Beyond Correlations: Physics-Aware Temporal Modeling and Conditional Diffusion

...

Unifying flows, attention, conditional diffusion models (stochastic interpolants)

Siddharth Kamlesh Sharma (2022B4A70601P)
Aashish Kapoor (2022B5A70786P)

Recap : CSDI objective

Input: a data sample x_0, trained denoising function ε_θ

Output: Imputed missing values x_0^{ta}

- Simply simulate the reverse process, using the equations as mentioned previously
- Note that the said equations can be rewritten based on the parameterisation wrt our learned conditional denoising function ε_θ

Recap - CSDI Training Structure

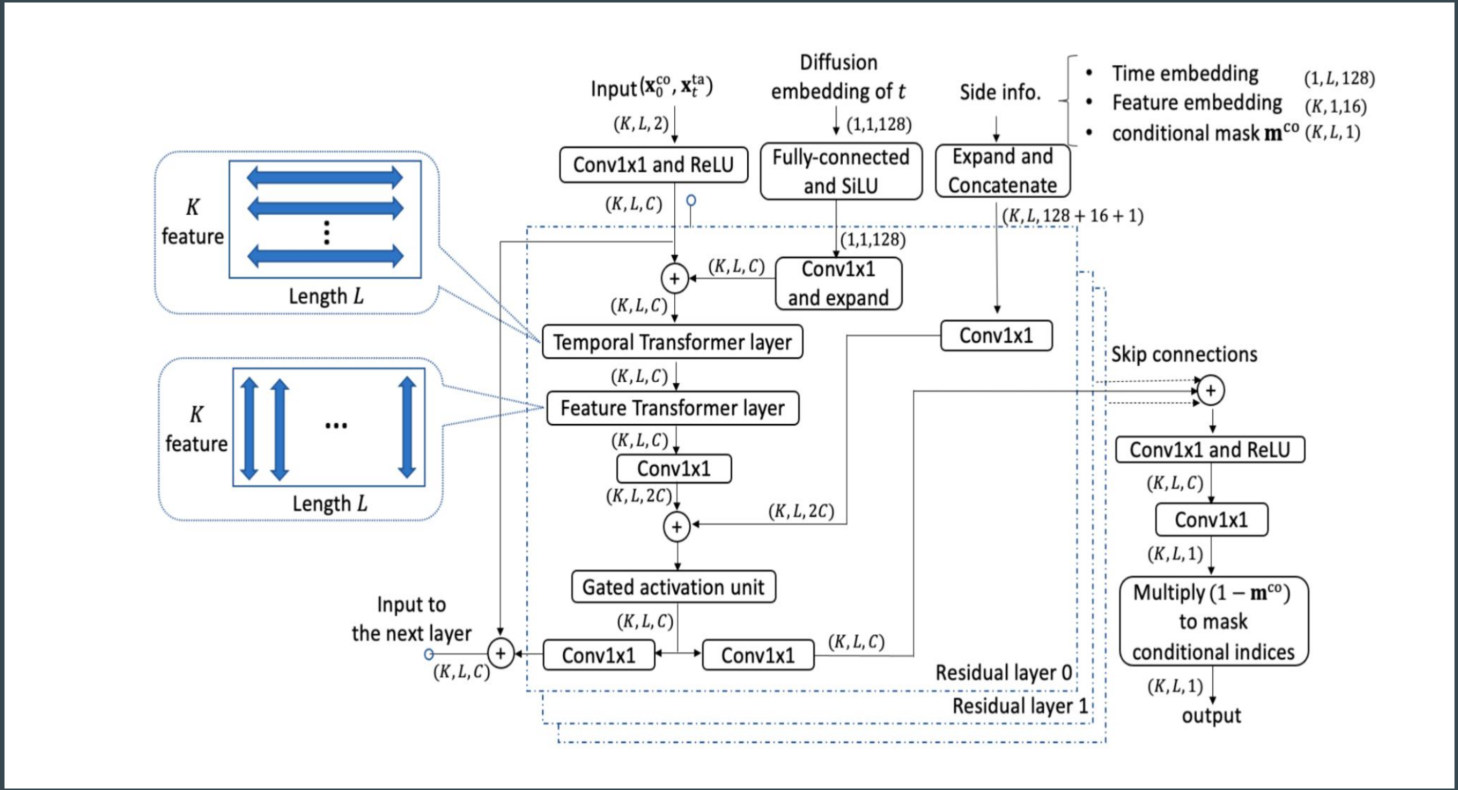
Given conditional observations x_0^{co} and imputation targets x_0^{ta}

We sample noisy targets $x_t^{\mathrm{ta}} = \sqrt{\alpha_t}\, x_0^{\mathrm{ta}} + (1 - \alpha_t)\, \varepsilon$

And train ε_θ by minimizing the following loss function:

$$\min_\theta \; \mathcal{L}(\theta) := \mathbb{E}_{x_0 \sim q(x_0),\, \varepsilon \sim \mathcal{N}(0,I),\, t} \left\| \varepsilon - \varepsilon_\theta\left(x_t^{\mathrm{ta}}, t \mid x_0^{\mathrm{co}}\right) \right\|_2^2 .$$

Recap - CSDI Architecture



Recap - **CSDI** Attention Mechanism

- Time series have two important correlations:
  - Temporal: How values change over time within each feature
  - Cross-sectional: How different variables affect each other at the same time point
- CSDI uses a smart 2D attention mechanism:
  - **Temporal Transformer Layer:** 1 layer transformer encoder to learn temporal dependencies across all K features.
  - **Feature Transformer Layer:** 1 layer transformer encoder to learn cross variable dependencies at each of the L time steps.

Physics-informed CSDI

- Introduce a plug-in physics loss acting on the denoised mean estimate, penalizing violations of known dynamical constraints during training.
- $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{phys}}\,\mathcal{L}_{\text{kinematic}}.$
- Apply a one step dataset specific projection at the end, pushing the sample closer to the physics manifold.

Physics-informed CSDI : Results

- Physics constraints substantially aid PDE-like fluids, especially under extreme sparsity.
- Improvements concentrate in distributional metrics (CRPS), highlighting feasibility enforcement rather than point error.
- For traffic, physics helps primarily when observations are scarce.
- Gains plateau due to CSDI's 1D architecture, which cannot fully leverage spatial physics (flattened representations limit physical expressivity).

Future Work

- Include physics in the diffusion process, not just as a mean penalty and a projection step.
- Better suited architecture for incorporating physics
    - UNet or ConvNet operating on (t, x, y) structure
    - Operator-Based Diffusion Models
    - Physics-Guided Noise Schedules
    - Hard Constraint Projection Within Each Diffusion Step
- Uncertainty aware physics corrections

Insight :
Transformers learn correlations while we want adherence to time *evolution*

...

HOW?

Background - Flows, Stochastic Interpolants

- A flow is a continuous-time transformation of a random variable. $dx_t/dt=f_\theta(x_t,t)$
- Flows move entire distributions. They are deterministic given initial conditions.
- Flow matching - Learns the vector field f_θ by matching velocities
- Key References:

1. Lipman et al., Flow Matching for Generative Modeling, arXiv:2210.02747
2. Kollovieh et al., Flow Matching with GP Priors, ICLR 2025
3. V Simkus & M Gutmann, CFMI: Flow Matching for Missing Data Imputation, arXiv:2506.09258

Why Did our Data Mining Attempt not Work?

- Using a plug-and-play loss estimate for enforcing physical laws clashes with the underlying structure used in CSDI
- The problem is the same - physical laws treat time not just as some index, rather than what is the basis of evolution
- Another problem - Inability to go from joint distribution to joint distribution ("cell to cell"), i.e. flattening of features leads to loss of important spatial context (velocity, pressure in 2D fluids/geographical data) necessary to enforce PDE-like physical constraints.
- A solution to the spatial problem has been explored by learning the copula (dependence structure) using transformers in this paper.

Our Proposed Solution - An outline

- Replace temporal attention with physics-aware flow / diffusion
- While retaining attention for conditioning (as done in CSDI)
- Aim to show: This separation enables consistent imputation and extrapolation under missing data while preserving flexibility in non-physical regimes.

Our technique is inspired by a combination of the following papers:

1. V Simkus & M Gutmann, CFMI: Flow Matching for Missing Data Imputation, arXiv:2506.09258
2. Bastek et al., Physics-Informed Diffusion Models, ICLR 2025
3. Li et al., Transformer-Modulated Diffusion Models for Probabilistic Multivariate Time Series Forecasting, ICLR 2024
4. Qian et al., Conditional Lagrangian Wasserstein Flow for Time Series Imputation, NeurIPS 2024 Reject

Testing Failure Modes - Hypotheses

- H1 - **Attention learns correlations and "disrespects" dynamics.** (Attention fails at global transport consistency).
- H2 - Flow matching based models **learn global operator structure** yet are *not adaptive to higher resolution imputation* (eg, local turbulences).
- H3 - **Soft physics penalties are architecturally insufficient.** (Loss-level penalties satisfy conditions in expectation but not per sample for which sample-wise hard projection is required)
- H4 (proposal) - In the broad framework of conditional diffusion models, our approach uses *flow for global dynamics* and *attention for local conditioning*.

Failure Modes - How? What Data?

- Dataset A: Linear Advection (Synthetic) : This is the easiest fluid flow adjacent setting that CSDI, CFMI, TMDM and PIDM should be capable of handling.
- Dataset B: ERA5 (ECMWF atmospheric reanalysis). This dataset represents transport-dominated system, and natural sparse observation patterns from weather stations. Moreover, it serves as a test in setting of absence of a clean analytical ground-truth solution (unlike above).

Failure Modes - (Additional) Metrics for Testing

1. **Semigroup (Operator) Consistency** (L2 norm of difference b/w model imputation output and direct "ground-truth" value):
   A low RMSE alongside high E_semi would imply the model *fits the data but does not encode the dynamics.*
2. **Conservation Error**
3. **Error as Horizon grows/Missingness increases**
4. **Wasserstein Distance** (Esp for H1 - penalises transport blurring which may be expected for attention based models)

Expected Behaviour of Models Across Metrics

- <u>Operator Consistency:</u>
    a. **CSDI, TMDM** : low RMSE, high E_semi (two-step rollout and direct imputation are mutually inconsistent)

    b. **CFMI, PIDM:** lower E_semi, competitive RMSE

- <u>Conservation Error:</u>

    a. Key idea would be check difference in mean and sample-wise error. CSDI, TMDM are expected to report lower error in the mean, but disproportionately higher sample-wise error.

    b. CFMI, PIDM are expected to lower the gap.

Expected Behaviour of Models Across Metrics

- Error Growth with Horizon / Sparsity :
    a. **CSDI, TMDM** : attention heavy models are expected to show a sharp increase after a point (especially in forecasting settings)

    b. **CFMI, PIDM:** lower E_semi, competitive RMSE

    c. TMDM degrades fastest when it has fewest anchors to average over.

- Wasserstein Distance :

    a. CSDI / TMDM: smear wind maximum across training-trajectory positions; W2 high even at moderate RMSE

Our Proposed Solution - Recap

- **Replace temporal attention with physics-aware flow / diffusion**
- **While retaining attention for conditioning (as done in CSDI)**
- Aim to show: This separation enables consistent imputation and extrapolation under missing data while preserving flexibility in non-physical regimes.
- If the mentioned failure modes are consistently identified and confirmed, our hybrid approach then has a clear motivation and simplifies architectural design.

Thank You