanjie Fu. PriSTI: A conditional diffusion framework for spatiotemporal imputation. In *Proc. Int. Conf. Data Eng.*, pp. 1927–1939, 2023a.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 30, pp. 1–15, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 29, pp. 1–13, 2016.

Song Liu, Jiahao Yu, Jack Simons, Mingxuan Yi, and Mark Beaumont. Minimizing f-divergences by interpolating velocity fields. In *Proc. Int. Conf. Mach. Learn.*, pp. 32308–32331. PMLR, 2024b.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–33, 2023b.

Yixin Liu, Thalaiyasingam Ajanthan, Hisham Husain, and Vu Nguyen. Self-supervision improves diffusion models for tabular data imputation. In *Proc. ACM Int. Conf. Inf. Knowl. Manag.*, pp. 1–10, 2024c.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary Transformers: Exploring the stationarity in time series forecasting. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 35, pp. 9881–9893, 2022b.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–23, 2025.

Qianli Ma, Sen Li, and Garrison W. Cottrell. Adversarial joint-learning recurrent neural network for incomplete time series classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4):1765–1776, 2022.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale Wasserstein gradient flows. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34, pp. 15243–15256, 2021.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *Proc. Int. Conf. Mach. Learn.*, pp. 7130–7140, 2020.

Kirill Neklyudov, Jannes Nys, Luca Thiede, Juan Carrasquilla, Qiang Liu, Max Welling, and Alireza Makhzani. Wasserstein quantum Monte Carlo: A novel approach for solving the quantum many-body Schrödinger equation. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 36, pp. 63461–63482, 2023.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with Transformers. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–24, 2023.

Bernt Øksendal. *Stochastic differential equations*. Springer, Berlin, Germany, 2003.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Commun. Partial Differ. Equ.*, 26(1-2):101–174, 2001.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 32, pp. 1–12, 2019.

Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat. Mach. Intell.*, 2(7):369–375, 2020.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proc. VLDB Endow.*, 17(9):2363–2377, 2024.

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, Berlin, Germany, 2015.

Filippo Santambrogio. {Euclidean, Metric, and Wasserstein} gradient flows: an overview. *Bull. Math. Sci.*, 7:87–154, 2017.

Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, Cambridge, UK, 2019.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with Python. *SciPy 2010*, 2010. doi: 10.25080/Majora-92bf1922-011. URL https://doi.org/10.25080/Majora-92bf1922-011.

Louis Sharrock, Lester Mackey, and Christopher Nemeth. Learning rate free sampling in constrained domains. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 36, pp. 65380–65415, 2023.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *Proc. Int. Conf. Mach. Learn.*, pp. 31016–31029. PMLR, 2023.

Lifeng Shen, Weiyu Chen, and James Kwok. Multi-resolution diffusion models for time series forecasting. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–19, 2024.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. In *Proc. Int. Conf. Mach. Learn.*, pp. 4644–4653, 2018.

Jiaxin Shi, Chang Liu, and Lester Mackey. Sampling with Mirrored Stein Operators. *Proc. Int. Conf. Learn. Represent.*, pp. 1–26, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–20, 2021.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proc. Conf. Uncertainty in Artificial Intelligence*, pp. 574–584, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–36, 2020b.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Proc. Adv. Neural Inf. Process. Syst.*, pp. 24804–24816, 2021.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

Fangyikang Wang, Hubery Yin, Shaobin Zhuang, Huminhao Zhu, Yinan Li, Lei Qian, Chao Zhang, Hanbin Zhao, Hui Qian, and Chen Li. Efficiently access diffusion fisher: Within the outer product span space. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–25, 2025a.

Hao Wang, Zhengnan li, Zhichao Chen, Xu Chen, Shuting He, Guangyi Liu, Haoxuan Li, and Zhouchen Lin. Iterative missing data imputation with model form adaptation and non-missing feature supervision. In *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1–25, 2025b.

Hao Wang, Zhengnan li, Haoxuan Li, Xu Chen, Mingming Gong, Bin Chen, and Zhichao Chen. Optimal transport for time series imputation. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–25, 2025c.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven C. H. Hoi. ETSformer: Exponential smoothing Transformers for time-series forecasting. *arxiv preprint*, pp. 1–18, 2022. URL https://arxiv.org/abs/2202.01381.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34, pp. 1–20, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–23, 2023.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by jko scheme. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 36, pp. 47379–47405, 2023a.

Jingwen Xu, Fei Lyu, and Pong C Yuen. Density-aware temporal attentive step-wise diffusion model for medical time series imputation. In *Proc. ACM Int. Conf. Inf. Knowl. Manag.*, pp. 2836–2845, 2023b.

Hanchen Yang, Jiannong Cao, Wengen Li, Yu Yang, Xiaoyi Li, Lingbai Kong, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. Towards robust and interpretable spatial-temporal graph modeling for traffic prediction. *ACM Trans. Knowl. Discov. Data.*, 19(9):1–20, 2025a.

Jie Yang, Kexin Zhang, Guibin Zhang, Philip S. Yu, and Kaize Ding. Glocal information bottleneck for time series imputation. In *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1–33, 2025b.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4):1–39, 2023.

Xinyu Yang, Yu Sun, Xiaojie Yuan, and Xinyang Chen. Frequency-aware generative models for multivariate time series imputation. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 37, pp. 52595–52623, 2024.

Jiahao Yu, Qizhen Ying, Leyang Wang, Ziyue Jiang, and Song Liu. Missing data imputation by reducing mutual information with rectified flows. In *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1–29, 2025.

Jiahao Yu, Qizhen Ying, Leyang Wang, Ziyue Jiang, and Song Liu. Missing data imputation by reducing mutual information with rectified flows. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 38, pp. 80324–80352, 2026.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers effective for time series forecasting? In *Proc. AAAI Conf. Artif. Intell.*, volume 37, pp. 11121–11128, 2023.

Fengming Zhang, Wenjie Du, Huan Zhang, Ke Yu, and Shen Qu. HELIX: Hybrid encoding with learnable identity and cross-dimensional synthesis for time series imputation. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–30. PMLR, 2026.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–33, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–21, 2023.

He Zhao, Ke Sun, Amir Dezfouli, and Edwin V Bonilla. Transformed distribution matching for missing value imputation. In *Proc. Int. Conf. Mach. Learn.*, pp. 42159–42186, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient Transformer for long sequence time-series forecasting. In *Proc. AAAI Conf. Artif. Intell.*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Xue Wang, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, and Rong Jin. FiLM: Frequency improved Legendre memory model for long-term time series forecasting. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 35, pp. 12677–12690, 2022.

CONTENTS FOR APPENDIX

A Nomenclature **17**

A.1 Abbreviations . . . 17

A.2 Data and Time Indices . . . 17

A.3 Distributions and Transport . . . 18

A.4 Continuous Dynamics and Variational Quantities . . . 18

A.5 SPIRIT Optimization and Algorithm . . . 18

B Additional Background Knowledge **20**

C Theoretical Derivation **21**

C.1 Derivation and Discussions of Proposition 3.1 . . . 22

C.2 Derivation of Lemma 3.2 . . . 27

C.3 Derivation of Proposition 3.3 . . . 29

C.4 Derivation of Proposition 3.4 . . . 32

C.5 Derivation of Proposition 3.5 and Related Discussions . . . 33

C.6 Discussions of the Convergence Property for SPIRIT Framework . . . 34

D Detailed Experimental Protocols **37**

D.1 Dataset Information . . . 37

D.2 Model Hyperparameters . . . 37

D.3 Evaluation Metrics . . . 38

D.4 Simulation of MCAR Scenario . . . 39

E Additional Experimental Results **39**

E.1 Statistics of Time-series Data . . . 39

E.2 Discussion of Outlier Preprocessing Strategy . . . 39

E.3 Sensitivity Analysis . . . 40

E.4 Additional Results on Block Missing . . . 41

E.5 Empirical Time Complexity . . . 42

E.6 Empirical Convergence Analysis . . . 42

F Discussions on Limitations & Future Research Directions **44**

A NOMENCLATURE

This section summarizes the abbreviations and mathematical notation used throughout the paper.

A.1 ABBREVIATIONS

CRPS	Continuous ranked probability score.
DM	Diffusion model.
DSM	Denoising score matching.
MCAR	Missing completely at random.
MAE, MSE	Mean absolute error, and mean squared error, respectively.
ODE	Ordinary differential equation.
OT	Optimal transport.
PDE	Partial differential equation.
PDF	Probability density function.
SPIRIT	Semi-Proximal Transport Regularized time-series Imputation, our proposed framework.
SPT	Semi-proximal transport, the transport discrepancy used as the proximal term in SPIRIT.
TSDI	Time-series data imputation.
VE-SDE	Variance-exploding stochastic differential equation.
VP-SDE	Variance-preserving stochastic differential equation.

A.2 DATA AND TIME INDICES

N, T, D	Number of samples, sequence length, and number of features, respectively. In particle-based derivations, N also denotes the number of particles representing an empirical measure.
n, t, d	Sample, dataset-time, and feature indices, respectively.
$\tau, \mathscr{T}$	Continuous diffusion or optimization time and its terminal value, respectively.
$\mathbf{X}^{\text{ideal}}$	Fully observed time-series data in $\mathbb{R}^{N\times T\times D}$.
$\mathbf{X}^{\text{obs}}$	Partially observed time-series data.
$\mathbf{X}^{\text{imp}}$	Imputed time-series data.
$\mathbf{x}^{\text{ideal}}, \mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{imp}}$	A fully observed, partially observed, and imputed sample, respectively.
$\mathscr{X}, \mathscr{O}$	Random variables corresponding to the imputed and observed components, respectively; that is, $\mathscr{X} = \mathbf{x}^{\text{imp}}$ and $\mathscr{O} = \mathbf{x}^{\text{obs}}$ in Proposition 3.5.
$\mathscr{S}(\mathscr{X}, \mathscr{O}), m(\mathscr{X})$	Conditional and marginal score functions, respectively, where $\mathscr{S}(\mathscr{X}, \mathscr{O}) = \nabla_{\mathscr{X}} \log p(\mathscr{X} \mid \mathscr{O})$ and $m(\mathscr{X}) = \nabla_{\mathscr{X}} \log p(\mathscr{X})$.
$\mathbf{M}$	Binary missingness mask; $\mathbf{M}_{n,t,d} = 1$ indicates a missing entry.
$\odot$	Hadamard (element-wise) product.
p_{miss}	Proportion of entries designated as missing in an evaluation setting.
$\mathscr{Q}$	Quantile operator; $\widehat{\mathscr{Q}}_{\alpha,n,t,d}$ denotes the empirical α-quantile of stochastic imputations at entry (n, t, d), while $\mathscr{Q}_{1,t}$ and $\mathscr{Q}_{3,t}$ denote the first and third quartiles used in the IQR analysis.
T_{patch}	Length of a time-series patch used by the score network; denoted by T when no ambiguity with sequence length arises.

A.3 Distributions and Transport

$p(\cdot)$	Data density; $p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}})$ denotes the conditional density of the missing entries given the observed entries.
$q(\cdot)$	Current or base density of imputed samples in a proximal recursion step.
$q'(\cdot)$	Candidate density produced by a proximal recursion step.
$q_\tau(\cdot)$	Time-indexed density transported by a continuous flow.
$q_\sigma(\widehat{\mathbf{x}}^{\text{imp}} \mid \mathbf{x}^{\text{imp}})$	Gaussian perturbation kernel used in DSM, with noise scale σ.
μ, ν	Source and target probability measures, respectively, in $\mathcal{P}_2(\mathbb{R}^D)$.
$\widetilde{\nu}$	Contaminated target measure, typically $\widetilde{\nu} = (1-\zeta)\nu + \zeta\delta_{\mathbf{z}}$.
ρ, σ	Generic intermediate or marginal measures. In the SPT decomposition, the intermediate target marginal may be denoted by σ; this use is distinct from the DSM noise scale.
$\mathcal{P}_2(\mathbb{R}^D)$	Space of probability measures on $\mathbb{R}^D$ with finite second moment.
π	Transport plan or coupling; $\pi_{\mathbf{x}}$ and $\pi_{\mathbf{y}}$ denote its first and second marginals, respectively.
$\Pi(\mu, \nu)$	Set of couplings with marginals μ and ν; $\Pi(\mu)$ fixes only the first marginal.
$\delta_{\mathbf{z}}$	Dirac measure concentrated at $\mathbf{z}$.
ζ	Contamination proportion in the outlier model, with $\zeta \in (0, 1)$.
$\mathbb{W}, \mathbb{W}_2^2(\mu, \nu)$	Wasserstein distance and the squared 2-Wasserstein distance between μ and ν, respectively.
$\mathbb{S}, \mathbb{S}(\mu, \nu)$	SPT discrepancy, which fixes the source marginal and relaxes the target-marginal constraint through a divergence penalty.
$\mathbb{E}, \mathbb{E}_q$	Expectation operator and expectation with respect to the distribution q, respectively.
$D_\psi(\rho, \nu)$	Generalized Bregman divergence induced by the potential ψ.
ψ	Strictly convex Bregman potential; SPIRIT uses the negative-entropy potential $\psi(\rho) = \int \rho(\mathbf{x})[\log \rho(\mathbf{x}) - 1]\, \mathrm{d}\mathbf{x}$.
$\mathbf{z}, \mathbf{y}^*$	Outlier location and a reference point in the support of the nominal target measure, respectively, in the robustness analysis.
$D(\mathbf{z}), C(\zeta)$	Outlier-dependent transport cost and contamination-dependent constant used in the SPT robustness bound, respectively.

A.4 Continuous Dynamics and Variational Quantities

$b(\mathbf{x}_\tau, \tau)$	Drift term of the stochastic differential equation.
$\sigma(\mathbf{x}_\tau, \tau)$	Volatility term of the stochastic differential equation; when written without arguments in DSM, σ denotes the Gaussian noise scale.
$f(\mathbf{x}, \tau), g(\tau)$	Drift and scalar diffusion coefficients in the unified diffusion-SDE notation.
$\mathrm{d}W_\tau$	Wiener-process increment.
$v_\tau(\mathbf{x})$	Velocity field in the continuity equation; $v_\tau^\star$ denotes the corresponding optimal or variationally induced field.
$\mathfrak{T}_\eta(\mathbf{x})$	Small-step transport map $\mathfrak{T}_\eta(\mathbf{x}) = \mathbf{x} - \eta\nabla_{\mathbf{x}}\varphi_q(\mathbf{x})$ induced by the Wasserstein-gradient-flow direction.
$g_\tau(\mathbf{x})$	Centered reaction field governing changes in particle weights; this notation is distinct from the scalar SDE diffusion coefficient $g(\tau)$.
$\mathcal{F}(q)$	Generic energy functional of a probability density q.
$\delta_q\mathcal{F}(q)$	First variation (functional derivative) of $\mathcal{F}$ with respect to q.
$\mathcal{J}(q)$	Energy functional used in the recursive-imputation convergence analysis.

A.5 SPIRIT Optimization and Algorithm

$\phi(\mathbf{x}^{\text{imp}})$	SDE-dependent dissipative-structure term identified in DM-based imputation and omitted from the SPIRIT objective.
η	Step size of the proximal recursion and recursive-imputation updates.
s_θ	Score network with parameters θ, trained to approximate $\nabla_{\mathbf{x}^{\text{imp}}} \log p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}})$.

$\mathcal{L}^{\mathrm{DSM}}(\theta)$	DSM objective used to train the score network.
$\widehat{\mathbf{x}}^{\mathrm{imp}}$	Noise-perturbed imputed sample used for DSM training.
$w_i, \widehat{w}_i$	Normalized weight of particle $\mathbf{x}_i$ and its intermediate unnormalized update, respectively; $w_i \geq 0$ and $\sum_{i=1}^{N} w_i = 1$.
$\boldsymbol{T}_{\mathbf{x}^{\mathrm{imp}}}$	Score-guided direction used to update imputed particle locations.
$\boldsymbol{T}_w$	Score-norm-guided direction used to update particle log-weights.
lr	Learning rate used to optimize the score-network parameters.
H_{s_θ}	Hidden dimension of the score network.
$\mathcal{E}_{\mathrm{score}}, \mathcal{E}_{\mathrm{imp}}$	Numbers of score-learning and recursive-imputation iterations, respectively.
`ApplyGrad`	Gradient-based update operator with inputs specifying an update direction, a step size, and the variable to be updated.

B Additional Background Knowledge

In this section, we present the background material used to derive the main theoretical results in this manuscript.

Differential Equation: Suppose we have the following SDE, which is known as an Itô process (Øksendal, 2003):

$$\mathrm{d}\mathbf{x} = f(\mathbf{x}, \tau)\mathrm{d}\tau + g(\tau)\mathrm{d}W_\tau. \tag{15}$$

Denote the marginal distribution of $\mathbf{x}$ at time τ as $q_\tau(\mathbf{x})$. The $q_\tau(\mathbf{x})$ satisfies the following partial differential equation (PDE), which is known as the Fokker-Planck equation (Särkkä & Solin, 2019):

$$\frac{\partial q_\tau(\mathbf{x})}{\partial \tau} = -\nabla_{\mathbf{x}} \cdot [f(\mathbf{x}, \tau) q_\tau(\mathbf{x})] + \frac{1}{2}\nabla_{\mathbf{x}} \cdot [g^2(\tau)\nabla_{\mathbf{x}} q_\tau(\mathbf{x})]. \tag{16}$$

A classical solution to Equation (16) requires $q_\tau(\mathbf{x})$ to be differentiable in both τ and $\mathbf{x}$. To address this issue, rather than working with classical (smooth) solutions of Equation (16), we consider a weak, measure-valued representation of $q_\tau(\mathbf{x})$ using a finite set of particles $\{\mathbf{x}_i\}_{i=1}^N$ as follows (Liu & Wang, 2016; Liu, 2017):

$$q_\tau(\mathbf{x}) \approx \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_i}. \tag{17}$$

Therefore, for each particle, we solve the PDE in Equation (16) via the ordinary differential equation (ODE) below. The method amounts to updating the spatial positions $\{\mathbf{x}_i\}_{i=1}^N$, and it provides a weak solution to Equation (16), as described in Section 4.1.2 of reference (Santambrogio, 2015):

$$\frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}\tau} = f(\mathbf{x}_i, \tau) - \frac{1}{2}g^2(\tau)\nabla_{\mathbf{x}} \log q_\tau(\mathbf{x}_i). \tag{18}$$

While the standard Itô process evolves the distribution $q_\tau(\mathbf{x})$ by transporting the particle locations $\{\mathbf{x}_i\}_{i=1}^N$, an alternative approach is to steer the density by adjusting the particle weights. Specifically, we formulate the empirical approximation of $q_\tau(\mathbf{x})$ as a weighted sum of Dirac measures:

$$q_\tau(\mathbf{x}) \approx \sum_{i=1}^N w_i \delta_{\mathbf{x}_i}, \quad \text{s.t.} \quad \underbrace{\sum_{i=1}^N w_i = 1, \quad w_i \geq 0}_{[w_1,\ldots,w_N]\in\Delta^{N-1}}. \tag{19}$$

In this framework, the shape of $q_\tau(\mathbf{x})$ is controlled by the time-varying weights $\{w_i\}_{i=1}^N$ while the particle locations remain fixed. The evolution of the probability density is governed by the following integro-differential equation:

$$\frac{\partial q_\tau(\mathbf{x})}{\partial \tau} = [\int g_\tau(\mathbf{x}')q_\tau(\mathbf{x}')\mathrm{d}\mathbf{x}' - g_\tau(\mathbf{x})]q_\tau(\mathbf{x}), \tag{20}$$

where $g_\tau : \mathbb{R}^D \to \mathbb{R}$ denotes the scalar function driving the reweighting process. Consequently, the continuous-time dynamics of the individual weights $\{w_i\}_{i=1}^N$ follow the ODE:

$$\frac{\mathrm{d}w_i}{\mathrm{d}\tau} = w_i[\int g_\tau(\mathbf{x})q_\tau(\mathbf{x})\mathrm{d}\mathbf{x} - g_\tau(\mathbf{x}_i)], \quad \forall i \in \{1, \ldots, N\}. \tag{21}$$

Using the fact that $\frac{\mathrm{d}\log w_i}{\mathrm{d}\tau} = \frac{1}{w_i}\frac{\mathrm{d}w_i}{\mathrm{d}\tau}$, Equation (21) can be further reformulated as follows:

$$\frac{\mathrm{d}\log w_i}{\mathrm{d}\tau} = [\int g_\tau(\mathbf{x})q_\tau(\mathbf{x})\mathrm{d}\mathbf{x} - g_\tau(\mathbf{x}_i)], \quad \forall i \in \{1, \ldots, N\}. \tag{22}$$

Notably, Equation (22) is the normalized equation for $\{w_i\}_{i=1}^N$, and its unnormalized counterpart is given by:

$$\frac{\mathrm{d}\log w_i}{\mathrm{d}\tau} = -g_\tau(\mathbf{x}_i), \quad \forall i \in \{1, \ldots, N\}. \tag{23}$$

Wasserstein Distance and Its Dynamic Formulation: Let $\mathcal{P}_2(\mathbb{R}^D)$ denote the space of probability measures on $\mathbb{R}^D$ with finite second moments, i.e., $\mathcal{P}_2(\mathbb{R}^D) \coloneqq \{\mu \in \mathcal{M}(\mathbb{R}^D) | \int \|\mathbf{x}\|^2 \mathrm{d}\mu(\mathbf{x}) < \infty\}$, where $\mathcal{M}(\mathbb{R}^D)$ denotes the set of all probability measures on $\mathbb{R}^D$. Considering any two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^D)$, we define the p-Wasserstein distance as follows (Villani, 2009):

$$\mathbb{W}_p^p(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \|\mathbf{x} - \mathbf{y}\|^p \mathrm{d}\pi(\mathbf{x}, \mathbf{y}). \tag{24}$$

Here, $\Pi(\mu, \nu)$ represents the collection of all joint distributions supported on $\mathbb{R}^D \times \mathbb{R}^D$ with marginals μ and ν. The integral formulation corresponds to the Kantorovich optimal transport problem, where the minimizer π^* represents the optimal transportation plan.

Even though the Kantorovich optimal transport problem provides a static view, the 2-Wasserstein distance allows for a dynamic interpretation via the *Benamou-Brenier formula* (Ambrosio et al., 2021). Specifically, the Benamou-Brenier formula considers a continuous curve of densities $(q_\tau)_{\tau \in [0,1]} : \mathbb{R}^D \to \mathbb{R}^+$ linking μ and ν, subject to the law of mass conservation described by the continuity equation as follows (Santambrogio, 2017):

$$\frac{\partial q_\tau(\mathbf{x})}{\partial \tau} + \nabla \cdot [q_\tau(\mathbf{x}) v_\tau(\mathbf{x})] = 0, \tag{25}$$

where $v_\tau : \mathbb{R}^D \to \mathbb{R}^D$ is the velocity field transporting the mass. The squared Wasserstein distance is identified as the minimal kinetic energy required for this transport:

$$\mathbb{W}_2^2(\mu, \nu) = \inf_{q_\tau, v_\tau} \int_0^1 \int_{\mathbb{R}^D} \|v_\tau(\mathbf{x})\|^2 q_\tau(\mathbf{x}) \mathrm{d}\mathbf{x} \mathrm{d}\tau, \tag{26}$$

where the 'inf' operator is taken over all pairs (q_τ, v_τ) satisfying Eq. (25) with boundary conditions $q_0 = \mu$ and $q_1 = \nu$. Based on this, denote the optimal transportation map $\boldsymbol{T}^\star : \mathbb{R}^D \to \mathbb{R}^D$ with an infinitesimal increment η as follows:

$$\boldsymbol{T}^\star(\mathbf{x}) \coloneqq \mathbf{x} + \eta v_\tau^*(\mathbf{x}), \tag{27}$$

we can reformulate the $\mathbb{W}_2^2(\mu, \nu)$ based on Equations (24) and (26) as follows:

$$\mathbb{W}_2^2(\mu, \nu) = \int \|\mathbf{x} - \boldsymbol{T}^\star(\mathbf{x})\|_2^2 \mathrm{d}\mu(\mathbf{x}) = \eta^2 \int \|v_\tau^*(\mathbf{x})\|_2^2 \mathrm{d}\mu(\mathbf{x}). \tag{28}$$

Functional Derivative: Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^D) \to \mathbb{R}^+$ be a functional of a PDF $q : \mathbb{R}^D \to \mathbb{R}^+$ in $\mathcal{P}_2(\mathbb{R}^D)$. To define the functional derivative, consider a small perturbation $h(\mathbf{x})$ of the PDF $q(\mathbf{x})$. The change in the functional $\mathcal{F}[q(\mathbf{x})]$ can be expressed via a linear expansion with the higher-order term $o(\|h(\mathbf{x})\|)$:

$$\mathcal{F}[q(\mathbf{x}) + h(\mathbf{x})] = \mathcal{F}[q(\mathbf{x})] + \mathrm{d}\mathcal{F}[h(\mathbf{x})] + o(\|h(\mathbf{x})\|), \text{where} \mathrm{d}\mathcal{F}[h(\mathbf{x})] \coloneqq \int h(\mathbf{x}) \delta_{q(\mathbf{x})} \mathcal{F}[q(\mathbf{x})] \mathrm{d}\mathbf{x}. \tag{29}$$

Here, $\mathrm{d}\mathcal{F}[h(\mathbf{x})]$ denotes the first variation of $\mathcal{F}[q(\mathbf{x})]$ in the direction $h(\mathbf{x})$, and the function $\delta_{q(\mathbf{x})} \mathcal{F}[q(\mathbf{x})]$ serves as the gradient of the functional with respect to the L^2 inner product.

C Theoretical Derivation

Organization of Section C. In this section, we provide detailed derivations for the theoretical results presented in the main manuscript. Specifically, in Section C.1, we prove Proposition 3.1, which characterizes what DMs implicitly do during the data-imputation process and highlights the issues that motivate this work. Building on this result, in Section C.2 we prove Lemma 3.2, establishing the robustness of the proposed SPT framework to outliers, in contrast to the vanilla OT formulation. We then introduce a novel functional for missing-data imputation and derive the coordinate-wise first-order updates for its splitting surrogate subproblems in Section C.3. Since the resulting update rule may not be well defined, we further study sufficient conditions to ensure well-posedness in Section C.4. Finally, to complete the overall workflow, we derive the learning objective for $\nabla \log p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}})$ in Section C.5 and present the corresponding block-wise and conditional continuous-time convergence analysis in Section C.6. In addition, to clarify the logical flow of the manuscript, we summarize the relationships among the motivating issues, the main theoretical results, their corresponding derivations in the appendix, and the overall convergence analysis in Figure 4.

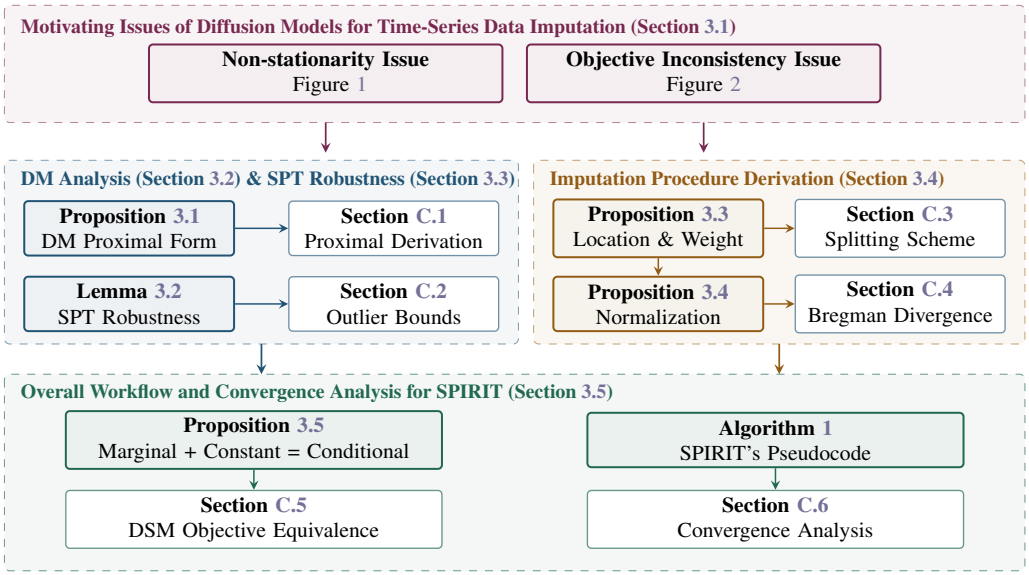


Figure 4: Logical dependency diagram of the theoretical contributions in Sections 3 and C. Arrows denote proof dependencies and workflow progression.

C.1 Derivation and Discussions of Proposition 3.1

C.1.1 Derivation of Proposition 3.1

Proposition (3.1). The imputation process for DMs can be formulated as an iterative solution of the upper bound of the following optimization problem in a proximal operator form:

$$\inf_{q'} \; -\mathbb{E}_{q'}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})] + \phi(\mathbf{x}^{\text{imp}}) + \frac{1}{\eta}\mathbb{W}_2^2(q', q), \tag{30}$$

where we abbreviate the candidate distribution (distribution for current iteration) $q'(\mathbf{x}^{\text{imp}})$ and the base distribution (distribution for previous iteration) $q(\mathbf{x}^{\text{imp}})$ as q' and q, respectively. The term $\frac{1}{\eta}$ is a predefined positive term determined by the noise schedule of DMs, and the term $\phi(\mathbf{x}^{\text{imp}})$ denotes the dissipative structure-related term, which depends on the underlying SDE and is specified as follows:

- Variance Preserving SDE (VP-SDE): $\phi(\mathbf{x}^{\text{imp}}) = \frac{1}{2}\mathbb{E}_{q'}[\log q'(\mathbf{x}^{\text{imp}})] - \frac{1}{4}\mathbb{E}_{q'}[\|\mathbf{x}^{\text{imp}}\|_2^2]$, and $\eta = \beta(\tau)$.
- Variance Exploding SDE (VE-SDE): $\phi(\mathbf{x}^{\text{imp}}) = \frac{1}{2}\mathbb{E}_{q'}[\log q'(\mathbf{x}^{\text{imp}})]$, and $\eta = \frac{1}{2}\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}$.

Proof. We begin by considering a more general problem (we omit the time index τ in the Fokker–Planck equation given by Equation (16)):

$$\inf_{q'(\mathbf{x})} \quad \mathcal{F}[q'(\mathbf{x})] + \frac{1}{2\eta}\mathbb{W}_2^2(q'(\mathbf{x}), q(\mathbf{x})), \tag{31}$$

where the functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^D) \to [0, +\infty)$ has the following form:

$$\mathcal{F}[q(\mathbf{x})] = \int q(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{32}$$

and $q'(\mathbf{x})$ is obtained from $q(\mathbf{x})$ by evolving the following continuity equation for one time step of length η:

$$\frac{\partial q(\mathbf{x})}{\partial \tau} = -\nabla_{\mathbf{x}} \cdot [q(\mathbf{x}) v_\tau(\mathbf{x})] \quad \Longrightarrow \quad q'(\mathbf{x}) = q(\mathbf{x}) - \eta \nabla_{\mathbf{x}} \cdot [q(\mathbf{x}) v_\tau(\mathbf{x})] + o(\eta). \tag{33}$$

On this basis, for the 2-Wasserstein distance, we have the following inequality:

$$\mathbb{W}_2^2(q'(\mathbf{x}), q(\mathbf{x})) = \int q(\mathbf{x})\|\mathbf{x} - \boldsymbol{T}^\star(\mathbf{x})\|_2^2 \mathrm{d}\mathbf{x} = \eta^2 \int q(\mathbf{x})\|v_\tau^*(\mathbf{x})\|_2^2 \mathrm{d}\mathbf{x} \leq \eta^2 \int q(\mathbf{x})\|v_\tau(\mathbf{x})\|_2^2 \mathrm{d}\mathbf{x}, \tag{34}$$

where $v_\tau(\mathbf{x})$ is the non-optimal transportation velocity field. Meanwhile, Equation (31) can be reformulated as follows:

$$\inf_{q'(\mathbf{x})} \quad \mathcal{F}[q'(\mathbf{x})] - \underbrace{\mathcal{F}[q(\mathbf{x})]}_{\text{constant}} + \frac{1}{2\eta}\mathbb{W}_2^2(q'(\mathbf{x}), q(\mathbf{x})). \tag{35}$$

Thus, we have the following upper bound for Equation (31) based on Equations (34) and (35):

$$\begin{aligned}
&\inf_{q'(\mathbf{x})} \quad \mathcal{F}[q'(\mathbf{x})] - \underbrace{\mathcal{F}[q(\mathbf{x})]}_{\text{constant}} + \frac{1}{2\eta}\mathbb{W}_2^2(q'(\mathbf{x}), q(\mathbf{x})) \\
\overset{\text{(i)}}{\Rightarrow} &\inf_{v_\tau(\mathbf{x})} \quad \cancel{\mathcal{F}[q(\mathbf{x})]} - \eta \int \nabla_{\mathbf{x}} \cdot [q(\mathbf{x})v_\tau(\mathbf{x})]\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}\mathrm{d}\mathbf{x} - \cancel{\mathcal{F}[q(\mathbf{x})]} + \frac{\eta}{2}\int q(\mathbf{x})\|v_\tau(\mathbf{x})\|_2^2 \mathrm{d}\mathbf{x} \\
\overset{\text{(ii)}}{\Rightarrow} &\inf_{v_\tau(\mathbf{x})} \quad \eta \int v_\tau^\top(\mathbf{x})\nabla_{\mathbf{x}}\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}q(\mathbf{x})\mathrm{d}\mathbf{x} + \frac{\eta}{2}\int q(\mathbf{x})\|v_\tau(\mathbf{x})\|_2^2 \mathrm{d}\mathbf{x} \\
\Rightarrow &\inf_{v_\tau(\mathbf{x})} \quad \eta \int v_\tau^\top(\mathbf{x})\nabla_{\mathbf{x}}\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}q(\mathbf{x})\mathrm{d}\mathbf{x} + \frac{\eta}{2}\int q(\mathbf{x})\|v_\tau(\mathbf{x})\|_2^2 \mathrm{d}\mathbf{x} \\
&\qquad\qquad + \underbrace{\frac{\eta}{2}\int q(\mathbf{x})\|\nabla_{\mathbf{x}}\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}\|_2^2 \mathrm{d}\mathbf{x}}_{\geq 0} \\
\Rightarrow &\inf_{v_\tau(\mathbf{x})} \quad \frac{\eta}{2}\mathbb{E}_{q(\mathbf{x})}[\|v_\tau(\mathbf{x}) + \nabla_{\mathbf{x}}\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}\|_2^2] \\
\Rightarrow &v_\tau^\star(\mathbf{x}) = -\nabla_{\mathbf{x}}\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\},
\end{aligned} \tag{36}$$

where "(i)" follows from the inequality in Equation (34), and "(ii)" follows from integration by parts under mild assumptions (Johnson & Zhang, 2021; Dong et al., 2022):

$$\begin{aligned}
&\int v_\tau^\top(\mathbf{x})\nabla_{\mathbf{x}}\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}q(\mathbf{x})\mathrm{d}\mathbf{x} + \int \nabla_{\mathbf{x}} \cdot [q(\mathbf{x})v_\tau(\mathbf{x})]\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}\mathrm{d}\mathbf{x} \\
=&\int \nabla_{\mathbf{x}} \cdot \{\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}q(\mathbf{x})v_\tau(\mathbf{x})\}\mathrm{d}\mathbf{x} \\
=&0.
\end{aligned} \tag{37}$$

Consider two empirical measures, $q(\mathbf{x})$ and $q'(\mathbf{x})$, and the velocity field $v_\tau^\star(\mathbf{x})$ that transports $q(\mathbf{x})$ to $q'(\mathbf{x})$ according to the following ODE, which can also be treated as a weak solution to Equation (25):

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\tau} = v_\tau^\star(\mathbf{x}), \tag{38}$$

we can view the velocity field as the solution to the proximal problem in Equation (31), with the Wasserstein distance serving as the "proximal term."

Following Song et al. (2020b), DMs can be broadly categorized based on their underlying stochastic dynamics, specifically the VP and VE SDEs. Despite their differences in noise scheduling, both formulations admit a unified SDE framework, which we leverage to analyze the energy functional when applying them to TSDI. Our analyses of the VP and VE SDEs are given below:

- **VP-SDE:** The DM-based imputation process that infers $\mathbf{x}^{\text{imp}}$ from $\mathbf{x}^{\text{obs}}$ is described by the following VP-SDE[1]:

$$\mathrm{d}\mathbf{x}^{\text{imp}} = [\frac{1}{2}\beta(\tau)\mathbf{x}^{\text{imp}} + \beta(\tau)\nabla_{\mathbf{x}^{\text{imp}}} \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})]\mathrm{d}\tau + \sqrt{\beta(\tau)}\mathrm{d}W_\tau. \tag{39}$$

[1]We reverse the time axis; thus, the coefficient of $\beta(\tau)$ is 1 rather than -1, as given by (Song et al., 2020b)

Based on Equation (16), the corresponding Fokker–Planck equation governing $q_\tau(\mathbf{x}^{\text{imp}})$ is given by:

$$\begin{aligned}&\frac{\partial q_\tau(\mathbf{x}^{\text{imp}})}{\partial \tau}\\ =&-\nabla_{\mathbf{x}^{\text{imp}}}\cdot[(\frac{1}{2}\beta(\tau)\mathbf{x}^{\text{imp}}+\beta(\tau)\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}}))q_\tau(\mathbf{x}^{\text{imp}})]+\frac{1}{2}\nabla_{\mathbf{x}^{\text{imp}}}\cdot[\beta(\tau)\nabla_{\mathbf{x}^{\text{imp}}}q_\tau(\mathbf{x}^{\text{imp}})],\end{aligned} \tag{40}$$

which is equivalent to the following continuity equation:

$$\frac{\partial q_\tau(\mathbf{x}^{\text{imp}})}{\partial \tau}=-\nabla_{\mathbf{x}^{\text{imp}}}\cdot[(\frac{1}{2}\beta(\tau)\mathbf{x}^{\text{imp}}+\beta(\tau)\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})-\frac{1}{2}\nabla_{\mathbf{x}^{\text{imp}}}\log q_\tau(\mathbf{x}^{\text{imp}}))q_\tau(\mathbf{x}^{\text{imp}})]. \tag{41}$$

The corresponding velocity field $v_\tau^\star(\mathbf{x}^{\text{imp}})$ is given by:

$$v_\tau^\star(\mathbf{x}^{\text{imp}})=\frac{1}{2}\beta(\tau)\mathbf{x}^{\text{imp}}+\beta(\tau)\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})-\frac{1}{2}\beta(\tau)\nabla_{\mathbf{x}^{\text{imp}}}\log q_\tau(\mathbf{x}^{\text{imp}}). \tag{42}$$

Consequently, the corresponding optimization problem is given by:

$$\begin{aligned}\inf_{q'(\mathbf{x}^{\text{imp}})}\ -\mathbb{E}_{q'(\mathbf{x}^{\text{imp}})}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})]+&\overbrace{\frac{1}{2}\mathbb{E}_{q'(\mathbf{x}^{\text{imp}})}[\log q'(\mathbf{x}^{\text{imp}})]-\frac{1}{4}\mathbb{E}_{q'(\mathbf{x}^{\text{imp}})}[\|\mathbf{x}^{\text{imp}}\|_2^2]}^{\phi(\mathbf{x}^{\text{imp}})}\\ &+\frac{1}{2\beta(\tau)}\mathbb{W}_2^2(q'(\mathbf{x}^{\text{imp}}),q(\mathbf{x}^{\text{imp}})),\end{aligned} \tag{43}$$

- **VE-SDE:** The DM-based imputation process that infers $\mathbf{x}^{\text{imp}}$ from $\mathbf{x}^{\text{obs}}$ is described by the following VE-SDE[2]:

$$\mathrm{d}\mathbf{x}^{\text{imp}}=\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\mathrm{d}\tau+\sqrt{\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}}\mathrm{d}W_\tau. \tag{44}$$

Based on Equation (16), the corresponding Fokker–Planck equation governing $q_\tau(\mathbf{x}^{\text{imp}})$ is given by:

$$\frac{\partial q_\tau(\mathbf{x}^{\text{imp}})}{\partial \tau}=-\nabla_{\mathbf{x}^{\text{imp}}}\cdot[(\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}}))q_\tau(\mathbf{x}^{\text{imp}})]+\frac{1}{2}\nabla_{\mathbf{x}^{\text{imp}}}\cdot[\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}\nabla_{\mathbf{x}^{\text{imp}}}q_\tau(\mathbf{x}^{\text{imp}})], \tag{45}$$

which is equivalent to the following continuity equation:

$$\frac{\partial q_\tau(\mathbf{x}^{\text{imp}})}{\partial \tau}=-\nabla_{\mathbf{x}^{\text{imp}}}\cdot[(\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})-\frac{1}{2}\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}\nabla_{\mathbf{x}^{\text{imp}}}\log q_\tau(\mathbf{x}^{\text{imp}}))q_\tau(\nabla_{\mathbf{x}^{\text{imp}}})]. \tag{46}$$

The corresponding velocity field $v_\tau^\star(\mathbf{x}^{\text{imp}})$ is given by:

$$v_\tau^\star(\mathbf{x}^{\text{imp}})=\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})-\frac{1}{2}\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}\nabla_{\mathbf{x}^{\text{imp}}}\log q_\tau(\mathbf{x}^{\text{imp}}). \tag{47}$$

Consequently, the corresponding optimization problem is given by:

$$\inf_{q'(\mathbf{x}^{\text{imp}})}\ -\mathbb{E}_{q'(\mathbf{x}^{\text{imp}})}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})]+\overbrace{\frac{1}{2}\mathbb{E}_{q'(\mathbf{x}^{\text{imp}})}[\log q'(\mathbf{x}^{\text{imp}})]}^{\phi(\mathbf{x}^{\text{imp}})}+\frac{1}{\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}}\mathbb{W}_2^2(q'(\mathbf{x}^{\text{imp}}),q(\mathbf{x}^{\text{imp}})). \tag{48}$$

By observing Equations (43) and (48), we arrive at the desired results. □

C.1.2 Discussions of Proposition 3.1

Extension to Flow-Matching-Based Approaches. Recent methods, such as MIRI (Yu et al., 2025), formulate time-series imputation through flow matching and parameterize a velocity field (Lipman

[2]We reverse the time axis; thus, $\frac{\mathrm{d}\sigma^2(\tau)}{\mathrm{d}\tau}$ is 1 rather than -1, as given by (Song et al., 2020b)

et al., 2023; Liu et al., 2023b) rather than a score function with parameter θ. We show that such a flow admits an equivalent score-based stochastic representation, allowing the proximal recursion interpretation in Proposition 3.1 to be applied under the same regularity assumptions. Let $\mathbf{x}^{\text{imp}}$ denote the target imputed component, let $\mathbf{x}^{\text{obs}}$ be the observed condition, and let $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Consider the linear interpolation

$$\mathbf{x}_\tau = \tau \mathbf{x}^{\text{imp}} + (1-\tau)\epsilon, \qquad \tau \in [0,1]. \tag{49}$$

The conditional velocity field is learned using the following objective:

$$\min_\theta \ \mathbb{E}_{\tau, \mathbf{x}^{\text{imp}}, \mathbf{x}^{\text{obs}}, \epsilon}[\frac{1}{2} \left\| v_\theta(\mathbf{x}_\tau, \tau, \mathbf{x}^{\text{obs}}) - (\mathbf{x}^{\text{imp}} - \epsilon) \right\|_2^2], \qquad \tau \sim \mathcal{U}[0,1]. \tag{50}$$

Here $\mathcal{U}[0,1]$ denotes the uniform distribution on $[0,1]$. At the population optimum, the learned field satisfies the following equation:

$$v^\star(\mathbf{x}, \tau, \mathbf{x}^{\text{obs}}) = \mathbb{E}[\mathbf{x}^{\text{imp}} - \epsilon \mid \mathbf{x}_\tau = \mathbf{x}, \mathbf{x}^{\text{obs}}]. \tag{51}$$

Starting from $\mathbf{x}_0 = \epsilon$, the corresponding probability-flow ODE is

$$\frac{\mathrm{d}\mathbf{x}_\tau}{\mathrm{d}\tau} = v^\star(\mathbf{x}_\tau, \tau, \mathbf{x}^{\text{obs}}), \tag{52}$$

and its terminal state $\mathbf{x}_1$ provides the imputed component, while the observed entries remain fixed.

Let $q_\tau(\mathbf{x} \mid \mathbf{x}^{\text{obs}})$ denote the conditional marginal density of $\mathbf{x}_\tau$. The Gaussian perturbation identity gives

$$\nabla_{\mathbf{x}} \log q_\tau(\mathbf{x} \mid \mathbf{x}^{\text{obs}}) = -\frac{\mathbb{E}[\epsilon \mid \mathbf{x}_\tau = \mathbf{x}, \mathbf{x}^{\text{obs}}]}{1-\tau}. \tag{53}$$

Since $\mathbf{x}_\tau = \epsilon + \tau(\mathbf{x}^{\text{imp}} - \epsilon)$, Equations (51) and (53) yield

$$\nabla_{\mathbf{x}} \log q_\tau(\mathbf{x} \mid \mathbf{x}^{\text{obs}}) = -\frac{\mathbf{x} - \tau v^\star(\mathbf{x}, \tau, \mathbf{x}^{\text{obs}})}{1-\tau}. \tag{54}$$

For any time-dependent diffusion rate $\sigma(\mathbf{x}_\tau, \tau)$, the SDE

$$\mathrm{d}\mathbf{x}_\tau = [v^\star(\mathbf{x}_\tau, \tau, \mathbf{x}^{\text{obs}}) + \frac{\sigma(\mathbf{x}_\tau, \tau)}{2} \nabla_{\mathbf{x}} \log q_\tau(\mathbf{x}_\tau \mid \mathbf{x}^{\text{obs}})]\mathrm{d}\tau + \sqrt{\sigma(\mathbf{x}_\tau, \tau)}\, \mathrm{d}W_\tau \tag{55}$$

has the same conditional marginal densities as the ODE in Equation (52). To recover the variance-preserving-type construction used in our original derivation, set

$$\sigma(\mathbf{x}_\tau, \tau) = \frac{2(1-\tau)}{\tau}. \tag{56}$$

Substituting Equations (54) and (56) into Equation (55) gives

$$v^\star(\mathbf{x}_\tau, \tau, \mathbf{x}^{\text{obs}}) + \frac{1-\tau}{\tau} \nabla_{\mathbf{x}} \log q_\tau(\mathbf{x}_\tau \mid \mathbf{x}^{\text{obs}}) = 2v^\star(\mathbf{x}_\tau, \tau, \mathbf{x}^{\text{obs}}) - \frac{\mathbf{x}_\tau}{\tau}. \tag{57}$$

Consequently, the flow admits the following VP-type SDE representation:

$$\mathrm{d}\mathbf{x}_\tau = [2v^\star(\mathbf{x}_\tau, \tau, \mathbf{x}^{\text{obs}}) - \frac{\mathbf{x}_\tau}{\tau}]\mathrm{d}\tau + \sqrt{\frac{2(1-\tau)}{\tau}}\, \mathrm{d}W_\tau. \tag{58}$$

This construction preserves the conditional marginals of the flow ODE and exhibits the linear restoring drift and time-dependent diffusion characteristic of a VP-type SDE (see Equation (39)). It therefore provides the connection required to extend the proximal recursion analysis to DM-based TSDI approaches trained by flow matching paradigm.

Theoretical Analysis. Notably, in Section C.1.1, we conclude that, in the Wasserstein space, simulating the velocity field $v_\tau^\star(\mathbf{x}) = -\nabla_{\mathbf{x}}\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}$ corresponds to the descent direction for the objective functional given in Equation (31). For convenience, we define the objective functional in Equation (31) as $\widehat{\mathcal{F}}[q'(\mathbf{x})]$:

$$\widehat{\mathcal{F}}[q'(\mathbf{x})] \coloneqq \mathcal{F}[q'(\mathbf{x})] + \frac{1}{2\eta} \mathbb{W}_2^2(q', q). \tag{59}$$

We next show that this direction decreases the original proximal objective for a sufficiently small step. At first, let

$$\varphi_q(\mathbf{x}) \coloneqq \delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})], \qquad v^\star(\mathbf{x}) = -\nabla_{\mathbf{x}}\varphi_q(\mathbf{x}). \tag{60}$$

On this basis, we define transportation map $\mathfrak{T}_\eta : \mathbb{R}^D \to \mathbb{R}^D$ as:

$$\mathfrak{T}_\eta(\mathbf{x}) \coloneqq \mathbf{x} - \eta\nabla_{\mathbf{x}}\varphi_q(\mathbf{x}), \qquad q_\eta \coloneqq (\mathfrak{T}_\eta)_{\#}q. \tag{61}$$

Suppose the q is absolutely continuous, φ_q is twice continuously differentiable, and the first variation of $\mathcal{F}$ admits a second-order remainder along $\mathfrak{T}_\eta$. Further when $\nabla^2\varphi_q$ is bounded above, so that, for sufficiently small $\eta > 0$, we have:

$$\mathbf{I} - \eta\nabla^2_{\mathbf{x}}\varphi_q(\mathbf{x}) \succeq 0 \qquad q\text{-a.e.} \tag{62}$$

Since

$$\mathfrak{T}_\eta(\mathbf{x}) = \nabla_{\mathbf{x}}[\frac{1}{2}\|\mathbf{x}\|_2^2 - \eta\varphi_q(\mathbf{x})], \tag{63}$$

condition (62) makes $\mathfrak{T}_\eta$ the gradient of a convex function. Hence, $\mathfrak{T}_\eta$ is the optimal quadratic-cost transport map from q to q_η (Ambrosio et al., 2005), and therefore we arrive at the following result:

$$\mathbb{W}_2^2(q_\eta, q) = \eta^2\mathbb{E}_{q(\mathbf{x})}[\|\nabla_{\mathbf{x}}\varphi_q(\mathbf{x})\|_2^2]. \tag{64}$$

The first-variation expansion along the same transport direction gives

$$\mathcal{F}[q_\eta] = \mathcal{F}[q] - \eta\mathbb{E}_{q(\mathbf{x})}[\|\nabla_{\mathbf{x}}\varphi_q(\mathbf{x})\|_2^2] + o(\eta^2). \tag{65}$$

Combining Equations (64) and (65), we have:

$$\widehat{\mathcal{F}}[q_\eta] = \mathcal{F}[q_\eta] + \frac{1}{2\eta}\mathbb{W}_2^2(q_\eta, q) = \mathcal{F}[q] - \frac{\eta}{2}\mathbb{E}_{q(\mathbf{x})}\|\nabla_{\mathbf{x}}\varphi_q(\mathbf{x})\|_2^2 + o(\eta^2). \tag{66}$$

Since $\widehat{\mathcal{F}}[q] = \mathcal{F}[q]$, Eq. (66) implies that, whenever $\mathbb{E}_{q(\mathbf{x})}[\|\nabla_{\mathbf{x}}\varphi_q\|_2^2] > 0$, the update generated by $v^\star$ rigorously decreases the original proximal objective for all sufficiently small η. At a stationary point, the first-order dissipation vanishes. Thus, under the stated regularity and step-size conditions, the velocity obtained from the upper-bound derivation is a locally valid descent direction for the original proximal objective.

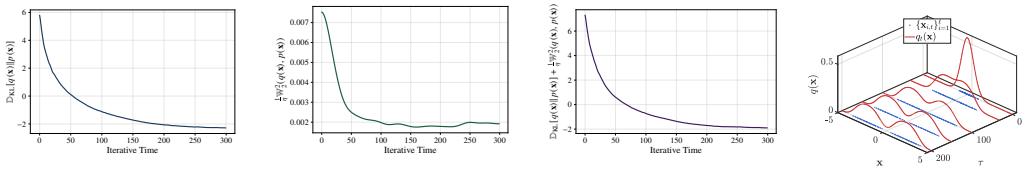
Toy Case Study. To further support our theoretical results, we conduct experiments on the following toy problem:

$$\inf_{q(\mathbf{x})} \quad \mathbb{D}_{\mathrm{KL}}[q(\mathbf{x})\|p(\mathbf{x})] + \frac{1}{\eta}\mathbb{W}_2^2(q(\mathbf{x}), p(\mathbf{x})), \tag{67}$$

where we set $p(\mathbf{x}) \propto \frac{1}{2}\mathcal{N}(\mathbf{x}| -2, 0.25) + \frac{1}{2}\mathcal{N}(\mathbf{x}|2, 0.25)$ and $\eta = 0.005$. We represent $q(\mathbf{x})$ using a finite set of particles $\{\mathbf{x}_i\}_{i=1}^N$, as described in Equation (17). We then optimize $\{\mathbf{x}_i\}_{i=1}^N$ using $v_\tau^\star(\mathbf{x}) = -\nabla_{\mathbf{x}}\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\}$. For the KL-divergence functional, $v_\tau^\star(\mathbf{x})$ is given by

$$v_\tau^\star(\mathbf{x}) = -\nabla_{\mathbf{x}}\{\delta_{q(\mathbf{x})}\mathcal{F}[q(\mathbf{x})]\} = \nabla_{\mathbf{x}}\log p(\mathbf{x}) - \nabla_{\mathbf{x}}\log q(\mathbf{x}).$$

The corresponding toy case study results are presented in Figure 5. As shown in Figures 5(a) to 5(c), both the Wasserstein distance and the KL divergence gradually decrease as τ evolves. In addition, as illustrated in Figure 5(d), the density progressively approaches the target distribution when τ increases. Consequently, the total functional $\widehat{\mathcal{F}}[q(\mathbf{x})]$ also decreases, thereby empirically supporting our theoretical justification.



(a) $\mathbb{D}_{\mathrm{KL}}[q(\mathbf{x})\|p(\mathbf{x})]$ vs. τ. (b) $\frac{1}{\eta}\mathbb{W}_2^2(q(\mathbf{x}), p(\mathbf{x}))$ vs. τ. (c) $\widehat{\mathcal{F}}[q(\mathbf{x})]$ vs. τ. (d) $q_\tau(\mathbf{x})$ w.r.t. τ.

Figure 5: Toy case study results for the optimizing problem Equation (67).

C.2 Derivation of Lemma 3.2

Lemma (3.2). Let μ and ν be probability measures on $\mathbb{R}^D$, and consider a contaminated target distribution $\tilde{\nu} = (1-\zeta)\nu + \zeta\delta_{\mathbf{z}}, \zeta \in (0,1)$, where $\delta_{\mathbf{z}}$ denotes a Dirac mass at the outlier location $\mathbf{z} \in \mathbb{R}^D$. The Wasserstein distance has the following lower bound:

$$\mathbb{W}_2^2(\mu,\tilde{\nu}) \geq \zeta\mathbb{W}_2^2(\mu,\nu) + (1-\zeta)[\|\mathbf{y}^* - \mathbf{z}\|_2^2 - g(\mathbf{y}^*) + \int h(\mathbf{z})\mu(\mathbf{z})\mathrm{d}\mathbf{z}], \tag{68}$$

for some $\mathbf{y}^*$ belonging to the support of ν, and where f and g are optimal dual potentials for $\mathbb{W}_2^2(\mu,\nu)$. Meanwhile, the SPT discrepancy with Bregman potential $\psi(\nu) \coloneqq \int \nu(\mathbf{y})[\log\nu(\mathbf{y}) - 1]\mathrm{d}\mathbf{y}$ under target contamination admits the bound as follows:

$$\mathbb{S}(\mu,\tilde{\nu}) \leq (1-\zeta)\mathbb{S}(\mu,\nu) + \zeta(1 - e^{-D(\mathbf{z})}) + C(\zeta), \tag{69}$$

where $D(\mathbf{z}) \coloneqq \int \|\mathbf{z}-\mathbf{x}\|_2^2\mu(\mathbf{x})\mathrm{d}\mathbf{x}$ is the average squared distance between $\mathbf{z}$ and samples from ν, and $C(\zeta)$ is a constant defined as $C(\zeta) \coloneqq (1-\zeta)\log\frac{1}{1-\zeta} - \zeta\log\zeta$.

Proof. Our proof is based on previous works (Fatras et al., 2021; Wang et al., 2025c), and is divided into two parts, namely the derivation of the lower bound given by Equation (68) and the derivation of the upper bound given by Equation (69).

Derivation of Equation (68): Let $C(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2$. The Kantorovich dual for $\mathbb{W}_2^2(\mu,\nu)$ reads

$$\mathbb{W}_2^2(\mu,\nu) = \sup_{h,g:\ h(\mathbf{x})+g(\mathbf{y})\leq C(\mathbf{x},\mathbf{y})} \{\int h(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) + \int g(\mathbf{y})\mathrm{d}\nu(\mathbf{y})\}, \tag{70}$$

where h and g are called optimal dual potentials. Let (h,g) be an optimal dual pair for (μ,ν). Consider the following contaminated target:

$$\tilde{\nu} = (1-\zeta)\nu + \zeta\delta_{\mathbf{z}}, \qquad \zeta \in (0,1), \tag{71}$$

with $\mathbf{z} \in \mathbb{R}^D$. Define a new potential $\tilde{g}$ by keeping g on the support $\mathrm{supp}(\nu)$ and extending it at $\mathbf{z}$ via the c-transform:

$$\tilde{g}(\mathbf{y}) = g(\mathbf{y}) \text{ for } \mathbf{y} \in \mathrm{supp}(\nu), \qquad \tilde{g}(z) = h^c(\mathbf{z}) \coloneqq \inf_{\mathbf{y}\in\mathbb{R}^D}\big(c(\mathbf{z},\mathbf{y}) - g(\mathbf{y})\big). \tag{72}$$

By construction of the c-transform, we have $h(\mathbf{x}) + \tilde{g}(\mathbf{z}) \leq c(\mathbf{x},\mathbf{z})$ for all $\mathbf{x}$, and on $\mathrm{supp}(\nu)$ we have $h(\mathbf{x}) + \tilde{g}(\mathbf{y}) = h(\mathbf{x}) + g(\mathbf{y}) \leq C(\mathbf{x},\mathbf{y})$. Hence $(h,\tilde{g})$ is feasible for the dual problem associated with $(\mu,\tilde{\nu})$. Therefore, we have:

$$\begin{aligned}
\mathbb{W}_2^2(\mu,\tilde{\nu}) &= \sup_{h',g':h'(\mathbf{x})+g'(\mathbf{y})\leq C(\mathbf{x},\mathbf{y})} \{\int h'(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) + \int g'(\mathbf{y})\mathrm{d}\tilde{\nu}(\mathbf{y})\} \\
&\geq \int h(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) + \int \tilde{g}(\mathbf{y})\mathrm{d}\tilde{\nu}(\mathbf{y}) \\
&= \int h(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) + (1-\zeta)\int g(\mathbf{y})\mathrm{d}\nu(\mathbf{y}) + \zeta\tilde{g}(\mathbf{z}) \\
&= (1-\zeta)[\int h(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) + \int g(\mathbf{y})\mathrm{d}\nu(\mathbf{y})] + \zeta[\tilde{g}(\mathbf{z}) + \int h(\mathbf{x})\mathrm{d}\mu(\mathbf{x})] \\
&= (1-\zeta)\mathbb{W}_2^2(\mu,\nu) + \zeta[h^c(\mathbf{z}) + \int h(\mathbf{x})\mathrm{d}\mu(\mathbf{x})],
\end{aligned} \tag{73}$$

where the last equality uses optimality of (h,g) for (μ,ν), i.e. $\int h\mathrm{d}\mu + \int g\mathrm{d}\nu = \mathbb{W}_2^2(\mu,\nu)$.

Finally, the infimum in $h^c(\mathbf{z})$ is achieved at some $\mathbf{y}^*$ in the support of ν, yielding

$$h^c(\mathbf{z}) = \inf_{\mathbf{y}} \|\mathbf{z}-\mathbf{y}\|_2^2 - g(\mathbf{y})) = \|\mathbf{z}-\mathbf{y}^*\|_2^2 - g(\mathbf{y}^*). \tag{74}$$

Plugging Equation (74) into Equation (73), we get the following result:

$$\mathbb{W}_2^2(\mu,\tilde{\nu}) \ \geq\ (1-\zeta)\mathbb{W}_2^2(\mu,\nu) \ +\ \zeta[\|\mathbf{z}-\mathbf{y}^*\|_2^2 - g(\mathbf{y}^*) + \int h(\mathbf{x})d\mu(\mathbf{x})]. \tag{75}$$

This proves the claimed lower bound.

Derivation of Equation (69): When we set the Bregman potential as $\psi(\nu) \coloneqq \int \nu(\mathbf{y})[\log \nu(\mathbf{y}) - 1]\mathrm{d}\mathbf{y}$, the SPT discrepancy can be reformulated as follows:

$$\begin{aligned}
&\mathbb{S}(\mu,\nu)\\
&= \inf_{\pi\in\Pi(\mu)} \int \|\mathbf{x}-\mathbf{y}\|^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y}) + \psi(\pi_{\mathbf{y}}) - \psi(\nu) - \langle \delta_\nu \psi(\nu), \pi_{\mathbf{y}} - \nu\rangle\\
&= \inf_{\pi\in\Pi(\mu)} \int \|\mathbf{x}-\mathbf{y}\|^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y}) + \int \pi_{\mathbf{y}}(\mathbf{y})[\log \pi_{\mathbf{y}}(\mathbf{y}) - 1]\mathrm{d}\mathbf{y} - \int \nu(\mathbf{y})[\log \nu(\mathbf{y}) - 1]\mathrm{d}\mathbf{y}\\
&\qquad\qquad - \int \log \nu(\mathbf{y})[\pi_{\mathbf{y}}(\mathbf{y}) - \nu(\mathbf{y})]\mathrm{d}\mathbf{y}\\
&= \inf_{\pi\in\Pi(\mu)} \int \|\mathbf{x}-\mathbf{y}\|^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y}) + \int \pi_{\mathbf{y}}(\mathbf{y}) \log \frac{\pi_{\mathbf{y}}(\mathbf{y})}{\nu(\mathbf{y})}\mathrm{d}\mathbf{y}.
\end{aligned} \tag{76}$$

On this basis, we consider a scalable contamination model by replacing $\tilde{\nu} = (1-\zeta)\nu + \zeta\delta_z$ with $\tilde{\nu}_\varphi = (1-\zeta)\nu + \zeta\varphi\delta_z$, where $\varphi \geq 0$ is an optimizable parameter. Let $\pi^\star \in \Pi(\mu)$ be an optimal plan for $\mathbb{S}(\mu,\nu)$, i.e.,

$$\mathbb{S}(\mu,\nu) = \int \|\mathbf{x}-\mathbf{y}\|^2 d\pi^\star(\mathbf{x},\mathbf{y}) + D_\psi(\pi^\star_y, \nu), \tag{77}$$

so that $\widehat{\pi}_{\mathbf{y}} = (1-\zeta)\pi^\star_{\mathbf{y}} + \zeta\delta_{\mathbf{z}}$. Let us fix any $\varphi \geq 0$ and set $\tilde{\nu}_\varphi \coloneqq (1-\zeta)\nu + \zeta\varphi\delta_{\mathbf{z}}$.

By linearity of the cost, we have the following result:

$$\int \|\mathbf{x}-\mathbf{y}\|_2^2 \mathrm{d}\widehat{\pi}(\mathbf{x},\mathbf{y}) = (1-\zeta)\int \|\mathbf{x}-\mathbf{y}\|^2 \mathrm{d}\pi^\star(\mathbf{x},\mathbf{y}) + \zeta \int \|\mathbf{x}-\mathbf{z}\|^2 \mu(\mathrm{d}\mathbf{x}). \tag{78}$$

Using the convexity of $p \mapsto D_\psi(p,q)$ in its first argument for fixed q, we have

$$D_\psi(\widehat{\pi}_{\mathbf{y}}, \tilde{\nu}_\varphi) \leq (1-\zeta) D_\psi(\pi^\star_{\mathbf{y}}, \tilde{\nu}_\varphi) + \zeta D_\psi(\delta_{\mathbf{z}}, \tilde{\nu}_\varphi). \tag{79}$$

Moreover, since $\mathbf{z}$ is an outlier, it is reasonable to assume that $\nu(\{\mathbf{z}\}) \approx 0$. Hence, we have:

$$D_\psi(\delta_{\mathbf{z}}, \tilde{\nu}_\varphi) = -\log(\zeta\varphi). \tag{80}$$

Since $\tilde{\nu}_\varphi \geq (1-\zeta)\nu$ as measures, we obtain the following result when $\pi^\star_{\mathbf{y}} \ll \nu$ (to ensure that the Bregman divergence is well defined):

$$D_\psi(\pi^\star_{\mathbf{y}}, \tilde{\nu}_\varphi) = \int \pi^\star_{\mathbf{y}}(\mathbf{y}) \log \frac{\pi^\star_{\mathbf{y}}(\mathbf{y})}{\tilde{\nu}_\varphi(\mathbf{y})} \leq \int \pi^\star_{\mathbf{y}} \log \frac{\pi^\star_{\mathbf{y}}(\mathbf{y})}{(1-\zeta)\nu(\mathbf{y})} = D_\psi(\pi^\star_{\mathbf{y}}, \nu) + \log \frac{1}{1-\zeta}. \tag{81}$$

Collecting terms and using the feasible coupling $\widehat{\pi}$, we obtain the following inequality for any $\varphi \geq 0$:

$$\mathbb{S}(\mu,\tilde{\nu}) \leq (1-\zeta)\mathbb{S}(\mu,\nu) + \zeta[\varphi D(\mathbf{z}) + (\varphi\log\varphi - \varphi + 1)] + C(\zeta), \tag{82}$$

where $D(\mathbf{z}) \coloneqq \int \|\mathbf{x}-\mathbf{z}\|^2 \mu(\mathrm{d}\mathbf{x})$, and $C(\zeta) \coloneqq (1-\zeta)\log\frac{1}{1-\zeta} - \zeta\log\zeta$ is a constant (for fixed ζ) independent of φ.

Finally, since Equation (82) holds for any $\varphi \geq 0$, we minimize the φ-dependent term. Define

$$e(\varphi) \coloneqq \varphi D(\mathbf{z}) + (\varphi\log\varphi - \varphi + 1), \tag{83}$$

Then take the first-order condition for the optimization problem, we have:

$$\frac{\mathrm{d}e(\varphi)}{\mathrm{d}\varphi} = D(\mathbf{z}) + \log\varphi = 0 \Rightarrow \varphi^\star = \exp(-D(\mathbf{z})). \tag{84}$$

On this basis, consider the second-order condition, we have:

$$\frac{\mathrm{d}^2 e(\varphi)}{\mathrm{d}\varphi^2} = \frac{1}{\varphi} > 0. \tag{85}$$

Thus, we have the following result:

$$\inf_{\varphi\geq 0} \quad e(\varphi) = D(z)e^{-D(z)} - d(z)e^{-D(z)} - e^{-D(z)} + 1 = 1 - e^{-D(z)}. \tag{86}$$

Plugging Equation (86) into Equation (82) yields the desired bound. □

C.3 Derivation of Proposition 3.3

Proposition (3.3). Assume $p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})$ is once continuously differentiable and $\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})$ is square-integrable under the measures considered. Let $\psi(\rho) = \int \rho(\mathbf{x})[\log \rho(\mathbf{x}) - 1]\mathrm{d}\mathbf{x}$. Represent q' by an empirical measure $q' = \sum_{i=1}^N w_i \delta_{\mathbf{x}_i}$ with $w_i \geq 0$ and $\sum_{i=1}^N w_i = 1$. Under the alternating splitting scheme detailed in Section C.3, the following coordinate-wise directions provide first-order updates for the resulting location and weight surrogate subproblems, respectively:

- **Location direction** (score-direction-guided), where $\{\mathbf{x}_i\}_{i=1}^N$ are updated by:

$$\boldsymbol{T}_{\mathbf{x}^{\text{imp}}}(\mathbf{x}^{\text{imp}}) = \nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}}). \tag{87}$$

- **Weight direction** (score-norm-guided), where logarithmic weights $\{\log w_i\}_{i=1}^N$ are updated by:

$$\boldsymbol{T}_w(\mathbf{x}^{\text{imp}}) = -2\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2 + 2\mathbb{E}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2]. \tag{88}$$

Proof. We first decompose the SPT discrepancy $\mathbb{S}$. Specifically, for $\mathbb{S}(\mu, \nu)$, when introducing an intermediate marginal distribution $\sigma := \pi_{\mathbf{y}}$ we can disintegrate the feasible set according to the $\mathbf{y}$-marginal:

$$\begin{aligned}
\mathbb{S}(\mu,\nu) &= \inf_{\pi\in\Pi(\mu)} \int \|\mathbf{x}-\mathbf{y}\|^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y}) + D_\psi(\pi_{\mathbf{y}},\nu) \\
\Rightarrow \mathbb{S}(\mu,\nu) &= \inf_{\sigma} \inf_{\pi:\ \pi_{\mathbf{x}}=\mu,\ \pi_{\mathbf{y}}=\sigma} \int \|\mathbf{x}-\mathbf{y}\|^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y}) + D_\psi(\pi_{\mathbf{y}},\nu) \\
\Rightarrow \mathbb{S}(\mu,\nu) &= \inf_{\sigma} \inf_{\pi:\ \pi_{\mathbf{x}}=\mu,\ \pi_{\mathbf{y}}=\sigma} \int \|\mathbf{x}-\mathbf{y}\|^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y}) + D_\psi(\sigma,\nu) \\
\Rightarrow \mathbb{S}(\mu,\nu) &= \inf_{\sigma} \left\{ \underbrace{\inf_{\pi\in\Pi(\mu,\sigma)} \int \|\mathbf{x}-\mathbf{y}\|^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y})}_{=\mathbb{W}_2^2(\mu,\sigma)} + D_\psi(\sigma,\nu) \right\} \\
\Rightarrow \mathbb{S}(\mu,\nu) &= \inf_{\sigma} \mathbb{W}_2^2(\mu,\sigma) + D_\psi(\sigma,\nu),
\end{aligned} \tag{89}$$

On this basis, when introducing the intermediate $\widehat{q}(\mathbf{x}^{\text{imp}})$, we have the following result:

$$\begin{aligned}
&\inf_{q'} \ -\mathbb{E}_{q'}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})] + \frac{1}{2}\mathbb{E}_{q'}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2] + \frac{1}{2\eta}\big\{\inf_{\widehat{q}} \mathbb{W}_2^2(q',\widehat{q}) + D_\psi(\widehat{q},q')\big\} \\
\Rightarrow &\inf_{q',\widehat{q}} \ -\mathbb{E}_{q'}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})] + \frac{1}{2}\mathbb{E}_{q'}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2] + \frac{1}{2\eta}[\mathbb{W}_2^2(q',\widehat{q}) + D_\psi(\widehat{q},q')].
\end{aligned} \tag{90}$$

To decouple location and weight updates, inspired by the Cole-Hopf transformation (Caluya & Halder, 2020; Chen et al., 2022; Caluya & Halder, 2021) and alternating direction method of multipliers-based optimization methods (Lin et al., 2022), we introduce two split measures for q' as:

$$\begin{cases} \tilde{q} = \sum_{i=1}^N w_i^{(k)} \delta_{\mathbf{x}_i^{\text{imp}}} \\ \widehat{q} = \sum_{i=1}^N \widehat{w}_i \delta_{\mathbf{x}_i^{\text{imp}}}^{(k+1)} \end{cases}, \tag{91}$$

where $\tilde{q}$ updates *locations* $\{\mathbf{x}_i^{\text{imp}}\}_{i=1}^N$ with fixed weights $\{w_i^{(k)}\}_{i=1}^N$, and $\widehat{q}$ updates *weights* $\{\widehat{w}_i\}$ on the fixed support $\{\delta_{\mathbf{x}_i^{\text{imp}}}^{(k+1)}\}_{i=1}^N$. Accordingly, we introduce the following definition:

$$\begin{cases} \tilde{q} \in \tilde{\mathcal{Q}} := \{\sum_{i=1}^N w_i^{(k)} \delta_{\mathbf{x}_i^{\text{imp}}}\} \\ \widehat{q} \in \widehat{\mathcal{Q}} := \{\sum_{i=1}^N w_i \delta_{\mathbf{x}_i^{\text{imp}}}^{(k+1)} : w_i \geq 0, \sum_{i=1}^N w_i = 1\} \end{cases}. \tag{92}$$

We then adopt an alternating splitting scheme. Given $q'^{(k)}$, we obtain the following surrogate objective functional:

$$\inf_{\tilde{q}\in\tilde{\mathcal{Q}},\widehat{q}\in\widehat{\mathcal{Q}}} \ -\mathbb{E}_{\tilde{q}}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})] + \frac{1}{2}\mathbb{E}_{\widehat{q}}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2] + \frac{1}{2\eta}[\mathbb{W}_2^2(\tilde{q},q) + D_\psi(\widehat{q},q')] \tag{93}$$

The two directions in Proposition 3.3 are obtained by separately optimizing the split measures in this surrogate: the transportation step optimizes $\tilde{q}$ with the weights fixed, whereas the teleportation step optimizes $\widehat{q}$ with the locations fixed. Thus, the result is explicitly conditional on this alternating splitting scheme and does not assert that the two directions are joint descent directions for the unsplit objective. Relating the split iterates to the unsplit objective requires the splitting discrepancy between $\tilde{q}$ and $\widehat{q}$ to vanish in the continuous-time limit.

On this basis, we can perform the following optimization process recursively:

- **Transportation Step:** In this step, we change the spatial location of samples $\{\mathbf{x}_i^{\text{imp}}\}_{i=1}^N$ by solving the following optimization problem:

$$\tilde{q}^{(k+1)} \in \arg\min_{\tilde{q}\in\tilde{\mathcal{Q}}} \left\{ -\mathbb{E}_{\tilde{q}}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})] + \frac{1}{2\eta}\mathbb{W}_2^2(\tilde{q}, q'^{(k)}) \right\}. \tag{94}$$

We have the following upper bound for Equation (31) based on Equations (34) and (35):

$$\begin{aligned}
&\inf_{\tilde{q}} \quad -\mathbb{E}_{\tilde{q}}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})] + \mathbb{E}_{q'^{(k)}}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})] + \frac{1}{2\eta}\mathbb{W}_2^2(\tilde{q}, q'^{(k)}) \\
\overset{\text{(i)}}{\Rightarrow} &\inf_{v_\tau(\mathbf{x}^{\text{imp}})} \quad \cancel{-\mathbb{E}_{q'^{(k)}}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})]} + \eta\int \nabla_{\mathbf{x}^{\text{imp}}}\cdot[q'^{(k)}(\mathbf{x}^{\text{imp}})v_\tau(\mathbf{x}^{\text{imp}})][\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})]\mathrm{d}\mathbf{x}^{\text{imp}} \\
&\qquad\qquad + \cancel{\mathbb{E}_{q'^{(k)}}[\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})]} + \frac{\eta}{2}\int q'^{(k)}(\mathbf{x}^{\text{imp}})\|v_\tau(\mathbf{x}^{\text{imp}})\|_2^2\mathrm{d}\mathbf{x}^{\text{imp}} \\
\overset{\text{(ii)}}{\Rightarrow} &\inf_{v_\tau(\mathbf{x}^{\text{imp}})} \quad -\eta\int v_\tau^\top(\mathbf{x}^{\text{imp}})\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})q'^{(k)}(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{x}^{\text{imp}} + \frac{\eta}{2}\int q'^{(k)}(\mathbf{x}^{\text{imp}})\|v_\tau(\mathbf{x}^{\text{imp}})\|_2^2\mathrm{d}\mathbf{x}^{\text{imp}} \\
\Rightarrow &\inf_{v_\tau(\mathbf{x}^{\text{imp}})} \quad -\eta\int v_\tau^\top(\mathbf{x}^{\text{imp}})\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})q'^{(k)}(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{x}^{\text{imp}} + \frac{\eta}{2}\int q'^{(k)}(\mathbf{x}^{\text{imp}})\|v_\tau(\mathbf{x}^{\text{imp}})\|_2^2\mathrm{d}\mathbf{x}^{\text{imp}} \\
&\qquad\qquad + \underbrace{\frac{\eta}{2}\int q'^{(k)}(\mathbf{x}^{\text{imp}})\|\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2\mathrm{d}\mathbf{x}}_{\geq 0} \\
\Rightarrow &\inf_{v_\tau(\mathbf{x}^{\text{imp}})} \quad \frac{\eta}{2}\mathbb{E}_{q'^{(k)}(\mathbf{x}^{\text{imp}})}[\|v_\tau(\mathbf{x}^{\text{imp}}) - \nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2] \\
\Rightarrow &v_\tau^\star(\mathbf{x}^{\text{imp}}) = \nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}}),
\end{aligned} \tag{95}$$

where "(i)" is based on the inequality given by Equation (34), and "(ii)" is based on the integration-by-parts under the mild assumptions (Johnson & Zhang, 2021; Dong et al., 2022):

$$\begin{aligned}
&\int v_\tau^\top(\mathbf{x}^{\text{imp}})\nabla_{\mathbf{x}^{\text{imp}}}\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})q'^{(k)}(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{x}^{\text{imp}} \\
&\qquad + \int \nabla_{\mathbf{x}^{\text{imp}}}\cdot[q'^{(k)}(\mathbf{x}^{\text{imp}})v_\tau(\mathbf{x}^{\text{imp}})]\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\mathrm{d}\mathbf{x}^{\text{imp}} \\
&= \int \nabla_{\mathbf{x}^{\text{imp}}}\cdot\{\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})q'^{(k)}(\mathbf{x}^{\text{imp}})v_\tau(\mathbf{x}^{\text{imp}})\}\mathrm{d}\mathbf{x}^{\text{imp}} = 0.
\end{aligned} \tag{96}$$

- **Teleportation Step:** In this step, we adjust the weights $\{w_i\}_{i=1}^N$ by solving the following optimization problem:

$$\widehat{q}^{(k+1)} \in \arg\min_{\widehat{q}\in\widehat{\mathcal{Q}}} \left\{ \frac{1}{2}\mathbb{E}_{\widehat{q}}[\|\nabla\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2] + \frac{1}{2\eta}D_\psi(\widehat{q}, \tilde{q}^{(k+1)}) \right\}, \tag{97}$$

When $\psi(\rho) = \int \rho(\mathbf{x})[\log\rho(\mathbf{x}) - 1]\mathrm{d}\mathbf{x}$, the optimization problem can be reformulated as follows:

$$\widehat{q}^{(k+1)} \in \arg\min_{\widehat{q}\in\widehat{\mathcal{Q}}} \left\{ \frac{1}{2}\mathbb{E}_{\widehat{q}}[\|\nabla\log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2] + \frac{1}{2\eta}\int \widehat{q}(\mathbf{x}^{\text{imp}})\log\frac{\widehat{q}(\mathbf{x}^{\text{imp}})}{\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})}\mathrm{d}\mathbf{x}^{\text{imp}} \right\}. \tag{98}$$

Consider the reaction PDE given by Equation (20), we have the following result:

$$\begin{cases}
\widehat{q}(\mathbf{x}^{\text{imp}}) = \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) - \eta\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})g_\tau(\mathbf{x}^{\text{imp}}) + O(\eta) \\
\log\widehat{q}(\mathbf{x}^{\text{imp}}) = \log\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) - \eta g_\tau(\mathbf{x}^{\text{imp}}) + \dfrac{\eta^2}{2}\dfrac{\partial^2\log\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})}{\partial\tau} + O(\eta^2)
\end{cases}. \tag{99}$$

Thus, for the Bregman divergence, we have the following result:

$$
\begin{aligned}
& D_{\psi}(\widehat{q}, \tilde{q}^{(k+1)}) \\
=& \int \widehat{q}(\mathbf{x}^{\text{imp}}) \log \frac{\widehat{q}(\mathbf{x}^{\text{imp}})}{\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})} \mathrm{d}\mathbf{x}^{\text{imp}} \\
=& \int \widehat{q}(\mathbf{x}^{\text{imp}})[-\eta g_{\tau}(\mathbf{x}^{\text{imp}}) + \frac{\eta^2}{2} \frac{\partial^2 \log \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})}{\partial \tau}] \mathrm{d}\mathbf{x}^{\text{imp}} + O(\eta^2) \\
\overset{\text{(i)}}{=}& -\eta \underbrace{\int \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) g_{\tau}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}}}_{=0} + \eta^2 \underbrace{\int \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) g_{\tau}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}}}_{=0} \\
& + \frac{\eta^2}{2} \int \frac{\partial^2 \log \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})}{\partial \tau} \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} + O(\eta^2) \\
\overset{\text{(ii)}}{=}& \int g_t^2(\mathbf{x}^{\text{imp}}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} + O(\eta^2),
\end{aligned}
\tag{100}
$$

where "(i)" is based on the following derivation:

$$
\begin{aligned}
& \int \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} = 1 \\
\Rightarrow & \frac{\mathrm{d}}{\mathrm{d}\tau} \int \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} = 0 \\
\Rightarrow & \int \frac{\partial \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})}{\partial \tau} \mathrm{d}\mathbf{x}^{\text{imp}} = \int [\int g_{\tau}(\mathbf{x}^{\text{imp}\prime}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}\prime}) \mathrm{d}\mathbf{x}^{\text{imp}\prime} - g_{\tau}(\mathbf{x}^{\text{imp}})] \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} = 0 \\
\Rightarrow & -\int g_{\tau}(\mathbf{x}^{\text{imp}}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} = 0,
\end{aligned}
\tag{101}
$$

and "(ii)" is based on the following derivation: Specifically, we have the following result based on the PDE defined by Equation (20):

$$
\begin{aligned}
& \frac{\partial \log \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})}{\partial \tau} = \frac{1}{\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})} \frac{\partial \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})}{\partial \tau} = -g_{\tau}(\mathbf{x}^{\text{imp}}) \\
\Rightarrow & \frac{\partial^2 \log \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})}{\partial \tau^2} = -\frac{\partial g_{\tau}(\mathbf{x}^{\text{imp}})}{\partial \tau}.
\end{aligned}
\tag{102}
$$

Differentiate the identity $\int g_{\tau}(\mathbf{x}^{\text{imp}}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} = 0$ in time, we have the following result:

$$
\begin{aligned}
& \frac{\partial}{\partial \tau} \int g_{\tau}(\mathbf{x}^{\text{imp}}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} = 0 \\
\Rightarrow & \int (\frac{\partial g_{\tau}(\mathbf{x}^{\text{imp}})}{\partial \tau}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} + \int g_{\tau} (\frac{\partial \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})}{\partial \tau}) \mathrm{d}\mathbf{x}^{\text{imp}} = 0 \\
\Rightarrow & \int (\frac{\partial g_{\tau}(\mathbf{x}^{\text{imp}})}{\partial \tau}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} - \int g_{\tau}^2(\mathbf{x}^{\text{imp}}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} = 0 \\
\Rightarrow & \int (\frac{\partial g_{\tau}(\mathbf{x}^{\text{imp}})}{\partial \tau}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} = \int g_{\tau}^2(\mathbf{x}^{\text{imp}}) \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}}.
\end{aligned}
\tag{103}
$$

On this basis, we have:

$$\begin{aligned}
&\inf_{\widehat{q}} \frac{1}{2}\mathbb{E}_{\widehat{q}}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2] + \frac{1}{2\eta}\int \widehat{q}(\mathbf{x}^{\text{imp}}) \log \frac{\widehat{q}(\mathbf{x}^{\text{imp}})}{\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})} \mathrm{d}\mathbf{x}^{\text{imp}} \\
\Rightarrow &\inf_{\widehat{q}} \cancel{\frac{1}{2}\mathbb{E}_{\tilde{q}^{(k+1)}}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2]} - \cancel{\frac{1}{2}\mathbb{E}_{\tilde{q}^{(k+1)}}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2]} \\
&\qquad + \frac{1}{\eta}\int \|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2 \tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{x}^{\text{imp}} + \frac{1}{4\eta}\int g_\tau^2(\mathbf{x}^{\text{imp}})\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{x}^{\text{imp}} \\
\Rightarrow &\inf_{\widehat{q}} \int \{\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2 - \mathbb{E}_{\tilde{q}^{(k+1)}}[\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})]\}\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{x}^{\text{imp}} \\
&\qquad + \frac{1}{4}\int g_\tau^2(\mathbf{x}^{\text{imp}})\tilde{q}^{(k+1)}(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{x}^{\text{imp}} \\
\Rightarrow &\inf_{\widehat{q}} \mathbb{E}_{\tilde{q}^{(k+1)}}\{[g_\tau(\mathbf{x}^{\text{imp}}) + 2\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2 - 2\mathbb{E}_{\tilde{q}^{(k+1)}}[\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})]]^2\}.
\end{aligned} \tag{104}$$

Thus, the optimal $g_\tau^\star(\mathbf{x}^{\text{imp}})$ can be given as follows:

$$g_\tau^\star(\mathbf{x}^{\text{imp}}) = -2\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2 + 2\mathbb{E}_{\tilde{q}^{(k+1)}}[\|\nabla \log p(\mathbf{x}^{\text{imp}}|\mathbf{x}^{\text{obs}})\|_2^2]. \tag{105}$$

Based on Equations (21) to (23), (95) and (105), we arrive at the desired results. □

C.4 Derivation of Proposition 3.4

Proposition (3.4). Let $\eta > 0$ and $\boldsymbol{T}_w \in \mathbb{R}^N \to \mathbb{R}$. Define the intermediate log-weights as $\log \widehat{w}_i^{(k+1)} \coloneqq \log w_i^{(k)} + \eta \boldsymbol{T}_w(\mathbf{x}^{\text{imp}})$. The normalized weights are given by: $w_i^{(k+1)} = \frac{\widehat{w}_i^{(k+1)}}{\sum_{j=1}^N \widehat{w}_j^{(k+1)}}$.

Proof. By denoting $\log \widehat{w} = [\log \widehat{w}_1, \ldots, \log \widehat{w}_N]^\top$, we can reformulate $\log \widehat{w}_i^{(k+1)} \coloneqq \log w_i^{(k)} + \eta \boldsymbol{T}_w(\mathbf{x}^{\text{imp}})$ as follows:

$$\begin{aligned}
&\log \widehat{w}_i^{(k+1)} \coloneqq \log w_i^{(k)} + \eta \boldsymbol{T}_w(\mathbf{x}^{\text{imp}}) \\
\Rightarrow &\log \widehat{w}^{(k+1)} = \arg\max\nolimits_{\log \widehat{w}} [\boldsymbol{T}_w(\mathbf{x}^{\text{imp}})]^\top [\log \widehat{w}] - \frac{1}{2\eta}\|\log \widehat{w}_i - \log w_i\|_2^2.
\end{aligned} \tag{106}$$

We observe that the main feasibility issue in the iteration result is the introduction of the Euclidean distance as the proximal term. Thus, the key for addressing this issue is replacing the proximal term with entropy-induced Bregman divergence. On this basis, we introduce the Bregman divergence for $\widehat{w}$ and w as follows:

$$\begin{aligned}
&D_\psi(\widehat{w}, w^{(k)}) = \psi(\widehat{w}) - \psi(w^{(k)}) - [\nabla\psi(w^{(k)})]^\top [\widehat{w} - w^{(k)}] \\
\overset{\psi(w):=w^\top[\log w - 1]}{\Longrightarrow} &D_\psi(\widehat{w}, w^{(k)}) = \sum_{i=1}^N \widehat{w}_i \log \frac{\widehat{w}_i}{w_i^{(k)}} - \sum_{i=1}^N \widehat{w}_i + \sum_{i=1}^N w_i^{(k)}.
\end{aligned} \tag{107}$$

The optimization problem is therefore given by:

$$\begin{aligned}
&\arg\max_{w} [\boldsymbol{T}_w(\mathbf{x}^{\text{imp}})]^\top [w] - \frac{1}{\eta}\{\sum_{i=1}^N w_i \log \frac{w_i}{w_i^{(k)}} - \sum_{i=1}^N w_i + \sum_{i=1}^N w_i^{(k)}\} \\
&\text{s.t.} \begin{cases} \sum_{j=1}^N w_j = 1 \\ w_j \geq 0, \forall j \in \{1, \ldots, N\} \end{cases}
\end{aligned}. \tag{108}$$

We introduce the Lagrange multipliers $\lambda \in \mathbb{R}$ and $\gamma \in \mathbb{R}^N$ for the constraints $\sum_{j=1}^N \widehat{w}_j = 1$ and $\widehat{w}_j \geq 0, \forall j \in \{1, \ldots, N\}$. The Lagrangian is given by:

$$\mathcal{L}(w, \lambda, \gamma) = [\boldsymbol{T}_w(\mathbf{x}^{\text{imp}})]^\top w - \frac{1}{\eta}\{\sum_{i=1}^N w_i \log \frac{w_i}{w_i^{(k)}} - \sum_{i=1}^N w_i + \sum_{i=1}^N w_i^{(k)}\} + \lambda(\sum_{i=1}^N w_i - 1) + \sum_{i=1}^N \gamma_i w_i. \tag{109}$$

For each $i \in \{1, \ldots, N\}$, using

$$\frac{\partial}{\partial w_i}(w_i \log \frac{w_i}{w_i^{(k)}}) = \log \frac{w_i}{w_i^{(k)}} + 1, \qquad \frac{\partial}{\partial w_i}(-\sum_j w_j) = -1, \tag{110}$$

we get the following result:

$$\frac{\partial \mathcal{L}}{\partial w_i} = \boldsymbol{T}_{w_i}(\mathbf{x}^{\text{imp}}) - \frac{1}{\eta}(\log \frac{w_i}{w_i^{(k)}} + 1 - 1) + \lambda + \gamma_i = \boldsymbol{T}_{w_i}(\mathbf{x}^{\text{imp}}) - \frac{1}{\eta} \log \frac{w_i}{w_i^{(k)}} + \lambda + \gamma_i. \tag{111}$$

Setting Equation (111) to 0, we have the following result:

$$0 = \boldsymbol{T}_{w_i}(\mathbf{x}^{\text{imp}}) - \frac{1}{\eta} \log \frac{w_i}{w_i^{(k)}} + \lambda + \gamma_i \quad \Longrightarrow \quad \log \frac{w_i}{w_i^{(k)}} = \eta(\boldsymbol{T}_{w_i}(\mathbf{x}^{\text{imp}}) + \lambda + \gamma_i). \tag{112}$$

The Karush–Kuhn–Tucker conditions (Boyd & Vandenberghe, 2004) include the following complementary-slackness conditions:

$$w_i \geq 0, \quad \gamma_i \geq 0, \quad \gamma_i w_i = 0, \quad \sum_i w_i = 1. \tag{113}$$

Since w_i can be given as follows based on Equation (112):

$$w_i = w_i^{(k)} \exp[\eta(\boldsymbol{T}_{w_i}(\mathbf{x}^{\text{imp}}) + \lambda + \gamma_i)], \tag{114}$$

we know that $\gamma_i = 0$ for all $i \in \{1, \ldots, N\}$. Then, we have:

$$w_i = w_i^{(k)} \exp[\eta(\boldsymbol{T}_{w_i}(\mathbf{x}^{\text{imp}}) + \lambda)]. \tag{115}$$

On this basis, using the fact that:

$$\widehat{w}_i^{(k+1)} \coloneqq w_i^{(k)} \exp(\eta \boldsymbol{T}_{w_i}(\mathbf{x}^{\text{imp}})), \tag{116}$$

we can further reformulate Equation (115) as follows:

$$w_i^{(k+1)} = \exp(\eta\lambda)\widehat{w}_i^{(k+1)}. \tag{117}$$

Imposing $\sum_{i=1}^N w_i^{(k+1)} = 1$, the coefficient $\exp(\eta\lambda)$ can be reformulated as follows:

$$\exp(\eta\lambda) = \frac{1}{\sum_{j=1}^N \widehat{w}_j^{(k+1)}}. \tag{118}$$

Finally, based on Equations (116) to (118), we arrive at the desired result. □

C.5 Derivation of Proposition 3.5 and Related Discussions

Proposition. Let $\mathscr{X} = \mathbf{x}^{\text{imp}}$ and $\mathscr{O} = \mathbf{x}^{\text{obs}}$, and define $\mathscr{S}(\mathscr{X}, \mathscr{O}) \coloneqq \nabla_{\mathscr{X}} \log p(\mathscr{X}|\mathscr{O})$, $m(\mathscr{X}) \coloneqq \nabla_{\mathscr{X}} \log p(\mathscr{X})$. Assume that differentiation and marginalization can be interchanged and that the scores are square-integrable. Then $m(\mathscr{X}) = \mathbb{E}[\mathscr{S}(\mathscr{X}, \mathscr{O}) \mid \mathscr{X}]$, we obtain:

$$\mathbb{E}_{\mathscr{X},\mathscr{O}}[\|s_\theta(\mathscr{X}) - \mathscr{S}(\mathscr{X}, \mathscr{O})\|_2^2] = \mathbb{E}_{\mathscr{X}}[\|s_\theta(\mathscr{X}) - m(\mathscr{X})\|_2^2] + \mathbb{E}_{\mathscr{X},\mathscr{O}}[\|\mathscr{S}(\mathscr{X}, \mathscr{O}) - m(\mathscr{X})\|_2^2]. \tag{119}$$

Hence, the two risks have the same minimizers over predictors depending only on $\mathscr{X}$.

Proof. Let $p_{\mathscr{X}}$ and $p_{\mathscr{O}}$ denote the marginal densities of $\mathscr{X}$ and $\mathscr{O}$, respectively. Marginalization gives the following result:

$$p_{\mathscr{X}}(\mathbf{x}^{\text{imp}}) = \int p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}}) p_{\mathscr{O}}(\mathbf{x}^{\text{obs}}) \, \mathrm{d}\mathbf{x}^{\text{obs}}. \tag{120}$$

By the assumed interchange of differentiation and integration, we have:

$$\begin{aligned} m(\mathbf{x}^{\text{imp}}) &= \nabla_{\mathbf{x}^{\text{imp}}} \log p_{\mathscr{X}}(\mathbf{x}^{\text{imp}}) \\ &= \frac{1}{p_{\mathscr{X}}(\mathbf{x}^{\text{imp}})} \int p_{\mathscr{O}}(\mathbf{x}^{\text{obs}}) \nabla_{\mathbf{x}^{\text{imp}}} p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}}) \, \mathrm{d}\mathbf{x}^{\text{obs}} \\ &= \int \frac{p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}}) p_{\mathscr{O}}(\mathbf{x}^{\text{obs}})}{p_{\mathscr{X}}(\mathbf{x}^{\text{imp}})} \nabla_{\mathbf{x}^{\text{imp}}} \log p(\mathbf{x}^{\text{imp}} \mid \mathbf{x}^{\text{obs}}) \, \mathrm{d}\mathbf{x}^{\text{obs}} \\ &= \int p(\mathbf{x}^{\text{obs}} \mid \mathbf{x}^{\text{imp}}) \mathscr{S}(\mathbf{x}^{\text{imp}}, \mathbf{x}^{\text{obs}}) \, \mathrm{d}\mathbf{x}^{\text{obs}} = \mathbb{E}[\mathscr{S}(\mathscr{X}, \mathscr{O}) \mid \mathscr{X} = \mathbf{x}^{\text{imp}}]. \end{aligned} \tag{121}$$

Thus, $m(\mathscr{X})$ is the conditional expectation of $\mathscr{S}(\mathscr{X}, \mathscr{O})$ given $\mathscr{X}$. Equivalently, in the Hilbert space L^2, $m(\mathscr{X})$ is the orthogonal projection of $\mathscr{S}(\mathscr{X}, \mathscr{O})$ onto the closed subspace of square-integrable functions measurable with respect to $\sigma(\mathscr{X})$.

To make this orthogonality explicit, we first define $r(\mathscr{X}, \mathscr{O})$ as follows:

$$r(\mathscr{X}, \mathscr{O}) \coloneqq \mathscr{S}(\mathscr{X}, \mathscr{O}) - m(\mathscr{X}). \tag{122}$$

Notably, Equation (121) yields the following result:

$$\mathbb{E}[r(\mathscr{X}, \mathscr{O}) \mid \mathscr{X}] = 0. \tag{123}$$

For any square-integrable predictor $s_\theta(\mathscr{X})$, the tower property therefore gives

$$\mathbb{E}[(s_\theta(\mathscr{X}) - m(\mathscr{X}))^\top r(\mathscr{X}, \mathscr{O})] = \mathbb{E}_{\mathscr{X}}[(s_\theta(\mathscr{X}) - m(\mathscr{X}))^\top \mathbb{E}[r(\mathscr{X}, \mathscr{O}) \mid \mathscr{X}]] = 0. \tag{124}$$

Finally, since the following equation holds, we get:

$$s_\theta(\mathscr{X}) - \mathscr{S}(\mathscr{X}, \mathscr{O}) = s_\theta(\mathscr{X}) - m(\mathscr{X}) - r(\mathscr{X}, \mathscr{O}). \tag{125}$$

As such, expanding the squared norm, taking expectations, and applying (124), we have:

$$\mathbb{E}_{\mathscr{X},\mathscr{O}}\|s_\theta(\mathscr{X}) - \mathscr{S}(\mathscr{X}, \mathscr{O})\|_2^2 = \mathbb{E}_{\mathscr{X}}\|s_\theta(\mathscr{X}) - m(\mathscr{X})\|_2^2 + \mathbb{E}_{\mathscr{X},\mathscr{O}}\|\mathscr{S}(\mathscr{X}, \mathscr{O}) - m(\mathscr{X})\|_2^2, \tag{126}$$

which is Equation (13). The final term is independent of θ and equals the irreducible conditional variation $\mathbb{E}[\mathrm{tr}\{\mathrm{Cov}(\mathscr{S}(\mathscr{X}, \mathscr{O}) \mid \mathscr{X})\}]$. Hence, both population risks have the same set of minimizers over predictors depending only on $\mathscr{X}$; this conclusion concerns equality of minimizers, not pointwise equality of the conditional and marginal scores. □

C.6 Discussions of the Convergence Property for SPIRIT Framework

Before discussing convergence, we present the pseudocode corresponding to Figure 3 in Algorithm 1.

Algorithm 1 The pseudocode for SPIRIT.

Input: $\mathbf{X}^{\text{obs}}$: observed data.
Parameters: η: proximal recursion step size, lr: the learning rate for the score network, σ: the coefficient for DSM, θ: the parameters of the score network, $\mathcal{E}_{\text{score}}$: the number of score-network training iterations, and $\mathcal{E}_{\text{imp}}$: the number of imputation iterations.
Output: $\mathbf{X}^{\text{imp}}$: imputed data.

Initialize the missing entries in $\mathbf{x}^{\text{imp}}$ and the score-network parameters θ; set $w_i^{(0)} = 1/N$ for $i = 1, \ldots, N$.
for $e_{\text{score}} = 1$ to $\mathcal{E}_{\text{score}}$ **do**
  $s_\theta \leftarrow$ Equation (14); ▷ Score Learning
  for $e_{\text{imp}} = 1$ to $\mathcal{E}_{\text{imp}}$ **do**
    $\boldsymbol{T}_{\mathbf{x}^{\text{imp}}}(\mathbf{x}^{\text{imp}}) \leftarrow$ Equation (11); ▷ Recursive Imputation
    $\boldsymbol{T}_w(\mathbf{x}^{\text{imp}}) \leftarrow$ Equation (12);
    $\log \widehat{w} \leftarrow \texttt{ApplyGrad}(\boldsymbol{T}_w(\mathbf{x}^{\text{imp}}), \eta, \log w)$;
    $w_i \leftarrow \dfrac{\widehat{w}_i}{\sum_{j=1}^N \widehat{w}_j}$
    $\mathbf{x}^{\text{imp}} \leftarrow \texttt{ApplyGrad}(w\boldsymbol{T}_{\mathbf{x}^{\text{imp}}}(\mathbf{x}^{\text{imp}}), \eta, \mathbf{x}^{\text{imp}})$;
    $\mathbf{X}^{\text{imp}} \leftarrow \mathbf{X}^{\text{obs}} \odot (1 - \mathbf{M}) + \mathbf{X}^{\text{imp}} \odot \mathbf{M}$;
  end for
end for

Interpretation of the SPIRIT implementation. Algorithm 1 alternates between "score learning" and "recursive imputation". During recursive imputation, the learned score determines the transport direction of each imputed particle, while the squared score norm provides a soft reweighting signal. Specifically, particles with larger score norms receive smaller normalized weights and consequently make smaller effective transport updates. The normalized log-weight update is an entropy-induced Bregman proximal step on the probability simplex, preserving nonnegative weights that sum to one. Thus, the method continuously modulates particle contributions rather than using hard acceptance, rejection, or rollback. Finally, the masking operation restores the observed entries after every recursive update, ensuring that optimization is restricted to the missing entries. Here, `ApplyGrad` denotes the

optimizer update used in our implementation, with inputs specifying the update direction, nominal step size, and variable to be updated.

Convergence Analysis of SPIRIT. The remainder of this subsection analyzes the convergence behavior of the two stages illustrated in Figure 3. For the "Score Learning" stage, convergence to a stationary point follows from standard optimization results under mild regularity conditions (Bottou et al., 2018). For the "Recursive Imputation" stage, we first establish sufficient-decrease properties for the two frozen-coordinate surrogate blocks and then analyze the idealized continuous-time limit of the split transport–reaction dynamics. Define the energy functional

$$\mathcal{J}[q'] \coloneqq -\mathbb{E}_{q'}[\log p(\mathbf{x}^{\mathrm{imp}} \mid \mathbf{x}^{\mathrm{obs}})] + \frac{1}{2}\mathbb{E}_{q'}[\big\|\nabla \log p(\mathbf{x}^{\mathrm{imp}} \mid \mathbf{x}^{\mathrm{obs}})\big\|_2^2]. \tag{127}$$

The continuous-time analysis relies on the following assumptions:

(A1) The energy $\mathcal{J}$ is lower bounded and admits the first variations required by the transport–reaction dynamics.
(A2) The boundary terms arising from integration by parts vanish.
(A3) The discrepancy between the two split measures is $O(\eta)$ as $\eta \to 0$.
(A4) The limiting transport and reaction directions are generated by the same energy $\mathcal{J}$.

Assumption (A1) holds, for example, when $-\log p(\mathbf{x}^{\mathrm{imp}} \mid \mathbf{x}^{\mathrm{obs}})$ is bounded below on the support of q' and both terms in Equation (127) are integrable. Assumption (A2) is satisfied under compact support or sufficiently fast decay of the density and flux at infinity. Assumption (A3) is a first-order consistency condition for the alternating splitting and holds when the two subflows are sufficiently regular and stable. Assumption (A4) links the split dynamics to the energy-dissipation identity derived below. As such, under Assumptions (A1)–(A4), the value of $\mathcal{J}$ is non-increasing along the limiting flow. With the assumptions for recursive imputation established, we now analyze the two stages in the workflow order shown in Figure 3 and Algorithm 1: First "Score Learning" and then "Recursive Imputation".

Convergence Analysis of "Score Learning". At first, we define the $\mathcal{L}^{\mathrm{DSM}}$ as follows:

$$\mathcal{L}^{\mathrm{DSM}}(\theta) \coloneqq \mathbb{E}_{q_\sigma(\widehat{\mathbf{x}}^{\mathrm{imp}}|\mathbf{x}^{\mathrm{imp}})}[\|s_\theta(\mathbf{x}^{\mathrm{imp}}) - \nabla \log q_\sigma(\widehat{\mathbf{x}}^{\mathrm{imp}}|\mathbf{x}^{\mathrm{imp}})\|_2^2] \tag{128}$$

For the following mild assumptions, we get the convergence promising of the "Score Learning" stage:

1. **Assumption 1 (Lower bound):** $\mathcal{L}^{\mathrm{DSM}}(\theta) \geq \mathcal{L}^{\mathrm{DSM}}(\theta^\star) \geq 0 > -\infty$ for all θ.
2. **Assumption 2 (L-smooth):** $\nabla\mathcal{L}^{\mathrm{DSM}}(\theta)$ is L-Lipschitz: $\|\nabla\mathcal{L}^{\mathrm{DSM}}(\theta) - \nabla\mathcal{L}^{\mathrm{DSM}}(\theta')\| \leq L\|\theta - \theta'\|$, $\forall\theta, \theta'$.

The "Lower bound" condition holds since $\mathcal{L}^{\mathrm{DSM}}(\theta)$ is an expectation of a squared ℓ_2 error and thus $\mathcal{L}^{\mathrm{DSM}}(\theta) \geq 0$. The "$L$-smoothness" condition (i.e., Lipschitz continuity of $\nabla\mathcal{L}^{\mathrm{DSM}}$) is a standard assumption in gradient-descent convergence analyses, for example (Bach, 2024). A sufficient (but not necessary) set of conditions for L-smoothness is to impose appropriate boundedness/spectral-norm constraints on the network so that $\nabla\mathcal{L}^{\mathrm{DSM}}$ becomes Lipschitz; gradient clipping is a practical heuristic that controls update magnitudes but does not imply L-smoothness.

Under Assumptions 1 and 2, for a learning rate $lr \in (0, \frac{2}{L})$, the gradient descent iterates satisfy:

$$\|\nabla\mathcal{L}^{\mathrm{DSM}}(\theta_e)\| \to 0 \quad \text{as } e \to \infty. \tag{129}$$

Because $\mathcal{L}^{\mathrm{DSM}}(\theta)$ is L-smooth, we have the following result for any θ, θ',

$$\mathcal{L}^{\mathrm{DSM}}(\theta') \leq \mathcal{L}^{\mathrm{DSM}}(\theta) + \langle\nabla\mathcal{L}^{\mathrm{DSM}}(\theta), \theta' - \theta\rangle + \frac{L}{2}\|\theta' - \theta\|^2. \tag{130}$$

Apply this with $\theta = \theta_e$ and $\theta' = \theta_{e+1} = \theta_e - lr\nabla\mathcal{L}^{\mathrm{DSM}}(\theta_e)$. Then

$$\begin{aligned}
&\mathcal{L}^{\mathrm{DSM}}(\theta_{e+1}) \\
\leq&\mathcal{L}^{\mathrm{DSM}}(\theta_e) + \left\langle\mathcal{L}^{\mathrm{DSM}}(\theta_e), -lr\nabla\mathcal{L}^{\mathrm{DSM}}(\theta_e)\right\rangle + \frac{L}{2}\| - \xi\nabla\mathcal{L}^{\mathrm{DSM}}(\theta_e)\|^2 \\
=&\mathcal{L}^{\mathrm{DSM}}(\theta_e) - lr\|\nabla\mathcal{L}^{\mathrm{DSM}}(\theta_e)\|^2 + \frac{L \times lr^2}{2}\|\mathcal{L}^{\mathrm{DSM}}(\theta_e)\|^2 \\
=&\mathcal{L}^{\mathrm{DSM}}(\theta_e) - lr(1 - \frac{L \times lr}{2})\|\nabla\mathcal{L}^{\mathrm{DSM}}(\theta_e)\|^2.
\end{aligned} \tag{131}$$

Since $lr \in (0, \frac{2}{L})$, we get the following results:

$$\mathcal{L}^{\text{DSM}}(\theta_e) - \mathcal{L}^{\text{DSM}}(\theta_{e+1}) \geq lr(1 - \frac{L \times lr}{2})\|\nabla \mathcal{L}^{\text{DSM}}(\theta_e)\|^2. \tag{132}$$

Summing from $t = 0$ to $\mathcal{E}_{\text{score}} - 1$ yields a telescoping sum:

$$\sum_{e=0}^{\mathcal{E}_{\text{score}}-1} c\|\nabla \mathcal{L}^{\text{DSM}}(\theta_e)\|^2 \leq \mathcal{L}^{\text{DSM}}(\theta_e) - \mathcal{L}^{\text{DSM}}(\theta_{\mathscr{T}}) \leq \mathcal{L}^{\text{DSM}}(\theta_e) - \mathcal{L}^{\text{DSM}}(\theta^\star). \tag{133}$$

Letting $\mathcal{E}_{\text{score}} \to \infty$,

$$\sum_{t=0}^{\infty} \|\nabla \mathcal{L}^{\text{DSM}}(\theta_e)\|^2 < \infty. \tag{134}$$

A series of nonnegative terms is finite only if the terms go to zero, hence

$$\lim_{e \to \infty} \|\nabla \mathcal{L}^{\text{DSM}}(\theta_e)\| = 0. \tag{135}$$

This proves convergence to a stationary point.

Convergence Analysis of "Recursive Imputation". We first state what can be guaranteed for the two frozen-coordinate blocks. Let $f(\mathbf{x}) = -\log p(\mathbf{x} \mid \mathbf{x}^{\text{obs}})$, let $a_i = \|\nabla \log p(\mathbf{x}_i^{\text{imp}} \mid \mathbf{x}^{\text{obs}})\|_2^2$, and condition on a fixed score network during one recursive-imputation stage. If ∇f is L-Lipschitz and the location update uses $\alpha_i^{(k)} = \eta w_i^{(k)} \in (0, \frac{2}{L})$, the descent lemma gives the following inequality:

$$\sum_{i=1}^{N} w_i^{(k)} f(\mathbf{x}_i^{(k+1)}) \leq \sum_{i=1}^{N} w_i^{(k)} f(\mathbf{x}_i^{(k)}) - \sum_{i=1}^{N} w_i^{(k)} \alpha_i^{(k)} [1 - \frac{L\alpha_i^{(k)}}{2}] \|\nabla f(\mathbf{x}_i^{(k)})\|_2^2. \tag{136}$$

For fixed particle locations, let $\psi(w) = \sum_{i=1}^{N} w_i(\log w_i - 1)$. The normalized multiplicative update in Proposition 3.4 is the exact solution of the following optimization problem:

$$w^{(k+1)} = \arg\min_{w \in \Delta^{N-1}} \left\{ [w]^\top [a] + \frac{1}{2\eta} D_\psi(w, w^{(k)}) \right\}, \tag{137}$$

and therefore satisfies

$$[w^{(k+1)}]^\top [a] + \frac{1}{2\eta} D_\psi(w^{(k+1)}, w^{(k)}) \leq [w^{(k)}]^\top [a]. \tag{138}$$

Thus, each frozen-coordinate surrogate block has a sufficient-decrease certificate. These inequalities do not by themselves establish descent of the coupled unsplit objective, because the score-norm and transport terms also vary when particle locations change.

When $\eta \to 0$, based on Section B, we are implicitly simulating the following PDE using the Forward-Euler method:

$$\frac{\partial q_\tau(\mathbf{x}^{\text{imp}})}{\partial \tau} = -\nabla \cdot (v_\tau^\star(\mathbf{x}^{\text{imp}}) q_\tau(\mathbf{x}^{\text{imp}})) + g_\tau^\star(\mathbf{x}^{\text{imp}}). \tag{139}$$

On this basis, the derivative of the functional along the flow is:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\tau} \mathcal{J}(q_\tau') \\ &= -\int \nabla[\cdot(q_\tau'(\mathbf{x}^{\text{imp}}) v_\tau^\star(\mathbf{x}^{\text{imp}}))\delta_{q_\tau'} \mathcal{J}(q_\tau')] \mathrm{d}\mathbf{x}^{\text{imp}} + \int q_\tau'(\mathbf{x}^{\text{imp}}) g_\tau(\mathbf{x}^{\text{imp}}) \delta_{q_\tau'} \mathcal{J}(q_\tau') \mathrm{d}\mathbf{X}^{\text{imp}} \\ &\overset{\text{(i)}}{=} \int [v_\tau^\star(\mathbf{x}^{\text{imp}})]^\top [\nabla \delta_{q_\tau'} \mathcal{J}(q_\tau')] q_\tau'(\mathbf{x}^{\text{imp}}) \mathrm{d}\mathbf{x}^{\text{imp}} + \int q_\tau'(\mathbf{x}^{\text{imp}}) g_\tau(\mathbf{x}^{\text{imp}}) \delta_{q_\tau'} \mathcal{J}(q_\tau') \mathrm{d}\mathbf{X}^{\text{imp}}, \end{aligned} \tag{140}$$

where "(i)" follows from integration by parts (Johnson & Zhang, 2021; Dong et al., 2022). According to the theoretical derivation, as $\eta \to 0$, we have:

$$v_\tau^\star(\mathbf{x}^{\text{imp}}) = -\nabla \delta_{q_\tau'} \mathcal{J}(q_\tau'), \tag{141a}$$

$$g_\tau^\star(\mathbf{x}^{\text{imp}}) = -\delta_{q_\tau'} \mathcal{J}(q_\tau') + \mathbb{E}_{q_\tau'}[\delta_{q_\tau'} \mathcal{J}(q_\tau')]. \tag{141b}$$

Plugging Equations (141a) and (141b) into Equation (140), we have the following result (Neklyudov et al., 2023):

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{J}(q'_\tau) = &-\int \|\nabla\delta_{q'_\tau}\mathcal{J}(q'_\tau)\|_2^2 \mathrm{d}\mathbf{x}^{\text{imp}} - \int \{\delta_{q'_\tau}\mathcal{J}(q'_\tau) - \mathbb{E}_{q'_\tau}[\delta_{q'_\tau}\mathcal{J}(q'_\tau)]\}^2 q'_\tau(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{X}^{\text{imp}} \\ &\underbrace{- \int \{\delta_{q'_\tau}\mathcal{J}(q'_\tau) - \mathbb{E}_{q'_\tau}[\delta_{q'_\tau}\mathcal{J}(q'_\tau)]\}\{\mathbb{E}_{q'_\tau}[\delta_{q'_\tau}\mathcal{J}(q'_\tau)]\} q'_\tau(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{x}^{\text{imp}}}_{=0} \\ \leq &0,\end{aligned} \tag{142}$$

Integrating both sides of the reaction PDE in Equation (20), we obtain:

$$\int \frac{\partial q'_\tau(\mathbf{x}^{\text{imp}})}{\partial \tau}\mathrm{d}\mathbf{x}^{\text{imp}} = -\int \{\delta_{q'_\tau}\mathcal{J}(q'_\tau) - \mathbb{E}_{q'_\tau}[\delta_{q'_\tau}\mathcal{J}(q'_\tau)]\} q'_\tau(\mathbf{x}^{\text{imp}})\mathrm{d}\mathbf{x} = 0, \tag{143}$$

Under Assumptions (A1)–(A4), the functional $\mathcal{J}[q']$ decreases along the PDE flow: when the density q'_τ evolves according to Equation (139), $\mathcal{J}[q'_\tau]$ is non-increasing and the dynamics converge to a stationary solution characterized by the following derivation:

$$\left\|\nabla_x \delta_{q'_\tau}\mathcal{J}[q'_\tau]\right\|_2^2 = 0 \Rightarrow \nabla_x \delta_{q'_\tau}\mathcal{J}[q'_\tau] = 0 \Rightarrow \delta_{q'_\tau}\mathcal{J}[q'_\tau] \equiv C, \tag{144}$$

where the last implication follows because a function with zero spatial gradient is a constant, and $C < \infty$ is a constant. Therefore, any limiting distribution of the idealized continuous-time transport–reaction flow is stationary.

D Detailed Experimental Protocols

D.1 Dataset Information

Our empirical evaluation is conducted on a diverse collection of widely used time-series benchmarks. Each dataset has distinct characteristics in terms of dimensionality and temporal resolution. A summary is provided in Table 4. Details of these datasets are listed below:

- **ETT** (Zhou et al., 2021): Contains seven metrics related to electricity transformers, recorded from July 2016 to July 2018. It is divided into four subsets based on sampling frequency: ETTh1 and ETTh2 (hourly), and ETTm1 and ETTm2 (every 15 minutes).
- **Exchange** (Wu et al., 2021): This dataset collects the daily exchange rates of 8 countries, including Australia, British, Canada, Switzerland, China, Japan, New Zealand and Singapore, from 1990.01 to 2016.12.
- **Illness** (Wu et al., 2021): This dataset describes the ratio of patients seen with influenza-like illness and the total number of the patients. It includes the weekly data from the Centers for Disease Control and Prevention of the United States from 2002 to 2021.
- **Traffic** (Wu et al., 2021): Documents the hourly occupancy rates of 862 sensors on San Francisco Bay Area freeways, spanning from 2015 to 2016.

D.2 Model Hyperparameters

In this manuscript, we adopt a set of widely used TSDI methods as baselines to evaluate the effectiveness of the proposed SPIRIT framework, including Crossformer (Zhang & Yan, 2023), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), Autoformer (Wu et al., 2021), ETSformer (Woo et al., 2022), FiLM (Zhou et al., 2022), DLinear (Zeng et al., 2023), GP-VAE (Fortuin et al., 2020), CSDI (Tashiro et al., 2021), Glocal (Yang et al., 2025b), Sinkhorn (Muzellec et al., 2020), TDM (Zhao et al., 2023), HELIX (Zhang et al., 2026), and PSW-I (Wang et al., 2025c). For Crossformer (Zhang & Yan, 2023), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), Autoformer (Wu et al., 2021), ETSformer (Woo et al., 2022), FiLM (Zhou et al., 2022), DLinear (Zeng et al., 2023), GP-VAE (Fortuin et al., 2020), and CSDI (Tashiro et al., 2021), we adopt the optimized hyperparameters reported in (Du et al., 2024). For Glocal, we use PatchTST as the backbone. For the matching-based baselines, the Sinkhorn batch size is 256, while those for TDM and PSW-I are 512 and 200, respectively;

Table 4: Dataset information.

Dataset	D	Patch Length	Dataset Size	Frequency	Domain
ETT-h1	7	24	17420	Hourly	Temperature for transformer oil
ETT-h2	7	24	17420	Hourly	Temperature for transformer oil
ETT-m1	7	24	69680	15-minute	Temperature for transformer oil
ETT-m2	7	24	69680	15-minute	Temperature for transformer oil
Exchange	8	24	7558	Daily	Financial
Illness	7	24	966	Weekly	Health
Traffic	862	24	236	Hourly	Transportation

Kindly Note: *D* denotes the number of variates. *Frequency* denotes the sampling interval of time points.

the corresponding step sizes are 0.001 (Sinkhorn), 0.001 (TDM), and 0.002 (PSW-I). For PSW-I, the unbalanced OT subproblem is solved via majorization–minimization with coefficient 10.0. For SPIRIT, we set the sampling step size to 0.002 and the score-network learning rate to 0.001. We parameterize the score function s_θ in SPIRIT and velocity field in MIRI using a three-layer MLP with adaptive layer normalization (Liu et al., 2025), with hidden dimension 256. All experiments are repeated with four random seeds to ensure the reliability of the reported results and are conducted on a workstation equipped with AMD EPYC 7742 CPUs and four NVIDIA A100 GPUs.

D.3 Evaluation Metrics

Following previous work (Wang et al., 2025c; Yang et al., 2025b;a), we evaluate our model's performance using mean absolute error (MAE) and mean squared error (MSE). The definitions of these two evaluation metrics are given below:

$$\text{MAE} := \frac{\sum_{i=1}^{N}\sum_{j=1}^{T}\sum_{k=1}^{D}[|\mathbf{X}^{\text{ideal}}_{i,j,k} - \mathbf{X}^{\text{imp}}_{i,j,k}| \odot (\mathbf{1}_{N\times T\times D} - \mathbf{M})_{i,j,k}]}{\sum_{i=1}^{N}\sum_{j=1}^{T}\sum_{k=1}^{D}(\mathbf{1}_{N\times T\times D} - \mathbf{M})_{i,j,k}},$$

$$\text{MSE} := \frac{\sum_{i=1}^{N}\sum_{j=1}^{T}\sum_{k=1}^{D}[\|\mathbf{X}^{\text{ideal}}_{i,j,k} - \mathbf{X}^{\text{imp}}_{i,j,k}\|_2^2 \odot (\mathbf{1}_{N\times T\times D} - \mathbf{M})_{i,j,k}]}{\sum_{i=1}^{N}\sum_{j=1}^{T}\sum_{k=1}^{D}(\mathbf{1}_{N\times T\times D} - \mathbf{M})_{i,j,k}},$$

In addition, we evaluate uncertainty quantification using the Continuous Ranked Probability Score (CRPS). Given a predictive cumulative distribution function $\mathscr{F}(\mathbf{x})$ and the ground-truth value $\mathbf{x}^{\text{ideal}}$, CRPS is defined as

$$\text{CRPS}(\mathscr{F}, \mathbf{x}^{\text{ideal}}) = \int_{-\infty}^{+\infty} [\mathscr{F}(\mathbf{x}) - \mathbb{I}\{\mathbf{x} \geq \mathbf{x}^{\text{ideal}}\}]^2 \, \mathrm{d}\mathbf{x}.$$

Given N predictive samples $\{\mathbf{x}_i\}_{i=1}^N$, CRPS admits the following empirical form:

$$\text{CRPS} = \frac{1}{N}\sum_{n=1}^{N}\left|\mathbf{x}_n - \mathbf{x}^{\text{ideal}}\right| - \frac{1}{2N^2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\left|\mathbf{x}_n - \mathbf{x}_{n'}\right|.$$

Based on this, we evaluate CRPS only at the artificially masked entries. Following the implementation of Du et al. (2026), we use a normalized finite-quantile approximation. Let S denote the number of stochastic imputations, let $\widehat{\mathscr{Q}}_{\alpha,n,t,d}$ be the empirical α-quantile of $\{\mathbf{X}^{\text{imp},(s)}_{n,t,d}\}_{s=1}^S$, and let $\mathcal{A} = \{0.05, 0.10, \ldots, 0.95\}$. With the pinball loss $\rho_\alpha(u) = u[\alpha - \mathbb{I}\{u < 0\}]$, the reported score can be given follows:

$$\text{CRPS} = \frac{2}{|\mathcal{A}|}\frac{\sum_{\alpha\in\mathcal{A}}\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{d=1}^{D}\mathbf{M}_{n,t,d}\,\rho_\alpha[\mathbf{X}^{\text{ideal}}_{n,t,d} - \widehat{\mathscr{Q}}_{\alpha,n,t,d}]}{\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{d=1}^{D}\left|\mathbf{X}^{\text{ideal}}_{n,t,d} \times \mathbf{M}_{n,t,d}\right|}. \tag{145}$$

D.4 Simulation of MCAR Scenario

Following the established protocol of Wang et al. (2025c) and our experimental setting, we simulate missing data as outlined by Jarrett et al. (2022). First, a random subset of features is selected to remain non-missing. The remaining features are masked using a logistic model with the non-missing features as predictors. The model uses randomly selected weights, and its bias is adjusted to achieve the desired missingness rate.

E Additional Experimental Results

E.1 Statistics of Time-series Data

In Section 3.1, our main motivation is that the time-series exhibits non-stationarity and outliers. Therefore, we propose SPT as a proximal regularizer to modify the vanilla diffusion model-based imputation process. However, an important question remains: *do the non-stationarity and outliers truly exist in our time-series data?* To answer this, we conduct statistical tests, including the ADF-KPSS joint stationarity test, the Jarque-Bera (JB) normality test, and a seasonal decomposition variance analysis (Seabold & Perktold, 2010). The results are reported in Table 5.

Table 5: Dataset statistical results analysis.

	ETT-h1	ETT-h2	ETT-m1	ETT-m2	Exchange	Illness	Traffic
Non-stationary columns	100%	100%	100%	100%	100%	71%	26%
Stationarity (ADF+KPSS)	T-NS	T-NS	T-NS	T-NS	NS	T-NS	S
p_{KPSS} (average)	0.01	0.01	0.01	0.01	0.01	0.0357	0.075
Trend var ratio (%)	60.6843	90.81	85.22	96.04	99.3388	61.34	15.5844
Excess kurtosis (JB)	1.3684	3.1627	1.3348	3.1535	-0.5275	5.5628	4.0856
Skewness (avg)	-0.3069	-0.2729	-0.3069	-0.2744	0.3487	1.8898	1.3747
Challenge type	NS	NS+O	NS	NS+O	NS	NS+O	O
Robustness valid.	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: "T-NS" abbreviates trend non-stationarity, "NS" denotes non-stationarity, "S" denotes stationarity, "O" denotes outliers only, and "NS+O" denotes non-stationarity with outliers.

From Table 5, we observe that non-stationarity is prevalent across multiple datasets. Specifically, ETT-h1, ETT-h2, ETT-m1, ETT-m2, and Exchange all show 100% non-stationary columns, indicating that their statistical properties change over time. For Illness and Traffic, the test results reveal a mixed pattern, where some variables are non-stationary while others are stationary, yielding "NS" and "S" categories accordingly. Moreover, outliers are also present in a dataset-dependent manner. As summarized in the "Challenge type" row, Illness primarily falls into the "NS+O" setting, whereas Traffic contains both "O" and "NS+O" cases. This indicates that abnormal extreme behaviors coexist with distributional shifts for certain variables. Overall, these findings verify that our motivating assumptions, non-stationarity and outliers, are indeed reflected in the real datasets, providing empirical support for introducing SPT into the imputation framework.

E.2 Discussion of Outlier Preprocessing Strategy

A natural alternative to modeling outliers within the imputation objective is to remove them before imputation. To evaluate this strategy, we construct a preprocessing baseline that applies an interquartile-range (IQR) filter independently to each time-series variable. At time step t, we compute the first and third quartiles, $\mathscr{Q}_{1,t}$ and $\mathscr{Q}_{3,t}$, from the trailing window containing at most 12 observations and define $\text{IQR}_t = \mathscr{Q}_{3,t} - \mathscr{Q}_{1,t}$. An observation is identified as an outlier when it falls outside $[\mathscr{Q}_{1,t} - \text{IQR}_t, \mathscr{Q}_{3,t} + \text{IQR}_t]$. Detected outliers are replaced by linear interpolation over the non-outlier observations in the same variable; values outside the valid interpolation range are filled using the nearest valid endpoint. If fewer than two valid observations remain, the corresponding variable is left unchanged. We then apply the same SPIRIT imputation pipeline to the preprocessed data, thereby isolating the effect of explicit outlier removal.

As shown in Table 6, explicit outlier preprocessing consistently degrades both MAE and MSE on all four ETT datasets. The deterioration is moderate on ETT-h1 and ETT-h2 but becomes substantially larger on ETT-m1 and ETT-m2; in particular, preprocessing increases MSE by 47.405% on ETT-m1 and 106.488% on ETT-m2. This result suggests that locally extreme observations are not necessarily

Table 6: Effect of IQR-based outlier preprocessing on MAE and MSE.

Preprocessing	ETT-h1				ETT-h2				ETT-m1				ETT-m2			
	MAE		MSE		MAE		MSE		MAE		MSE		MAE		MSE	
	Value	Δ(↑)	Value	Δ(↑)	Value	Δ(↑)	Value	Δ(↑)	Value	Δ(↑)	Value	Δ(↑)	Value	Δ(↑)	Value	Δ(↑)
✓	0.232*	10.919%	0.140*	29.348%	0.148*	12.660%	0.048*	25.528%	0.165*	39.659%	0.055*	47.405%	0.129*	57.279%	0.034*	106.488%
✗	0.209	-	0.108	-	0.131	-	0.038	-	0.118	-	0.037	-	0.082	-	0.016	-

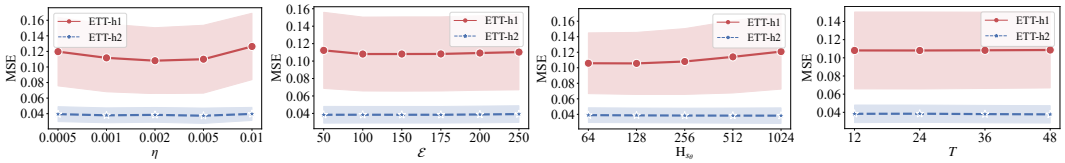
Note: Each entry is averaged over $p_{\text{miss}} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. The symbols ✓ and ✗ denote SPIRIT with and without IQR preprocessing, respectively. $\Delta(\uparrow)$ denotes the percentage degradation relative to the corresponding result without preprocessing. "*" indicates a significant difference under a paired t-test ($p < 0.05$).

spurious corruptions: they may encode legitimate regime changes, transient dynamics, or informative tail behavior. Hard detection followed by interpolation can therefore oversmooth the series and discard structures that are useful for reconstructing missing entries. In contrast, SPIRIT retains the original observations and handles unreliable mass through the soft matching mechanism of SPT. The comparison supports modeling non-stationarity and outlier sensitivity within the imputation objective rather than removing extreme values as a separate preprocessing step.

E.3 Sensitivity Analysis

In this subsection, we conduct the sensitivity analysis with respect to the step size η, number of iterations $\mathcal{E}$, hidden dimension of the score network H_{s_θ}, and patch length T. The corresponding results are presented in Figure 6.

As shown in Figure 6(a), SPIRIT's performance first improves and then degrades as the step size η increases. Since the strength of proximal regularization scales with $1/\eta$, a moderate η appropriately relaxes the proximal constraint and emphasizes the objective functional, whereas an excessively large η weakens the stabilizing effect of the proximal operator. Figure 6(b) shows that performance remains stable as the number of iterations $\mathcal{E}$ increases, suggesting fast convergence. Moreover, Figure 6(c) shows that increasing the hidden dimension H_{s_θ} degrades performance, likely because excessive model capacity causes overfitting and weakens generalization. Finally, Figure 6(d) indicates that varying the patch length has only a minor effect on performance, demonstrating SPIRIT's applicability and robustness across different patch-length settings. Overall, SPIRIT favors a moderate step size and a moderate-to-small hidden dimension to balance proximal regularization and prevent overfitting, while remaining insensitive to the number of iterations and patch length.



(a) Sensitivity results, η. (b) Sensitivity results, $\mathcal{E}$. (c) Sensitivity results, H_{s_θ}. (d) Sensitivity results, T.

Figure 6: Sensitivity analysis results. The points and shaded areas indicate the mean and $\pm$ 1.0 standard deviation from the mean, respectively.

Moreover, in the sensitivity method, we find that the shaded areas in Figure 6 for the ETT-h1 and ETT-h2 datasets are noticeably different. To understand why, we conduct trend and seasonal decomposition, the outcomes of which are summarized in Table 7. From Table 7, we observe that, compared with ETT-h2, ETT-h1 exhibits stronger seasonality (16.44% v.s. 2.82%) and is less trend-dominated (60.68% v.s. 90.81%). As a result, it is harder to fit consistently across random seeds, leading to a larger standard-deviation error than that of the ETT-h2 dataset in Figure 6.

Table 7: Variance composition statistics of ETT-h1 and ETT-h2.

Metric	**ETT-h1**	**ETT-h2**
Trend Variance Ratio (%)	60.68	90.81
Seasonal Variance Ratio (%)	16.44	2.82

E.4 Additional Results on Block Missing

Block missingness is common in real-world time series and is more challenging than pointwise missingness because consecutive observations are unavailable. We therefore evaluate SPIRIT under contiguous missing patches with lengths 3 and 4, and report the results in Tables 8 and 9.

Overall, SPIRIT remains strongly competitive across both patch lengths: it consistently delivers the lowest or near-lowest MAE and MSE across most dataset–metric pairs. When the missing patch becomes longer (from length 3 to 4), SPIRIT's performance degrades only mildly, indicating that the proposed stochastic score-based imputation and soft transport matching are robust to extended gaps.

On the ETT benchmarks, SPIRIT frequently ranks among the top methods on MAE/MSE, and is especially competitive on ETT-h2, ETT-m1, and ETT-m2, where it achieves strong MAE and/or MSE with limited sensitivity to the gap length. In comparison, several deterministic forecasting-oriented or transport-style baselines (e.g., Autoformer, Sinkhorn, and TDM) deteriorate more noticeably under contiguous missingness. On Exchange and Illness, SPIRIT maintains consistently low errors, while CSDI becomes numerically unstable (e.g., extremely large MSE values on Illness for patch length 3 and 4), leading to significant degradation. On Traffic, SPIRIT also yields low MAE and stable MSE, outperforming (or matching the best) non-stochastic baselines.

Overall, these results suggest that SPIRIT is robust across different contiguous gap lengths and data domains, supporting the effectiveness of stochastic score-based imputation and soft transport matching for reconstructing extended missing segments.

Table 8: Block-Missing Results with Missing Patch Length 3.

Dataset	ETT-h1		ETT-h2		ETT-m1		ETT-m2		Exchange		Illness		Traffic	
Metric	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Crossformer	0.240	**0.121**	0.188	0.072	0.144	0.047	0.255	0.157	0.160	0.061	0.359	0.363	0.252	0.238
TimesNet	0.277	0.152	0.224	0.100	0.192	0.075	0.182	0.067	0.239	0.126	0.321	0.293	0.301	0.287
PatchTST	0.271	0.149	0.207	0.086	0.144	0.049	0.126	0.033	0.158	0.054	0.432	0.440	0.300	0.310
Autoformer	0.684	0.933	0.691	1.050	0.723	1.185	0.768	1.474	0.627	0.769	0.661	0.935	0.487	0.597
ETSformer	0.329	0.248	0.219	0.102	0.242	0.140	0.147	0.046	0.155	0.046	0.293	0.237	0.299	**0.228**
FiLM	0.448	0.451	0.262	0.136	0.249	0.131	0.205	0.085	0.176	0.060	0.313	0.299	0.771	1.150
DLinear	0.303	0.203	0.229	0.113	0.249	0.141	0.199	0.084	0.162	0.051	0.216	0.153	0.279	0.229
Glocal	0.281	0.166	0.202	0.082	0.151	0.053	0.128	0.033	0.155	0.049	0.301	0.243	0.369	0.428
HELIX	**0.226**	0.124	0.147	0.053	0.139	0.053	0.102	0.028	0.042	0.004	0.183	0.132	0.373	0.400
GP-VAE	0.330	0.211	0.255	0.123	0.261	0.136	0.216	0.091	0.294	0.165	0.426	0.421	0.621	0.817
CSDI	0.240	0.151	0.213	0.714	**0.126**	**0.041**	**0.072**	0.025	0.049	0.007	3.368	71.677	8.081	228
MIRI	0.488	0.485	0.329	0.326	0.378	0.345	0.259	0.257	0.275	0.178	0.247	0.190	0.321	0.298
Sinkhorn	0.748	0.979	0.748	1.001	0.753	0.998	0.752	1.011	0.824	0.993	0.718	1.003	0.785	1.052
TDM	0.747	0.977	0.746	0.999	0.752	0.998	0.751	1.010	0.820	0.988	0.707	0.988	0.780	1.048
PSW-I	0.251	0.153	0.148	0.049	0.137	0.053	0.095	0.025	0.028	0.003	0.155	0.113	0.383	0.505
SPIRIT (Ours)	0.235	0.140	**0.136**	**0.043**	0.131	0.046	0.094	**0.020**	**0.028**	**0.003**	**0.140**	**0.097**	**0.245**	0.239
Win Counts	14	13	15	15	14	14	14	15	15	15	15	15	15	12

Kindly Note: Best results are marked with **bold**; second best results are marked with wavy underline.

Table 9: Block-Missing Results with Missing Patch Length 4.

Dataset	ETT-h1		ETT-h2		ETT-m1		ETT-m2		Exchange		Illness		Traffic	
Metric	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Crossformer	0.266	0.148	0.186	0.070	0.739	1.024	0.749	1.101	0.176	0.069	0.272	0.349	0.257	0.265
TimesNet	0.300	0.180	0.257	0.134	0.219	0.098	0.205	0.090	0.264	0.145	0.379	0.512	0.311	0.325
PatchTST	0.299	0.188	0.230	0.104	0.141	0.050	0.115	0.031	0.168	0.059	0.284	0.335	0.308	0.342
Autoformer	0.782	1.242	0.800	1.417	0.876	1.712	0.981	2.391	0.793	1.200	0.712	1.213	0.516	0.668
ETSformer	0.347	0.277	0.230	0.111	0.272	0.173	0.158	0.053	0.169	0.054	0.365	0.476	0.311	**0.256**
FiLM	0.459	0.458	0.276	0.150	0.259	0.142	0.214	0.095	0.189	0.065	0.354	0.431	0.776	1.153
DLinear	0.348	0.261	0.280	0.168	0.309	0.210	0.237	0.126	0.183	0.064	0.246	0.243	0.292	0.263
Glocal	0.321	0.210	0.226	0.104	0.163	0.063	0.135	0.037	0.166	0.054	0.344	0.407	0.377	0.463
HELIX	0.255	0.151	0.162	0.060	0.153	0.065	0.109	0.031	0.044	0.005	0.215	0.213	0.383	0.440
GP-VAE	0.350	0.232	0.264	0.131	0.288	0.163	0.221	0.095	0.306	0.179	0.475	0.570	0.681	0.934
CSDI	**0.250**	**0.141**	0.250	1.172	**0.132**	**0.047**	**0.084**	0.032	0.052	0.008	5.671	209.724	11.01	505.9
MIRI	0.504	0.498	0.344	0.356	0.388	0.358	0.263	0.274	0.298	0.192	0.268	0.242	0.333	0.335
Sinkhorn	0.746	0.977	0.752	1.039	0.754	0.996	0.747	0.998	0.826	1.002	0.760	1.334	0.791	1.089
TDM	0.745	0.974	0.750	1.036	0.754	0.995	0.747	0.997	0.822	0.997	0.750	1.318	0.787	1.087
PSW-I	0.277	0.190	0.158	0.057	0.147	0.060	0.101	0.029	0.031	0.004	0.187	0.162	0.439	0.622
SPIRIT (Ours)	0.261	0.175	**0.147**	**0.050**	0.137	0.050	0.096	**0.025**	**0.031**	**0.004**	**0.169**	**0.150**	**0.254**	0.273
Win Counts	13	12	15	15	14	14	14	15	15	15	15	15	15	12

Kindly Note: Best results are marked with **bold**; second best results are marked with wavy underline.

E.5 Empirical Time Complexity

We further report the empirical runtime of SPIRIT in Figure 7. As the dataset size increases, the "Score Learning" stage scales more steeply than the "Recursive Imputation" stage. This is expected since DSM training requires backpropagating through the network to compute gradients with respect to the inputs. Nevertheless, the overall runtime remains below 3 minutes even for the largest setting, demonstrating the computational efficiency of SPIRIT.

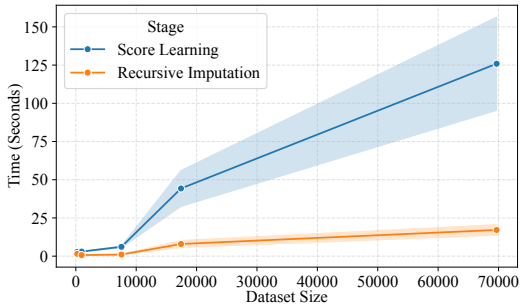
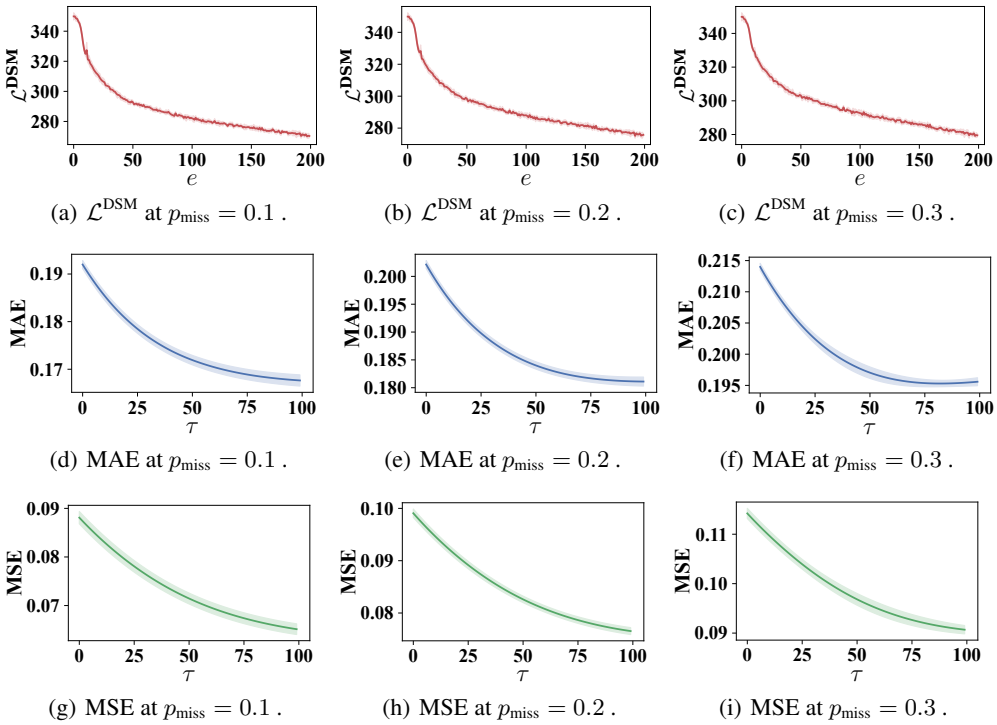


Figure 7: The computational time for the "Score Learning" and "Recursive Imputation" stages. The points and shaded areas indicate the mean and one standard deviation from the mean, respectively.

E.6 Empirical Convergence Analysis

In this subsection, we empirically validate the convergence claims discussed in Section C.6. Specifically, on the ETT-h1 dataset with missing ratios $p_{\text{miss}} \in \{0.1, 0.2, 0.3\}$, we track the score-matching loss $\mathcal{L}^{\text{DSM}}$ together with the imputation metrics (MAE/MSE) during the "Score Learning" and "Recursive Imputation" stages. As shown in Figures 8(a) to 8(c), $\mathcal{L}^{\text{DSM}}$ decreases steadily and then plateaus, indicating that the score-network optimization reaches a stable basin. Consistently, MAE and MSE in Figures 8(d) to 8(i) exhibit the same trend, stabilizing as training proceeds. Overall, these

results provide empirical evidence of SPIRIT's convergence behavior and support the discussion in Section C.6.



(a) $\mathcal{L}^{\text{DSM}}$ at $p_{\text{miss}} = 0.1$. (b) $\mathcal{L}^{\text{DSM}}$ at $p_{\text{miss}} = 0.2$. (c) $\mathcal{L}^{\text{DSM}}$ at $p_{\text{miss}} = 0.3$.

(d) MAE at $p_{\text{miss}} = 0.1$. (e) MAE at $p_{\text{miss}} = 0.2$. (f) MAE at $p_{\text{miss}} = 0.3$.

(g) MSE at $p_{\text{miss}} = 0.1$. (h) MSE at $p_{\text{miss}} = 0.2$. (i) MSE at $p_{\text{miss}} = 0.3$.

Figure 8: Convergence results comparison on ETT-h1 dataset. The lines and shaded areas indicate the mean and one standard deviation from the mean, respectively.

F Discussions on Limitations & Future Research Directions

Although SPIRIT achieves promising performance on TSDI, several limitations remain and may motivate future research:

- **Periodic structure and temporal dependence:** Following common practice (Shen & Kwok, 2023; Shen et al., 2024; Liu et al., 2025), we adopt techniques for image generation to handle time-series data. While effective, this design does not explicitly capture temporal autocorrelation or periodic patterns, which may limit its ability to handle temporal patterns, as evidenced by the empirical results in Section E.3.Future work may extend to frequency-domain diffusion models (Crabbé et al., 2024) or formulate the PSW (Wang et al., 2025c) as a proximal regularization term to more effectively preserve periodic structure.
- **Limited uncertainty quantification:** To prioritize accurate imputation, our derivation reduces the dissipative component, which may reduce sample diversity and, consequently, weaken uncertainty estimates. In the main text, we present a practical remedy inspired by ODE-based generative models, including the Denoising Diffusion Implicit Model (Song et al., 2021) and the Diffusion Exponential Integrator Sampler (Zhang & Chen, 2023). Specifically, we perturb the initial imputation values and generate multiple outputs by repeatedly running the iterative procedure. As reported in Section 4.2, this strategy maintains comparable imputation performance in the context of CRPS metric. Future work may explore multi-objective optimization (Gong et al., 2022) to explicitly balance accuracy and diversity. A complementary direction is to equip the network with a variational Bayesian last layer (Harrison et al., 2024; Brunzema et al., 2025), enabling posterior predictive sampling and more principled uncertainty estimation without redesigning the entire model.
- **Downstream-task awareness:** We focus on TSDI *per se*. Future work should investigate how to integrate SPIRIT with downstream models, for example, time-series forecasting (Qiu et al., 2024), to improve robustness under missing-data scenarios.
- **Beyond "Missing Completely At Random" to "Missing Not At Random" settings:** Our current formulation primarily targets missing completely at random. In many real applications, missingness depends on the unobserved values or the data-collection mechanism. Extending SPIRIT to MNAR settings may require explicitly modeling the missingness mechanism (Kyono et al., 2021) and incorporating distributionally robust optimization (Levy et al., 2020; Liu et al., 2022a) to improve reliability under mechanism shift.
- **Alternative score-learning strategies:** In this work, we adopt denoising score matching to learn the score function (Vincent, 2011). DSM requires computing gradients with respect to the inputs, which incurs additional backpropagation overhead and leads to higher runtime as the data scale increases. Future work could explore alternative score-learning objectives (Shi et al., 2018; Song et al., 2020a) or architectures that reduce input-gradient computation.