K^2VAE: A Koopman-Kalman Enhanced Variational AutoEncoder for Probabilistic Time Series Forecasting

Xingjian Wu [*1] **Xiangfei Qiu** [*1] **Hongfan Gao** [1] **Jilin Hu** [1] **Bin Yang** [1] **Chenjuan Guo** [1]

Abstract

Probabilistic Time Series Forecasting (PTSF) plays a crucial role in decision-making across various fields, including economics, energy, and transportation. Most existing methods excel at short-term forecasting, while overlooking the hurdles of Long-term Probabilistic Time Series Forecasting (LPTSF). As the forecast horizon extends, the inherent nonlinear dynamics have a significant adverse effect on prediction accuracy, and make generative models inefficient by increasing the cost of each iteration. To overcome these limitations, we introduce K^2VAE, an efficient VAE-based generative model that leverages a KoopmanNet to transform nonlinear time series into a linear dynamical system, and devises a KalmanNet to refine predictions and model uncertainty in such linear system, which reduces error accumulation in long-term forecasting. Extensive experiments demonstrate that K^2VAE outperforms state-of-the-art methods in both short- and long-term PTSF, providing a more efficient and accurate solution.

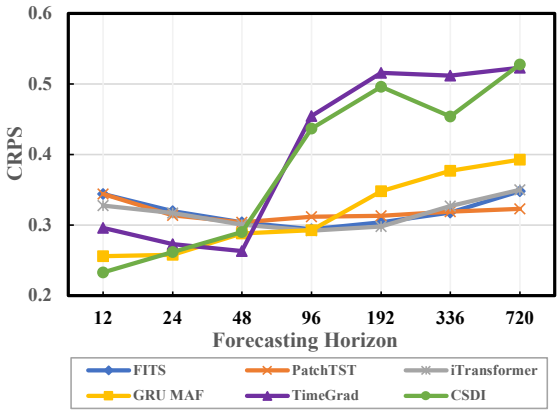

Figure 1. We compare three native probabilistic forecasting models including GRU MAF, TimeGrad, and CSDI with three point forecasting models equipped with distributional heads including FITS, PatchTST, and iTransformer on ETTh1. Longer forecasting horizons lead to rapid collapse of the CRPS metric (lower is better) on probabilistic forecasting models, even worse than point forecasting models.

1. Introduction

In recent years, time series analysis has seen remarkable progress, with key tasks such as anomaly detection (Wang et al., 2023; Liu & Paparrizos, 2024; Miao et al., 2025; Hu et al., 2024; Wu et al., 2025c), classification (Yao et al., 2024; Campos et al., 2023), and imputation (Gao et al., 2025; Wang et al., 2024a;c; Yu et al., 2025a), among others (Wang et al., 2024b; Miao et al., 2024a; Liu et al., 2025a; Huang et al., 2023; Yao et al., 2023), gaining attention. Among these, Probabilistic Time Series Forecasting (PTSF) is a crucial and widely studied task. By quantifying the stochastic temporal evolutions of multiple continuous variables, it provides significant support for decision-making in various fields such as economics (Sezer et al., 2020; Huang et al., 2022b), traffic (Wu et al., 2024b; 2025d; Pan et al., 2023a; Cirstea et al., 2022b; Fang et al., 2021; Yang et al., 2021; 2022), energy (Wang et al., 2024d; Guo et al., 2015; Sun et al., 2022), and AIOps (Lin et al., 2024a; Campos et al., 2022; Chen et al., 2023; Pan et al., 2023c; Lin et al., 2025). In these practical applications, an urgent need is to extend the prediction time to the distant future, known as Long-term Probabilistic Time Series Forecasting (LPTSF), which is highly meaningful for long-term planning and early warning. Most existing methods excell at short-term problem settings, such as predicting up to 48 steps or fewer (Rasul et al., 2021a; Kollovieh et al., 2023; Rasul et al., 2021b), while directly applying these methods to long-term forecasting tasks often results in poor performance–see Figure 1.

However, probabilistic forecasting faces numerous challenges in long-term forecasting tasks. *First, the inherent nonlinearity of time series challenges probabilistic models in modeling dynamic evolution.* Due to factors such as non-stationarity and complex interdependencies between

[*]Equal contribution [1]School of Data Science and Engineering, East China Normal University, Shanghai, China. Correspondence to: Bin Yang <byang@dase.ecnu.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

variables, time series typically exhibit nonlinear characteristics, complicating the construction of probabilistic models. Specifically, the nonlinearity makes it difficult for these models to derive a simple equation that precisely describes the state transition process. As a result, the uncertainties within the models are also hard to quantify, particularly in long-term forecasting tasks. *Second, as the forecasting horizon extends, the accuracy and efficiency become major bottlenecks.* Longer foreacasting horizons lead to more intricate target distributions, which causes remarkable error accumulation. It also makes the diffusion-based (Kollovieh et al., 2023; Rasul et al., 2021a) or flow-based models (Rasul et al., 2021b) difficult to find a clear probabilistic transition path and inefficient to perform each iteration, which results in more computational consumption but worse performance.

Since the nonlinearity in time series leads to the dynamic evolution of complex patterns, probabilistic models struggle to effectively capture these changes and accurately model their evolution. To tackle this thorny issue, Koopman Theory (Koopman, 1931) provides a linearization approach to transform the nonlinear time series into the space of measurement function, which is a theoretically infinite-dimeansional space characterizing all measurements of the dynamical system at each moment, and the transition process of these measurements can be captured by a linear Koopman Operator (Lan & Mezić, 2013). On the other hand, in order to accurately and efficiently model the process uncertainty and mitigate the error accumulation phenomenon in long-term forecasting, the Kalman Filter (Welch, 1995) provides a solution, which fuses observations from multiple sensors to extract the Kalman gains, to refine the prediction and process uncertainty. This inspires us to transform the probabilistic time series forecasting into modeling the process uncertainty of a linear dynamical system in the space of the measurement function.

In this study, we propose K^2VAE, a generative probabilistic forecasting model tailored for LPTSF–see Figure 2. First, to handle the nonlinearity and capture the underlying dynamics in time series, we patchify the time series into tokens and model them through the KoopmanNet. The KoopmanNet provides a way to simulate the Koopman Theory, which transforms the nonlinear time series into latent measurements, and fit the Koopman Operator to construct a "biased" linear dynamical system easy to describe and model. Second, to achieve accurate long-term forecasting performance, we design a KalmanNet in a data-driven manner based on the principle of Kalman Filter. Through integrating the residual nonlinear information as control inputs, while treating the biased linear dynamical system as the observation, the KalmanNet predicts and updates to model and refine the uncertainty with Kalman gain. This effectively mitigates the error accumulation of the linear system and helps construct the variational distribution in the space of the measurement function with clear semantics. Compared to diffusion-based models or flow-based models with longer generation processes, which cause more computational consumption and memory overhead, K^2VAE adopts a VAE-based structure composed of lightweight but effective KoopmanNet and KalmanNet, which contributes to fast one-step generation and lower memory occupation. The contributions are summarized as follows:

- To address PTSF, we propose an efficient framework called K^2VAE. It transforms nonlinear time series into a linear dynamical system. Through predicting and refining the process uncertainty of the system, K^2VAE demonstrates strong generative capability and excells in both the short- and long-term probabilistic forecasting.
- To distengle the complex nonlinearity in the time series, we design a KoopmanNet to fully exploit the underlying linear dynamical characteristics in the space of measurement function, simplify the modeling, and thus contributing to high model efficiency.
- To mitigate the error accumulation in LPTSF, we devise a KalmanNet to model, and refine the prediction and uncertainty iteratively.
- Comprehensive experiments on both short- and long-term PTSF show that K^2VAE outperforms state-of-the-art baselines. Additionally, all datasets and code are avaliable at `https://github.com/decisionintelligence/K2VAE`.

2. Preliminaries

Koopman Theory. Koopman Theory (Koopman, 1931; Lan & Mezić, 2013) is a widely used mathematical tool for dynamic system analysis, providing a way to linearize the nonlinear systems. For nonlinear system $x_{k+1} = f(x_k)$, where x_k denotes system state and f is a nonlinear function, it assumes that the system's state can be mapped into the space of measurement function ψ, where it can be modeled by an infinite-dimensional linear Koopman Operator $\mathcal{K}$:

$$\psi(x_{k+1}) = \psi(f(x_k)) = \mathcal{K} \circ \psi(x_k) \tag{1}$$

Koopman Theory helps understand the underlying dynamics of complex nonlinear systems and serves as a powerful tool to linearize them for ease of process.

Kalman Filter. Kalman Filter (Welch, 1995; Simon, 2001) is a recursive algorithm used for estimating the state of a linear dynamic system. It works in two steps: first, it predicts the current state x_k and uncertainty covariance matrix P_k based on the system's state transition equation; then, it updates the estimation by incorporating the difference between the measurement and prediction, known as Kalman gain K_k. The Kalman Filter effectively fuses information

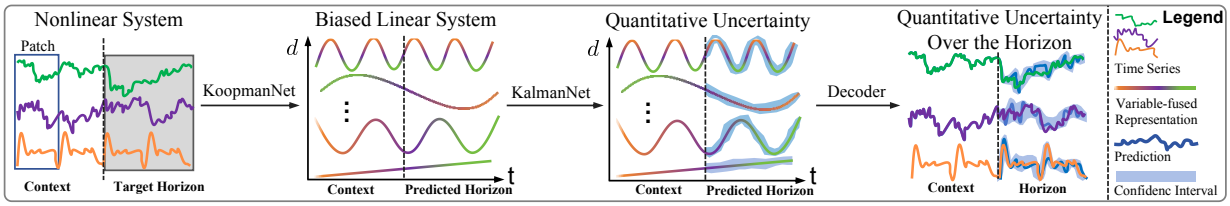

Figure 2. The data flow of K^2VAE. It models time series through the KoopmanNet, which constructs a biased linear system. Then the linear system is refined through the KalmanNet while the uncertainty is modeled. Finally, the target distributions over the horzion are predicted through the Decoder.

from multiple sensors to enhance estimation accuracy while modeling the uncertainty of the system.

VAE for Probabilistic Time Series Forecasting. PTSF can be treated as a conditional generative task, i.e., generating forecasting horizon $\hat{Y} = [\hat{x}_{T+1}, \hat{x}_{T+2}, \cdots, \hat{x}_L] \in \mathbb{R}^{N \times L}$ given context series $X = [x_1, x_2, \cdots, x_T] \in \mathbb{R}^{N \times T}$, where N denotes the number of variables, T denotes the context length, and L denotes the forecasting horizon. The objective is to model the conditional distribution $\mathbb{P}(Y|X)$ and sample from it to obtain $\hat{Y}$. When using Variational AutoEncoder (Higgins et al., 2017; Pu et al., 2016), the log-likehood objective is optimized through the Evidence Lower Bound (2) which is obtained by Jensen Inequality:

$$\mathcal{L}_{ELBO} = \\ -\mathbb{E}[\log \mathbb{P}(Y|Z,X)] + D_{KL}(\mathbb{Q}(Z|X)||\mathbb{P}(Z|X)) \quad (2)$$

In our proposed K^2VAE, we meticulously construct the variational distribution $\mathbb{Q}(Z|X)$, aligning it with the uncertainty of the dynamical system. This endows the latent space in K^2VAE with clear semantics, enhancing its generative capabilities in PTSF.

3. Methodology

3.1. K^2VAE Architecture

As demonstrated in Figure 3, K^2VAE consists of four main components: Input Token Embedding, KoopmanNet, KalmanNet, and Decoder. The KoopmanNet and KalmanNet consitute the Encoder of K^2VAE. To facilitate comprehension, we present the Data Flow–see Figure 2.

Overall, K^2VAE employs a meticulously designed pipeline to model the time series at the perspective of dynamic system. First, the Input Token Embedding module patchifys the time series into tokens. Then the KoopmanNet projects them into the space of measurement function, where the inherent nonlinearity and intricate joint distribution between variables are reconsidered for ease. Sequentially, the Koopman Operator is fitted and iterates over the first token to delineate a linear system. Obviously, the perfect measurement function which constructs an absolute linear system is the ideal objective, which means the series generated by the Koopman Operator is biased. We then design the KalmanNet to refine such biased linear system and model the uncertainty by outputting the covariance matrix of multi-dimensional state vector, which assigns the variational posterior distribution $\mathbb{Q}(Z|X)$ in the space of measurement function with clear semantics. The Decoder works as the inverse measurement function ψ^{-1} to map the samples from $\mathbb{Q}(Z|X)$ to the original space, which also serves as the decoder of VAE and models the target distribution $\mathbb{P}(Y|Z,X)$ of the forecasting horizon.

3.1.1. Input Token Embedding

Since Triformer (Cirstea et al., 2022a) first proposes the Patching technique, existing works (Nie et al., 2023; Wu et al., 2025c) demonstrate that considering a patch as the "token" retains most semantic information and helps establish meaningful state transition procedure for autoregressive models. Our proposed K^2VAE also works like an autoregressive dynamic system to model the state transition procedure. Different from those Channel-Independent models which divides patches for each channel and projects them independently, we consider multivariate patches as tokens to implicitly model the cross-variable interaction during state transition. We divide the context series $X = [x_1, x_2, \cdots, x_T] \in \mathbb{R}^{N \times T}$ into non-overlapping patches:

$$X^P = \left[x_1^P, x_2^P, \cdots, x_n^P\right] \in \mathbb{R}^{N \times s \times n}, \quad (3)$$

where $s = T/n$ denotes the patch size, n denotes the patch number, and $x_i^P \in \mathbb{R}^{N \times s}$ denotes a patch. Then X^P are embeded into high-dimensional hidden space:

$$X^{P'} = \text{Projection}(\text{Flatten}(X^P)), \quad (4)$$

where patches are first flattened into $\mathbb{R}^{(N \times s) \times n}$ and then mapped into embeddings $X^{P'} \in \mathbb{R}^{d \times n}$ through a linear projection to fuse the variable information.

3.1.2. K^2VAE Encoder

Linearizing with the KoopmanNet. Since there exists variable-wise periodic misalignment or temporal non-stationarity in realistic multivariate time series, yielding

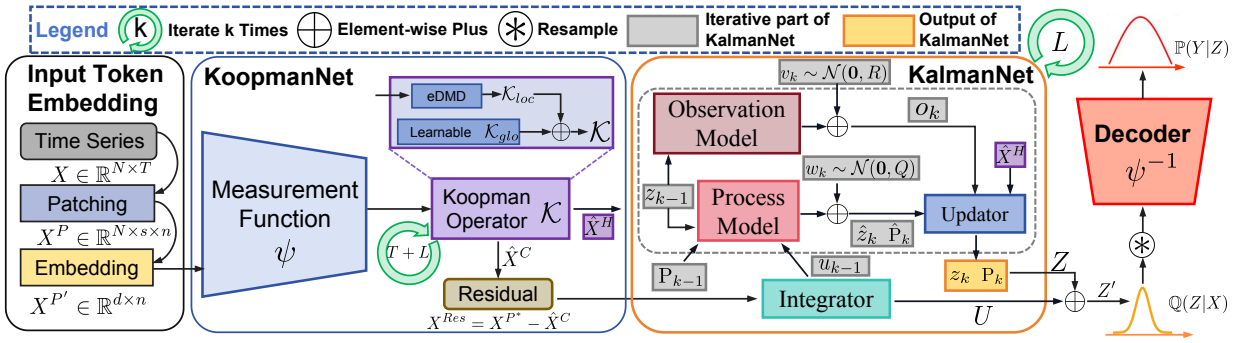

Figure 3. The architecture of K^2VAE. Input Token Embedding Module patchifys the time series into tokens and applies Embedding. Encoder Module consists of KoopmanNet and KalmanNet, transforming the tokens into linear system in the space of measurement function, refining it and modeling the process uncertainty as the variational distribution. After resampling from the variational distribution, Decoder module constructs the likehood distribution about the forecasting horizon.

non-linearity, K^2VAE applies Koopman Theory (Koopman, 1931) to construct the measurement function to project the system states into measurements which can be modeled as a linear system. Practically, we use a learnable MLP-based network to serve as the measurement function ψ:

$$X^{P^*} = \psi(X^{P'}) = \left[x_1^{P^*}, x_2^{P^*}, \cdots, x_n^{P^*}\right], \tag{5}$$

where $X^{P^*} \in \mathbb{R}^{d\times n}$ denotes the projected tokens in the measurement space. To capture the transition rule, we utilize the one-step eDMD (Schmid, 2010; Liu et al., 2023) over X^{P^*} to efficiently find the best fitted $\mathcal{K}_{loc}$:

$$X_{back}^{P^*} = \left[x_1^{P^*}, x_2^{P^*}, \cdots, x_{n-1}^{P^*}\right], \tag{6}$$

$$X_{fore}^{P^*} = \left[x_2^{P^*}, x_3^{P^*}, \cdots, x_n^{P^*}\right], \tag{7}$$

$$\mathcal{K}_{loc} = X_{fore}^{P^*}(X_{back}^{P^*})^{\dagger}, \tag{8}$$

where $(X_{back}^{P^*})^{\dagger}$ denotes the Moore-Penrose inverse of $X_{back}^{P^*}$. $\mathcal{K}_{loc}$ effectively captures the local transition rule in the space of current measurement function. However, when ψ is underfitted, the low quality of the space may cause numerical instability or guide the model to converge in a wrong direction. To mitigate this issue and capture the global-shared dynamics, we introduce a learnable part $\mathcal{K}_{glo}$. Then we delineate the system through the Koopman Operator $\mathcal{K} = \mathcal{K}_{loc} + \mathcal{K}_{glo}$:

$$\hat{X}^C = \left[\hat{x}_1^C, \hat{x}_2^C, \cdots, \hat{x}_n^C\right], \tag{9}$$

$$\hat{X}^H = \left[\hat{x}_1^H, \hat{x}_2^H, \cdots, \hat{x}_m^H\right], \tag{10}$$

$$\hat{x}_i^C = (\mathcal{K})^{i-1} x_1^{P^*}, \hat{x}_i^H = (\mathcal{K})^{i+n-1} x_1^{P^*}, \tag{11}$$

where $\hat{X}^C \in \mathbb{R}^{d\times n}$ denotes the reconstruction context generated by Koopman Operator $\mathcal{K} \in \mathbb{R}^{d\times d}$ and $\hat{X}^H \in \mathbb{R}^{d\times m}$ is the predicted horizon, $m = L/s$ means that predicting L steps in the original space is equivalent to predicting m steps in the space of measurement function.

Modeling the Uncertainty with the KalmanNet. Since we adopt a data-driven paradigm to model the measurement function ψ and Koopman Operator $\mathcal{K}$, it exists bias between the generated $\hat{X}^C$ and X^{P^*} during optimization, known as a biased linear system. Inspired by Kalman Filter (Welch, 1995; Simon, 2001) which is born to refine such biased linear sytem, we devise a KalmanNet to model and refine the uncertainty adaptively, aligning it with the variational distribution $\mathbb{Q}(Z|X)$ in the latent measurement space. Specifically, we first fully reuse the nonlinear residual through the Integrator based on an Encoder-Only Vanilla Transformer (Vaswani et al., 2017):

$$X^{Res} = X^{P^*} - \hat{X}^C, \tag{12}$$

$$U = \text{Integrator}(X^{Res}) = [u_1, u_2, \cdots, u_m], \tag{13}$$

where $U \in \mathbb{R}^{d\times m}$ denotes the output integrated by the Integrator. We then construct the Process Model of KalmanNet, which describes the state transition process:

$$z_k = Az_{k-1} + Bu_k + w_k, \tag{14}$$

$$z_0 = x_n^{P^*}, \tag{15}$$

where $A \in \mathbb{R}^{d\times d}$ is the state transition matrix, $B \in \mathbb{R}^{d\times d}$ is the control input matrix, and $w_k \sim \mathcal{N}(\mathbf{0}, Q)$ is the process noise and Q is its covariance matrix. Sequentially, we construct the Observation Model:

$$o_k = Hz_k + v_k, \tag{16}$$

where $H \in \mathbb{R}^{d\times d}$ is the observation matrix and we treat the prediction $\hat{X}^H$ as the prior observation in Update Step (20). $v_k \sim \mathcal{N}(\mathbf{0}, R)$ is the observation noise and R is its covariance matrix. Our goal is to reuse the information from the nonlinear residual, and integrate it into the linear system constructed by KoopmanNet, thus obtaining a more accurate linear system and modeling the uncertainty. In the KalmanNet, all the matrices are learnable. Additionally, we

initialize the covariance matrices Q and R as identity matrices and use lower triangular matrices L_Q and L_R to keep the positive definiteness: $Q = L_Q L_Q^T$ and $R = L_R L_R^T$.

Then we conduct the Prediction Step and Update Step iteratively, the Prediction Step can be formulated as:

$$\hat{z}_k = Az_{k-1} + Bu_k, \tag{17}$$

$$\hat{\mathrm{P}}_k = A\mathrm{P}_{k-1}A^T + Q, \tag{18}$$

where $\hat{z}_k$ is the predicted state and $\hat{\mathrm{P}}_k$ is the predicted covariance matrix of the process uncertainty. Then the Update Step measures the weight between observation and prediction through Kalman gain K_k to refine the system:

$$K_k = \hat{\mathrm{P}}_k H^T (H\hat{\mathrm{P}}_k H^T + R)^{-1}, \tag{19}$$

$$z_k = \hat{z}_k + K_k(\hat{x}_k^H - H\hat{z}_k), \tag{20}$$

$$\mathrm{P}_k = (I - K_k H)\hat{\mathrm{P}}_k, \tag{21}$$

where z_k and P_k is the refined state vector and covariance matrix. We then obtain the refined predictions $Z = [z_1, z_2, \cdots, z_m]$ and covariance matrices of each token $\mathrm{P} = [\mathrm{P}_1, \mathrm{P}_2, \cdots, \mathrm{P}_m]$, which describes the temporal process uncertainty in the dynamical system. We show that the process also obeys the basic assumptions of Koopman Theory in Section 3.2. To fully utilize the ability of the Integrator, we make a skip connection:

$$Z' = Z + U \tag{22}$$

During the training process, the model leverages the Integrator to integrate nonlinear information and gradually adjust the topological structure of the measurement space. Optimized by $\mathcal{L}_{Rec}$ (27), the deviation of the linear system constructed by the KoopmanNet gradually decreases, causing $U \to \mathbf{0}$. This facilitates a linear dynamical system in the measurement space and gradually reduces dependence on the Integrator.

3.1.3. K^2VAE Decoder

After obtaining the prediction Z', and the covariance matrix P of process uncertainty, the variational distribution is formulated as $\mathbb{Q}(Z|X) = \mathcal{N}(Z', \mathrm{P})$. During training, we conduct reparameterization sampling from it to keep the ensure the propagation of the gradient. Finally, we utilize the Decoder to map the samples back to the original space and model the $\mathbb{P}(Y|Z)$ with an isotropic Gaussian distribution. Specifically, the Decoder consists of two same MLP structures as the inverse of the Koopman Encoder ψ, we formalize them as ψ_μ^{-1} and ψ_σ^{-1}:

$$Z^{sample} = \text{Resample}(\mathbb{Q}(Z|X)), \tag{23}$$

$$\mu = \psi_\mu^{-1}(Z^{sample}), \sigma = \psi_\sigma^{-1}(Z^{sample}), \tag{24}$$

$$X^{Rec} = \psi_\mu^{-1}(\hat{X}^C), \tag{25}$$

so that the $\mathbb{P}(Y|Z) = \mathcal{N}(\mu, \sigma)$ is modeled. We also map back the Koopman reconstruction $\hat{X}^C$ from the measurement space to optimize the $\mathcal{L}_{Rec}$ (27), which helps measurement function ψ to build a linear system.

3.1.4. Overall Learning Objective

The overall learning objective is weightsumed by $\mathcal{L}_{ELBO}$ and $\mathcal{L}_{Rec}$:

$$\mathcal{L}_{ELBO} = -\mathbb{E}[\log \mathbb{P}(Y|Z,X)] + D_{KL}(\mathbb{Q}(Z|X)||\mathbb{P}(Z|X)), \tag{26}$$

$$\mathcal{L}_{Rec} = ||X - X^{Rec}||_2^2, \tag{27}$$

where the $\mathcal{L}_{ELBO}$ ensures the fundmental mechanism of K^2VAE. The prior distribution is $\mathbb{P}(Z|X) = \mathcal{N}(\mathbf{0}, I)$, where we hope the linear system in measurement space converge to a stable state. $\mathcal{L}_{Rec}$ facilitates the linearization of the measurement space.

3.2. Theoretical Analysis

3.2.1. The Stability of KalmanNet

Since the proposed KalmanNet works in a data-driven manner, the floating-point operation error may cause the covariance matrix P losing positive definiteness, which often occurs in the step (21). To mitigate this, we utilize a numerically stable form for this step.

Theorem 3.1. *The positive-definiteness of covariance matrix P_k during the update step $\mathrm{P}_k = (I - K_k H_k)\hat{\mathrm{P}}_k$ can be retained through a numerically stable form:*

$$\mathrm{P}_k = \frac{1}{2}(\mathrm{P}_k + \mathrm{P}_k^T), \tag{28}$$

$$\mathrm{P}_k^{dual} = (I - K_k H_k)\hat{\mathrm{P}}_k(I - K_k H_k)^T + K_k R_k K_k^T, \tag{29}$$

where (28) ensures the symmetry, (29) stabilizes the positive-definiteness by decomposing the formula into the sum of two positive definite terms, which better ensures positive definiteness during floating operation.

3.2.2. The Convergence of K^2VAE

Since K^2VAE models a linear dynamical system in the measurement space where the Koopman Operator serves as the state transition equation, we hope that the convergence state of the KalmanNet does not violate the assumptions of Koopman Theory. In K^2VAE, we meticulously design the KalmanNet by making it gradually converge to the Koopman Operator in the forecasting horizon.

Theorem 3.2. *When $U \to \mathbf{0}$, the state transition equation of the KalmanNet in K^2VAE gradually converges to the Koopman Operator.*

We provide the proof of Theorem 3.1–3.2 in Appendix A.

Table 1. Statistical information of the datasets.

Horizon	Dataset	#var.	range	freq.	timesteps	Description
Long-term	ETTh1/h2-L	7	$\mathbb{R}^+$	H	17,420	Electricity transformer temperature per hour
	ETTm1/m2-L	7	$\mathbb{R}^+$	15min	69,680	Electricity transformer temperature every 15 min
	Electricity-L	321	$\mathbb{R}^+$	H	26,304	Electricity consumption (Kwh)
	Traffic-L	862	(0,1)	H	17,544	Road occupancy rates
	Exchange-L	8	$\mathbb{R}^+$	Busi. Day	7,588	Daily exchange rates of 8 countries
	ILI-L	7	(0,1)	W	966	Ratio of patients seen with influenza-like illness
	Weather-L	21	$\mathbb{R}^+$	10min	52,696	Local climatological data
Short-term	ETTh1/h2-S	7	$\mathbb{R}^+$	H	17,420	Electricity transformer temperature per hour
	ETTm1/m2-S	7	$\mathbb{R}^+$	15min	69,680	Electricity transformer temperature every 15 min
	Exchange-S	8	$\mathbb{R}^+$	Busi. Day	6,071	Daily exchange rates of 8 countries
	Solar-S	137	$\mathbb{R}^+$	H	7,009	Solar power production records
	Electricity-S	370	$\mathbb{R}^+$	H	5,833	Electricity consumption
	Traffic-S	963	(0,1)	H	4,001	Road occupancy rates

Table 2. Comparison on short-term probabilistic forecasting scenarios across eight real-world datasets. Lower CRPS or NMAE values indicate better predictions. The means and standard errors are based on 5 independent runs of retraining and evaluation. **Red**: the best, Blue: the 2nd best.

Model	Metric	Exchange-S	Solar-S	Electricity-S	Traffic-S	ETTh1-S	ETTh2-S	ETTm1-S	ETTm2-S
FITS	CRPS	0.012±0.002	0.516±0.011	0.068±0.003	0.298±0.022	0.320±0.017	0.212±0.012	0.193±0.005	0.199±0.003
	NMAE	0.017±0.003	0.701±0.014	0.092±0.004	0.392±0.028	0.423±0.033	0.278±0.009	0.249±0.007	0.260±0.011
PatchTST	CRPS	0.052±0.016	0.491±0.008	0.063±0.003	0.278±0.018	0.314±0.022	0.207±0.006	0.234±0.011	0.212±0.018
	NMAE	0.069±0.013	0.663±0.010	0.085±0.006	0.363±0.023	0.407±0.030	0.260±0.009	0.271±0.009	0.257±0.011
iTransformer	CRPS	0.059±0.018	0.504±0.012	0.066±0.004	0.244±0.011	0.317±0.020	0.219±0.008	0.254±0.012	0.201±0.018
	NMAE	0.081±0.022	0.695±0.017	0.087±0.006	0.319±0.019	0.408±0.028	0.276±0.017	0.291±0.017	0.242±0.009
Koopa	CRPS	0.012±0.001	0.545±0.016	0.085±0.014	0.253±0.018	0.326±0.013	0.211±0.019	0.288±0.022	0.220±0.015
	NMAE	0.015±0.002	0.742±0.022	0.112±0.019	0.330±0.019	0.423±0.017	0.266±0.022	0.329±0.026	0.278±0.022
TSDiff	CRPS	0.077±0.019	0.568±0.015	0.111±0.013	0.189±0.009	0.304±0.016	0.204±0.006	0.209±0.013	0.124±0.008
	NMAE	0.096±0.024	0.635±0.012	0.115±0.018	0.206±0.011	0.400±0.025	0.272±0.015	0.276±0.008	0.162±0.008
D^3VAE	CRPS	0.011±0.002	0.769±0.029	0.071±0.009	0.143±0.008	0.324±0.019	0.216±0.015	0.198±0.015	0.303±0.024
	NMAE	0.012±0.002	0.998±0.049	0.092±0.013	0.178±0.013	0.410±0.016	0.267±0.018	0.250±0.018	0.378±0.031
GRU NVP	CRPS	0.019±0.006	0.530±0.008	0.062±0.003	0.168±0.008	0.398±0.034	0.309±0.023	0.455±0.029	0.276±0.014
	NMAE	0.024±0.007	0.670±0.011	0.081±0.006	0.209±0.013	0.477±0.040	0.375±0.024	0.584±0.047	0.349±0.028
GRU MAF	CRPS	0.012±0.003	0.486±0.007	0.056±0.002	0.144±0.022	0.258±0.013	0.160±0.008	0.151±0.009	0.146±0.011
	NMAE	0.016±0.002	0.603±0.009	0.073±0.004	0.182±0.029	0.326±0.016	0.208±0.003	0.198±0.004	0.193±0.008
Trans MAF	CRPS	0.012±0.001	0.442±0.011	0.054±0.002	0.133±0.004	0.309±0.009	0.200±0.012	0.139±0.005	0.180±0.010
	NMAE	0.016±0.001	0.577±0.014	0.071±0.003	0.160±0.006	0.400±0.011	0.256±0.009	0.162±0.006	0.224±0.009
TimeGrad	CRPS	0.009±0.001	0.465±0.016	0.057±0.002	0.130±0.005	0.273±0.007	0.184±0.006	0.186±0.003	0.148±0.004
	NMAE	0.012±0.002	0.609±0.015	0.073±0.004	**0.155**±0.007	0.356±0.013	0.224±0.014	0.246±0.007	0.189±0.006
CSDI	CRPS	0.009±0.001	0.392±0.006	**0.051**±0.001	0.147±0.014	0.262±0.012	0.133±0.006	0.140±0.012	0.144±0.018
	NMAE	0.013±0.001	0.533±0.007	**0.066**±0.001	0.175±0.013	0.339±0.009	0.161±0.013	0.169±0.021	0.181±0.024
K^2VAE	CRPS	**0.009**±0.001	**0.367**±0.005	0.053±0.002	**0.129**±0.004	**0.256**±0.008	**0.128**±0.006	**0.135**±0.008	**0.122**±0.008
	NMAE	**0.009**±0.001	**0.480**±0.008	0.068±0.002	0.157±0.007	**0.312**±0.008	**0.140**±0.007	**0.152**±0.007	**0.146**±0.009

4. Experiments

In this section, we provide empirical results to show the strong performance of K^2VAE against state-of-art baselines on both short- and long-term probabilistic forecasting tasks. We also analyze the model efficiency and the key parameters of K^2VAE as the proof of architectural superiority.

4.1. Experimental Setup

Datasets. We conduct experiments on 8 datasets of short-term forecasting and 9 datasets of long-term forecasting based on ProbTS (Zhang et al., 2024a), a comprehensive benchmark used to evaluate probabilistic forecasting tasks. Specifically, we use the datasets ETTh1-S, ETTh2-S, ETTm1-S, ETTm2-S, Electricity-S, Solar-S, Traffic-S, and Exchange-S for short-term forecasting, of which the context length T is equivalent to forecasting horizon L with $T = L = 30$ for Exchange-S and $T = L = 24$ for the others. For long-term forecasting, we use the datasets ETTh1-L, ETTh2-L, ETTm1-L, ETTm2-L, Electricity-L, Traffic-L, Exchange-L, Weather-L, and ILI-L with forecasting horizon $L \in \{24, 36, 48, 60\}$ for ILI-L and $L \in \{96, 192, 336, 720\}$ for the others. Note that we fix the context length of all the models with $T = 36$ for ILI-L and $T = 96$ for the others to

Table 3. Comparison on long-term probabilistic forecasting (forecasting horizon L=720) scenarios across nine real-world datasets. Lower CRPS or NMAE values indicate better predictions. The means and standard errors are based on 5 independent runs of retraining and evaluation. **Red**: the best, Blue: the 2nd best. The full results of all four horizons 96, 192, 336, 720 are listed in Table 9, 10 in Appendix C.5.

Model	Metric	ETTm1-L	ETTm2-L	ETTh1-L	ETTh2-L	Electricity-L	Traffic-L	Weather-L	Exchange-L	ILI-L
FITS	CRPS	0.305±0.024	0.449±0.034	0.348±0.025	0.314±0.022	0.115±0.024	0.374±0.004	0.267±0.003	0.074±0.011	0.211±0.011
	NMAE	0.406±0.072	0.540±0.052	0.468±0.012	0.401±0.022	0.149±0.012	0.453±0.022	0.317±0.021	0.097±0.011	0.245±0.017
PatchTST	CRPS	0.304±0.029	0.229±0.036	0.323±0.020	0.304±0.018	0.127±0.015	0.214±0.001	0.142±0.005	0.097±0.007	0.233±0.019
	NMAE	0.382±0.066	0.288±0.034	0.428±0.024	0.371±0.021	0.164±0.024	0.253±0.012	0.152±0.029	0.126±0.001	0.287±0.023
iTransformer	CRPS	0.455±0.021	0.311±0.024	0.350±0.019	0.542±0.015	0.109±0.044	0.284±0.004	0.133±0.004	0.087±0.023	0.222±0.020
	NMAE	0.490±0.038	0.385±0.042	0.449±0.022	0.667±0.012	0.140±0.009	0.361±0.030	0.147±0.019	0.113±0.015	0.278±0.017
Koopa	CRPS	0.295±0.027	0.233±0.025	0.318±0.009	0.293±0.026	0.113±0.018	0.358±0.022	0.140±0.007	0.091±0.012	0.228±0.022
	NMAE	0.377±0.037	0.290±0.033	0.412±0.008	0.286±0.042	0.149±0.025	0.432±0.032	0.162±0.009	0.116±0.022	0.288±0.031
TSDiff	CRPS	0.478±0.027	0.344±0.046	0.516±0.027	0.406±0.056	0.478±0.005	0.391±0.002	0.152±0.003	0.082±0.010	0.263±0.022
	NMAE	0.622±0.045	0.416±0.065	0.657±0.017	0.482±0.022	0.622±0.142	0.478±0.006	0.141±0.026	0.142±0.009	0.272±0.020
GRU NVP	CRPS	0.546±0.036	0.561±0.273	0.502±0.039	0.539±0.090	0.114±0.013	0.211±0.004	0.110±0.004	0.079±0.009	0.307±0.005
	NMAE	0.707±0.050	0.749±0.385	0.643±0.046	0.688±0.161	0.144±0.017	0.264±0.006	0.135±0.008	0.103±0.009	0.333±0.005
GRU MAF	CRPS	0.536±0.033	0.272±0.029	0.393±0.043	0.990±0.023	0.106±0.007	-	0.122±0.006	0.160±0.019	0.172±0.034
	NMAE	0.711±0.081	0.355±0.048	0.496±0.019	1.092±0.019	0.136±0.098	-	0.149±0.034	0.182±0.010	0.216±0.014
Trans MAF	CRPS	0.688±0.043	0.355±0.043	0.363±0.053	0.327±0.033	-	-	0.113±0.004	0.148±0.017	0.155±0.018
	NMAE	0.822±0.034	0.475±0.029	0.455±0.025	0.412±0.020	-	-	0.148±0.040	0.191±0.006	0.183±0.019
TimeGrad	CRPS	0.621±0.037	0.470±0.054	0.523±0.027	0.445±0.016	0.108±0.003	0.220±0.002	0.113±0.011	0.099±0.015	0.295±0.083
	NMAE	0.793±0.034	0.561±0.044	0.672±0.015	0.550±0.018	0.134±0.004	0.263±0.001	0.136±0.020	0.113±0.016	0.325±0.068
CSDI	CRPS	0.448±0.038	0.239±0.035	0.528±0.012	0.302±0.040	-	-	0.087±0.003	0.143±0.020	0.283±0.012
	NMAE	0.578±0.051	0.306±0.040	0.657±0.014	0.382±0.030	-	-	0.102±0.005	0.173±0.020	0.299±0.013
K^2VAE	CRPS	**0.294**±0.026	**0.221**±0.023	**0.314**±0.011	**0.280**±0.014	**0.057**±0.005	**0.200**±0.001	**0.084**±0.003	**0.069**±0.005	**0.142**±0.008
	NMAE	**0.373**±0.032	**0.275**±0.035	**0.396**±0.012	**0.278**±0.020	**0.117**±0.019	**0.248**±0.010	**0.099**±0.009	**0.084**±0.017	**0.167**±0.007

Due to the excessive time and memory consumption, some results are unavailable and denoted as -.

ensure a fair comparison. Details are shown in Table 1.

Baselines. We compare K^2VAE with 11 strong baselines, including 4 point forecasting models: FITS (Xu et al., 2024), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024), and Koopa (Liu et al., 2023), as well as 7 generative models: TSDiff (Kollovieh et al., 2023), D^3VAE (Li et al., 2022), GRU NVP, GRU MAF, Trans MAF (Rasul et al., 2021b), TimeGrad (Rasul et al., 2021a), and CSDI (Tashiro et al., 2021), in both short-term and long-term probabilistic forecasting scenarios. The point forecasting models are equipped with gaussian heads to predict the distributions. Detailed descriptions of these models can be found in Appendix C.2.

Evaluation Metrics. We use two commonly-used metrics CPRS (Continuous Ranked Probability Score) and NMAE (Normalized Mean Absolute Error) to evaluate the probabilistic forecasts. Detailed descriptions of these metrics can be found in Appendix C.3.

4.2. Main Results

Comprehensive probabilistic forecasting results are listed in Table 2 and Table 3 with the best in red and the second in blue. We have the following observations:

1) K^2VAE outperforms state-of-the-art baselines, showing notable improvements in predictive performance. In short-term scenarios, it achieves a *7.3%* reduction in CRPS and *14.5%* reduction in NMAE compared to the second-best baseline, CSDI. In long-term scenarios, it surpasses PatchTST with improvements of *20.9%* and *19.9%*.

2) K^2VAE shows significant advantage on nonstationary time seris datasets such as Exchange-S and Exchange-L. There exists distribution drift phenomenon in these datasets, which causes non-linearity and hinders the prediction and uncertainty modeling. Though diffusion-based and flow-based models are theoretically capable of fitting any distributions, they struggle to construct explict probability transition paths to reach such complex destinations. While K^2VAE simplifys this difficulty through modeling the time series in a linear dynamical system, where the uncertainty is more explicit and easier to be modeled.

3) K^2VAE also shows stable and strong performance with respect to the varying forecasting horizons–see Table 9 and Table 10 in Appendix C.5. The performance of most baselines drops significantly as the forecasting horizon extends, while K^2VAE maintains superiority. One potential reason is that K^2VAE utilizes the KoopmanNet to effectively handle the inherent nonlinearity in long-term forecasting. Another reason is that the KalmanNet mitigates the error accumulation by integrating diverse information.

4.3. Ablation Studies

4.3.1. Variants of Koopman Operator

We design the local Koopman Operator $\mathcal{K}_{loc}$ obtained by one-step eDMD and a learnable part $\mathcal{K}_{glo}$. As one-step

eDMD estimates the $\mathcal{K}_{loc}$ through matrix calculation, which relys on the local quality of the space of measurement function. Once the initilization leads to an ill-conditioned topological structure, the one-step eDMD suffers from the numerical calculation error or guides the model to converge in a wrong direction, which often occurs in long-term probabilistic forecasting scenarios and hinders the performance. To enhance the robustness, we adopt a global learnable part $\mathcal{K}_{glo}$ to mitigate this phenomenon while capturing the global-shared dynamics. As shown in Table 4, mixed Koopman Operator $\mathcal{K} = \mathcal{K}_{loc} + \mathcal{K}_{glo}$ demonstrates better performance on both short- and long-term tasks while providing roubustness to avoid calculation error if $\mathcal{K}_{loc}$ fails. Complete experimental results are provided in Table 12 in Appendix C.6.

Table 4. Comparison on different Koopman Operators. Lower values indicate better performance. **Red**: the best. L: the forecasting horizon. The full results can be found in Table 12 in Appendix C.6.

Koopman Operator	Metrics	Electricity-L (L = 720)	ETTm1-L (L = 720)	Exchange-S (L = 30)	Solar-S (L = 24)
$\mathcal{K}_{loc}$	CRPS	-	-	0.012±0.002	0.450±0.012
	NMAE	-	-	0.014±0.002	0.566±0.015
$\mathcal{K}_{glo}$	CRPS	0.065±0.007	0.311±0.024	0.011±0.001	0.374±0.004
	NMAE	0.130±0.024	0.395±0.027	0.013±0.001	0.488±0.008
$\mathcal{K}_{loc} + \mathcal{K}_{glo}$	CRPS	**0.057**±0.005	**0.294**±0.026	**0.009**±0.001	**0.367**±0.005
	NMAE	**0.117**±0.019	**0.373**±0.032	**0.009**±0.001	**0.480**±0.008

Due to the numerical instability, some results are unavailable and denoted as -.

4.3.2. Connections in KalmanNet

We adopt an Integrator to assist the KalmanNet for faster convergence, which potentially helps tune the topological structure of the space of measurement function into the linear dynamical system. Specifically, the Integrator integrates the non-linear residual into the control input of KalmanNet, which is proved not to affect the prior of Koopman Theory in Section 3.2. We also make a skip connection between Integrator and the KalmanNet for the final prediction, this constraints Integrator predicting residuals from residuals. Since the space of the measurement space is optimized to converge to a linear system, the Integrator serves as an assistant and gradually stop helping the model. We showcase the different variants in Table 5, to which only some of the features mentioned above are applied.

We observe that our adopted Mixed variant outperforms others, because it provides gains for the KalmanNet, which integrates non-linear information for adaption, and provides constraints for the Integrator, which makes full use of it without destroying the assumptions of Koopman Theory. The variant "w/o skip connection" completely depends on the linear fitting ability of the KalmanNet, which is hard to disentangle the nonlinear components in the early stages of training, thus potentially hindering the modeling of process uncertainty. On the other hand, the variant "w/o control input" gives too little support for KalmanNet to adaptively refine the prediction and process uncertainty. Complete experimental results are provided in Table 13 in Appendix C.5.

Table 5. Comparison on different connections in KalmanNet. Lower values indicate better performance. **Red**: the best. L: the forecasting horizon. The full results can be found in Table 13 in Appendix C.6.

Connections in KalmanNet	Metrics	Electricity-L (L = 720)	ETTm1-L (L = 720)	Exchange-S (L = 30)	Solar-S (L = 24)
w/o Integrator	CRPS	0.082±0.011	0.359±0.024	0.015±0.002	0.398±0.005
	NMAE	0.188±0.028	0.442±0.029	0.022±0.001	0.531±0.010
w/o skip connection	CRPS	0.063±0.007	0.315±0.016	0.011±0.001	0.388±0.006
	NMAE	0.131±0.015	0.402±0.030	0.013±0.004	0.511±0.008
w/o control input	CRPS	0.069±0.005	0.322±0.017	0.013±0.006	0.423±0.005
	NMAE	0.142±0.018	0.418±0.022	0.017±0.003	0.560±0.009
Mixed	CRPS	**0.057**±0.005	**0.294**±0.026	**0.009**±0.001	**0.367**±0.005
	NMAE	**0.117**±0.019	**0.373**±0.032	**0.009**±0.001	**0.480**±0.008

4.3.3. Ablations of KoopmanNet & KalmanNet

As the most important modules, the KoopmanNet and KalmanNet jointly contribute to state-of-the-art performance of K^2VAE. To evaluate their impactment, we conduct ablation studies and the results are shown in Table 6.

It is observed that both the KoopmanNet and KalmanNet show indispensability in probabilistic forecasting. Since the KoopmanNet ensures the linearization of the modeling, it shows greater impact in forecasting performance. Another reason is that KalmanNet does not excell at non-linear modeling because it is based on the linear kalman filter.

Table 6. Ablations on KoopmanNet and KalmanNet. Lower values indicate better performance. **Red**: the best. L: the forecasting horizon. The full results can be found in Table 14 in Appendix C.6.

Variants	Metrics	Electricity-L (L = 720)	ETTm1-L (L = 720)	Exchange-S (L = 30)	Solar-S (L = 24)
w/o KoopmanNet	CRPS	0.074±0.009	0.443±0.034	0.014±0.002	0.385±0.008
	NMAE	0.162±0.015	0.601±0.058	0.016±0.001	0.528±0.014
w/o KalmanNet	CRPS	0.089±0.011	0.398±0.038	0.011±0.001	0.375±0.005
	NMAE	0.192±0.023	0.539±0.044	0.012±0.001	0.499±0.009
K^2VAE	CRPS	**0.057**±0.005	**0.294**±0.026	**0.009**±0.001	**0.367**±0.005
	NMAE	**0.117**±0.019	**0.373**±0.032	**0.009**±0.001	**0.480**±0.008

4.4. Model Efficiency

We evaluate the model efficiency from three aspects: probabilistic forecasting performance (CRPS), inference time (sec/sample), and max gpu memory (GB). Figure 4 showcases a common scenario on Electricity-L (96-96), which reflects the overall relative relationships on above-mentioned three aspects. K^2VAE achieves best forecasting performance while occupying the minimum gpu memory and having the fastest inference speed. One reason is that K^2VAE applies KoopmanNet and KalmanNet, composed of several lightweight MLPs or linear layers, to efficiently build the variational distribution as process uncertainty, thus enhancing generation capability of K^2VAE. Another reason

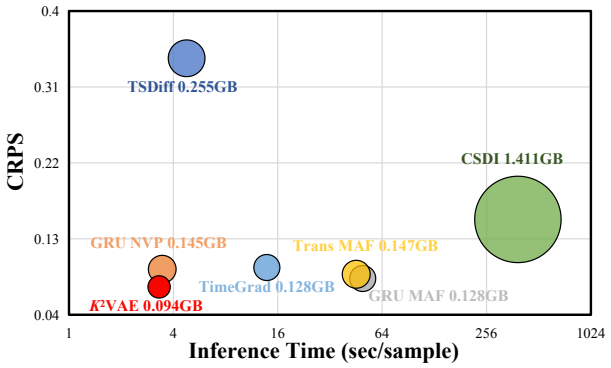

Figure 4. Model efficiency comparison. All the statistical data is obtained on the Electricity-L ($T = L = 96$). Sample-wise inference time and max gpu memory is obtained with batch size equals 1. Lower values of CRPS indicate better performance.

is that K^2VAE utilizes the VAE architecture and obeys the one-step-generation paradigm, while diffusion-based or flow-based models have longer probabilistic transition paths, which produces more intermediate results and consumes longer duration. More evidence of model efficiency is provided in Table 11 in Appendix C.5.

5. Related Works

5.1. Probabilistic Time Series Forecasting

Probabilistic forecasting aims to provide the predictive distribution of the target variable. With the rapid development of deep learning, new methods are continually emerging. DeepAR (Salinas et al., 2020) uses recurrent neural networks (RNNs) to model the transitions of hidden states and generates a Gaussian distribution for predictions. Following the autoregressive paradigm, DeepState (Rangapuram et al., 2018) and DSSMF (Li et al., 2019) combine state space models with deep learning to improve forecasting accuracy. MANF (Feng et al., 2024) and ProTran (Tang & Matteson, 2021) introduced attention-based methods that enhance the model's ability to capture long-range dependencies, further improving forecasting accuracy. Diffusion models, such as those proposed by TimeGrad (Rasul et al., 2021a), TSDiff (Kollovieh et al., 2023), and CSDI (Tashiro et al., 2021), approach the forecasting task as a denoising process, excelling in handling high-dimensional data. Another approach involves using more complex distribution forms, such as normalizing flows (Rasul et al., 2021b), to further enhance forecasting performance. Compared with RNN-based or State Space models, K^2VAE also autoregressively models the time series in a linear dynamical system, but mitigates the error accumulation through KalmanNet. Compared with generative diffusion-based or flow-based models, K^2VAE adopts the VAE structure and follows single-step-generation principle, achieves faster inference speed, lower memory occupation, and better performance.

5.2. VAE for Time Series

Variational Autoencoders (VAEs) (Kingma & Welling, 2014) have found wide applicability across various time series tasks. In time series generation, VAEs synthesize time series by encoding the data into a lower-dimensional latent space and then decoding it to recreate similar sequences, which helps preserve the statistical properties of the original data, making VAEs valuable for data augmentation (Li et al., 2023a; Desai et al., 2021). In time series imputation, VAEs recover missing values by learning the underlying latent structure of the data (Boquet et al., 2019; Li et al., 2021). By capturing temporal dependencies and relationships, they help restore incomplete time series with high accuracy. In time series anomaly detection, VAEs are used to learn the expected patterns within time series data and flag deviations that indicate anomalous behavior (Huang et al., 2022a; Wang et al., 2024f). In time series forecasting, TimeVAE (Desai et al., 2021) and D^3VAE (Li et al., 2022) are tailored for short-term probabilistic foercasting tasks. Koopa (Liu et al., 2023), as a strong baseline based on Koopman Theory, is tailored for long-term deterministic forecasting by adopting multi-scale MLP structures, which also falls short in probabilistic forecasting. Compared to these methods, K^2VAE is tailored for LPTSF, which considers the inherent nonlinearity of time series through a KoopmanNet and tackles the error accumulation through a KalmanNet, thus enhancing the ability to predict long-term future distributions. This facilitates better decision-making in dynamic and uncertain environments.

6. Conclusion

In this work, we propose a VAE-based probabilistic forecasting model called K^2VAE to solve PTSF. By leveraging the KoopmanNet, K^2VAE transforms nonlinear time series into a linear dynamical system, which allows for a more effective representation of state transitions and the inherent process uncertainties. Furthermore, the KalmanNet provides a solution to model the uncertainty in the linear dynamical system, mitigating the error accumuation in long-term forecasting tasks. Through comprehensive experiments, we demonstrate that K^2VAE not only outperforms existing state-of-the-art methods in both short- and long-term probabilistic forecasting tasks, but also achieves fascinating model efficiency.

In the future, we hope to continuously study the one-step generation paradigm in time series probabilistic modeling to further improve the model performance and efficiency. Another primary direction is to pionner the exploration of foundation probabilistic time series forecasting models, which can work effectively in zero-shot scenarios.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (62472174, 62372179). Bin Yang is the corresponding author of the work.

References

Boquet, G., Vicario, J. L., Morell, A., and Serrano, J. Missing data in traffic estimation: A variational autoencoder imputation method. In *ICASSP*, pp. 2882–2886, 2019.

Box, G. E. and Pierce, D. A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332):1509–1526, 1970.

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Campos, D., Kieu, T., Guo, C., Huang, F., Zheng, K., Yang, B., and Jensen, C. S. Unsupervised time series outlier detection with diversity-driven convolutional ensembles. *Proc. VLDB Endow.*, 15(3):611–623, 2022.

Campos, D., Zhang, M., Yang, B., Kieu, T., Guo, C., and Jensen, C. S. Lightts: Lightweight time series classification with adaptive ensemble distillation. *Proc. ACM Manag. Data*, 1(2):171:1–171:27, 2023.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *SIGKDD*, pp. 785–794, 2016.

Chen, Y., Huang, X., Zhang, Q., Li, W., Zhu, M., Yan, Q., Li, S., Chen, H., Hu, H., Yang, J., et al. Gim: A million-scale benchmark for generative image manipulation detection and localization. *arXiv preprint arXiv:2406.16531*, 2024.

Chen, Y., Huang, S., Cheng, Y., Chen, P., Rao, Z., Shu, Y., Yang, B., Pan, L., and Guo, C. AimTS: Augmented series and image contrastive learning for time series classification. In *ICDE*, 2025.

Chen, Z., Ding, L., Chu, Z., Qi, Y., Huang, J., and Wang, H. Monotonic neural ordinary differential equation: Time-series forecasting for cumulative data. In *CIKM*, pp. 4523–4529, 2023.

Cirstea, R., Guo, C., Yang, B., Kieu, T., Dong, X., and Pan, S. Triformer: Triangular, variable-specific attentions for long sequence multivariate time series forecasting. In *IJCAI*, pp. 1994–2001, 2022a.

Cirstea, R.-G., Yang, B., Guo, C., Kieu, T., and Pan, S. Towards spatio-temporal aware traffic time series forecasting. In *ICDE*, pp. 2900–2913, 2022b.

Cui, K., Liu, S., Feng, W., Deng, X., Gao, L., Cheng, M., Lu, H., and Yang, L. T. Correlation-aware cross-modal attention network for fashion compatibility modeling in ugc systems. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024a.

Cui, K., Tang, W., Zhu, R., Wang, M., Larsen, G. D., Pauca, V. P., Alqahtani, S., Yang, F., Segurado, D., Fine, P., et al. Real-time localization and bimodal point pattern analysis of palms using uav imagery. *arXiv preprint arXiv:2410.11124*, 2024b.

Cui, K., Zhu, R., Wang, M., Tang, W., Larsen, G. D., Pauca, V. P., Alqahtani, S., Yang, F., Segurado, D., Lutz, D., et al. Detection and geographic localization of natural objects in the wild: A case study on palms. *arXiv preprint arXiv:2502.13023*, 2025.

Dai, T., Wu, B., Liu, P., Li, N., Bao, J., Jiang, Y., and Xia, S.-T. Periodicity decoupling framework for long-term series forecasting. In *ICLR*, 2024.

Desai, A., Freeman, C., Wang, Z., and Beaver, I. Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

Fang, Z., Pan, L., Chen, L., Du, Y., and Gao, Y. MDTP: A multi-source deep traffic prediction framework over spatio-temporal trajectory data. *Proc. VLDB Endow.*, 14 (8):1289–1297, 2021.

Feng, S., Miao, C., Xu, K., Wu, J., Wu, P., Zhang, Y., and Zhao, P. Multi-scale attention flow for probabilistic time series forecasting. *IEEE Trans. Knowl. Data Eng.*, 36(5): 2056–2068, 2024.

Gao, H., Shen, W., Qiu, X., Xu, R., Yang, B., and Hu, J. Ssdts: Exploring the potential of linear state space models for diffusion models in time series imputation. In *SIGKDD*, 2025.

Godahewa, R., Bergmeir, C., Webb, G. I., Hyndman, R. J., and Montero-Manso, P. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

Guo, C., Yang, B., Andersen, O., Jensen, C. S., and Torp, K. Ecomark 2.0: empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data. *GeoInformatica*, 19:567–599, 2015.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

Hu, S., Zhao, K., Qiu, X., Shu, Y., Hu, J., Yang, B., and Guo, C. Multirc: Joint learning for time series anomaly prediction and detection with multi-scale reconstructive contrast. *arXiv preprint arXiv:2410.15997*, 2024.

Hu, Y., Liu, P., Zhu, P., Cheng, D., and Dai, T. Adaptive multi-scale decomposition framework for time series forecasting. In *AAAI*, 2025a.

Hu, Y., Zhang, G., Liu, P., Lan, D., Li, N., Cheng, D., Dai, T., Xia, S.-T., and Pan, S. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. *ICML*, 2025b.

Huang, Q., Shen, L., Zhang, R., Ding, S., Wang, B., Zhou, Z., and Wang, Y. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. In *NeurIPS*, pp. 46885–46902, 2023.

Huang, T., Chen, P., and Li, R. A semi-supervised vae based active anomaly detection framework in multivariate time series for online systems. In *WWW*, pp. 1797–1806, 2022a.

Huang, X., Yang, Y., Wang, Y., Wang, C., Zhang, Z., Xu, J., Chen, L., and Vazirgiannis, M. Dgraph: A large-scale financial dataset for graph anomaly detection. In *NeurIPS*, pp. 22765–22777, 2022b.

Hyndman, R., Koehler, A. B., Ord, J. K., and Snyder, R. D. *Forecasting with exponential smoothing: the state space approach*. 2008.

Jing, P., Cui, K., Guan, W., Nie, L., and Su, Y. Category-aware multimodal attention network for fashion compatibility modeling. *IEEE Transactions on Multimedia*, 25: 9120–9131, 2023.

Jing, P., Cui, K., Zhang, J., Li, Y., and Su, Y. Multimodal high-order relationship inference network for fashion compatibility modeling in internet of multimedia things. *IEEE Internet of Things Journal*, 11(1):353–365, 2024.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.

Kollovieh, M., Ansari, A. F., Bohlke-Schneider, M., Zschiegner, J., Wang, H., and Wang, Y. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. In *NeurIPS*, 2023.

Koopman, B. O. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

Lan, Y. and Mezić, I. Linearization in the large of nonlinear systems and koopman operator spectrum. *Physica D: Nonlinear Phenomena*, 242(1):42–53, 2013.

Li, H., Yu, S., and Principe, J. Causal recurrent variational autoencoder for medical time series generation. In *AAAI*, volume 37, pp. 8562–8570, 2023a.

Li, J., Ren, W., and Han, M. Variational auto-encoders based on the shift correction for imputation of specific missing in multivariate time series. *Measurement*, 186: 110055, 2021.

Li, L., Yan, J., Yang, X., and Jin, Y. Learning interpretable deep state space model for probabilistic time series forecasting. In *IJCAI*, pp. 2901–2908, 2019.

Li, Y., Lu, X., Wang, Y., and Dou, D. Generative time series forecasting with diffusion, denoise, and disentanglement. *Advances in Neural Information Processing Systems*, 35: 23009–23022, 2022.

Li, Y., Wang, H., Li, Z., Wang, S., Dev, S., and Zuo, G. Daanet: Dual attention aggregating network for salient object detection. In *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1–7, 2023b.

Li, Y., Wang, H., Xu, J., Ma, Z., Wu, P., Wang, S., and Dev, S. CP2M: Clustered-Patch-Mixed Mosaic Augmentation for Aerial Image Segmentation. *arXiv preprint arXiv:2501.15389*, 2025a.

Li, Y., Wang, H., Xu, J., Wu, P., Xiao, Y., Wang, S., and Dev, S. DDUNet: Dual Dynamic U-Net for Highly-Efficient Cloud Segmentation. *arXiv preprint arXiv:2501.15385*, 2025b.

Li, Z., Qiu, X., Chen, P., Wang, Y., Cheng, H., Shu, Y., Hu, J., Guo, C., Zhou, A., Jensen, C. S., and Yang, B. TSMF-Bench: Comprehensive and unified benchmarking of foundation models for time series forecasting. In *SIGKDD*, 2025c.

Lin, S., Lin, W., Wu, W., Zhao, F., Mo, R., and Zhang, H. Segrnn: Segment recurrent neural network for long-term time series forecasting. *arXiv preprint arXiv:2308.11200*, 2023.

Lin, S., Lin, W., Wu, K., Wang, S., Xu, M., and Wang, J. Z. Cocv: A compression algorithm for time-series data with continuous constant values in iot-based monitoring systems. *Internet of Things*, 25:101049, 2024a.

Lin, S., Lin, W., Wu, W., Chen, H., and Yang, J. SparseTSF: Modeling long-term time series forecasting with 1k parameters. In *ICML*, pp. 30211–30226, 2024b.

Lin, S., Lin, W., Xinyi, H., Wu, W., Mo, R., and Zhong, H. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. In *NeurIPS*, 2024c.

Lin, S., Lin, W., Zhao, F., and Chen, H. Benchmarking and revisiting time series forecasting methods in cloud workload prediction. *Cluster Computing*, 28(1):71, 2025.

Liu, C., Xu, Q., Miao, H., Yang, S., Zhang, L., Long, C., Li, Z., and Zhao, R. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *AAAI*, volume 39, pp. 18780–18788, 2025a.

Liu, P., Guo, H., Dai, T., Li, N., Bao, J., Ren, X., Jiang, Y., and Xia, S.-T. Calf: Aligning llms for time series forecasting via cross-modal fine-tuning. *AAAI*, 39(18): 18915–18923, 2025b.

Liu, P., Wu, B., Hu, Y., Li, N., Dai, T., Bao, J., and Xia, S.-T. Timebridge: Non-stationarity matters for long-term time series forecasting. In *ICML*, 2025c.

Liu, Q. and Paparrizos, J. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *NeurIPS*, 2024.

Liu, Y., Li, C., Wang, J., and Long, M. Koopa: Learning non-stationary time series dynamics with koopman predictors. *Advances in neural information processing systems*, 36: 12271–12290, 2023.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. itransformer: Inverted transformers are effective for time series forecasting. In *ICLR*, 2024.

Matheson, J. E. and Winkler, R. L. Scoring rules for continuous probability distributions. *Management science*, 22 (10):1087–1096, 1976.

Miao, H., Liu, Z., Zhao, Y., Guo, C., Yang, B., Zheng, K., and Jensen, C. S. Less is more: Efficient time series dataset condensation via two-fold modal matching. *PVLDB*, 18(2):226–238, 2024a.

Miao, H., Zhao, Y., Guo, C., Yang, B., Zheng, K., Huang, F., Xie, J., and Jensen, C. S. A unified replay-based continuous learning framework for spatio-temporal prediction on streaming data. In *ICDE*, pp. 1050–1062, 2024b.

Miao, H., Xu, R., Zhao, Y., Wang, S., Wang, J., Yu, P. S., and Jensen, C. S. A parameter-efficient federated framework for streaming time series anomaly detection via lightweight adaptation. *TMC*, 2025.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

Pan, Z., Sharma, A., Hu, J. Y.-C., Liu, Z., Li, A., Liu, H., Huang, M., and Geng, T. Ising-traffic: Using ising machine learning to predict traffic congestion under uncertainty. In *AAAI*, volume 37, pp. 9354–9363, 2023a.

Pan, Z., Wang, Y., Zhang, Y., Yang, S. B., Cheng, Y., Chen, P., Guo, C., Wen, Q., Tian, X., Dou, Y., et al. Magicscaler: Uncertainty-aware, predictive autoscaling. In *Proc. VLDB Endow.*, 2023b.

Pan, Z., Wang, Y., Zhang, Y., Yang, S. B., Cheng, Y., Chen, P., Guo, C., Wen, Q., Tian, X., Dou, Y., et al. Magicscaler: Uncertainty-aware, predictive autoscaling. *Proc. VLDB Endow.*, 16(12):3808–3821, 2023c.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., and Carin, L. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems*, 29, 2016.

Qiu, X., Hu, J., Zhou, L., Wu, X., Du, J., Zhang, B., Guo, C., Zhou, A., Jensen, C. S., Sheng, Z., and Yang, B. TFB: towards comprehensive and fair benchmarking of time series forecasting methods. *Proc. VLDB Endow.*, 17(9): 2363–2377, 2024.

Qiu, X., Cheng, H., Wu, X., Hu, J., and Guo, C. A comprehensive survey of deep learning for multivariate time series forecasting: A channel strategy perspective. *arXiv preprint arXiv:2502.10721*, 2025a.

Qiu, X., Li, X., Pang, R., Pan, Z., Wu, X., Yang, L., Hu, J., Shu, Y., Lu, X., Yang, C., Guo, C., Zhou, A., Jensen, C. S., and Yang, B. Easytime: Time series forecasting made easy. In *ICDE*, 2025b.

Qiu, X., Li, Z., Qiu, W., Hu, S., Zhou, L., Wu, X., Li, Z., Guo, C., Zhou, A., Sheng, Z., Hu, J., Jensen, C. S., and Yang, B. TAB: Unified benchmarking of time series anomaly detection methods. In *Proc. VLDB Endow.*, 2025c.

Qiu, X., Wu, X., Lin, Y., Guo, C., Hu, J., and Yang, B. DUET: Dual clustering enhanced multivariate time series forecasting. In *SIGKDD*, pp. 1185–1196, 2025d.

Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., and Januschowski, T. Deep state space models for time series forecasting. In *NeurIPS*, pp. 7796–7805, 2018.

Rasul, K., Seward, C., Schuster, I., and Vollgraf, R. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *ICML*, volume 139, pp. 8857–8868, 2021a.

Rasul, K., Sheikh, A., Schuster, I., Bergmann, U. M., and Vollgraf, R. Multivariate probabilistic time series forecasting via conditioned normalizing flows. In *ICLR*, 2021b.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656: 5–28, 2010.

Sezer, O. B., Gudelek, M. U., and Özbayoglu, A. M. Financial time series forecasting with deep learning : A systematic literature review: 2005-2019. *Appl. Soft Comput.*, 90:106181, 2020.

Simon, D. Kalman filtering. *Embedded systems programming*, 14(6):72–79, 2001.

Sun, Y., Xie, Z., Wang, H., Huang, X., and Hu, Q. Solar wind speed prediction via graph attention network. *Space Weather*, 20(7), 2022.

Tang, B. and Matteson, D. S. Probabilistic transformer for time series analysis. In *NeurIPS*, pp. 23592–23608, 2021.

Tashiro, Y., Song, J., Song, Y., and Ermon, S. CSDI: conditional score-based diffusion models for probabilistic time series imputation. In *NeurIPS*, pp. 24804–24816, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NeurIPS*, pp. 5998–6008, 2017.

Wang, C., Zhuang, Z., Qi, Q., Wang, J., Wang, X., Sun, H., and Liao, J. Drift doesn't matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection. In *NeurIPS*, pp. 10758–10774, 2023.

Wang, H., Chen, Z., Liu, Z., Li, H., Yang, D., Liu, X., and Li, H. Entire space counterfactual learning for reliable content recommendations. *IEEE Trans. Autom. Sci. Eng.*, pp. 1–12, 2024a.

Wang, H., Chen, Z., Liu, Z., Li, H., Yang, D., Liu, X., and Li, H. Entire space counterfactual learning for reliable content recommendations. *IEEE Transactions on Information Forensics and Security*, 2024b.

Wang, H., Chen, Z., Liu, Z., Pan, L., Xu, H., Liao, Y., Li, H., and Liu, X. Spot-i: Similarity preserved optimal transport for industrial iot data imputation. *IEEE Transactions on Industrial Informatics*, 2024c.

Wang, H., Wang, Z., Niu, Y., Liu, Z., Li, H., Liao, Y., Huang, Y., and Liu, X. An accurate and interpretable framework for trustworthy process monitoring. *IEEE Trans. Artif. Intell.*, 5(5):2241–2252, 2024d.

Wang, H., Chen, Z., Zhang, H., Li, Z., Pan, L., Li, H., and Gong, M. Debiased recommendation via wasserstein causal balancing. *ACM Transactions on Information Systems*, 2025a.

Wang, J., Zhang, Z., He, Y., Song, Y., Shi, T., Li, Y., Xu, H., Wu, K., Qian, G., Chen, Q., et al. Enhancing code llms with reinforcement learning in code generation. *arXiv preprint arXiv:2412.20367*, 2024e.

Wang, Y., Qiu, Y., Shu, Y., Rao, Z., Pan, L., Yang, B., and Chenjuan, G. Lightgts: A lightweight general time series forecasting model. In *ICML*, 2025b.

Wang, Y., Qiu, Y., Zhao, k., Shu, Y., Rao, Z., Pan, L., Yang, B., and Chenjuan, G. Towards a general time series forecasting model with unified representation and adaptive transfer. In *ICML*, 2025c.

Wang, Z., Pei, C., Ma, M., Wang, X., Li, Z., Pei, D., Rajmohan, S., Zhang, D., Lin, Q., Zhang, H., et al. Revisiting vae for unsupervised time series anomaly detection: A frequency perspective. In *WWW*, pp. 3096–3105, 2024f.

Welch, G. An introduction to the kalman filter. 1995.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *ICLR*, 2023a.

Wu, W., Qiu, X., Song, S., Huang, X., Ma, F., and Xiao, J. Prompt categories cluster for weakly supervised semantic segmentation. *arXiv preprint arXiv:2412.13823*, 2024a.

Wu, W., Dai, T., Chen, Z., Huang, X., Ma, F., and Xiao, J. Generative prompt controlled diffusion for weakly supervised semantic segmentation. *Neurocomputing*, pp. 130103, 2025a.

Wu, W., Song, S., Qiu, X., Huang, X., Ma, F., and Xiao, J. Image fusion for cross-domain sequential recommendation. In *Companion Proceedings of the ACM Web Conference 2025*, 2025b.

Wu, X., Zhang, D., Zhang, M., Guo, C., Yang, B., and Jensen, C. S. AutoCTS+: Joint neural architecture and hyperparameter search for correlated time series forecasting. *Proc. ACM Manag. Data*, 1(1):97:1–97:26, 2023b.

Wu, X., Wu, X., Yang, B., Zhou, L., Guo, C., Qiu, X., Hu, J., Sheng, Z., and Jensen, C. S. AutoCTS++: zero-shot joint neural architecture and hyperparameter search for correlated time series forecasting. *The VLDB Journal*, 33 (5):1743–1770, 2024b.

Wu, X., Qiu, X., Li, Z., Wang, Y., Hu, J., Guo, C., Xiong, H., and Yang, B. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *ICLR*, 2025c.

Wu, X., Wu, X., Zhang, D., Zhang, M., Guo, C., Yang, B., and Jensen, C. S. Fully automated correlated time series forecasting in minutes. In *Proc. VLDB Endow.*, volume 18, pp. 144–157, 2025d.

Xu, Z., Zeng, A., and Xu, Q. FITS: modeling time series with 10k parameters. In *ICLR*, 2024.

Yang, S. B., Guo, C., Hu, J., Tang, J., and Yang, B. Unsupervised path representation learning with curriculum negative sampling. In *IJCAI*, pp. 3286–3292, 2021.

Yang, S. B., Guo, C., and Yang, B. Context-aware path ranking in road networks. *IEEE Trans. Knowl. Data Eng.*, 34(7):3153–3168, 2022.

Yao, Y., Li, D., Jie, H., Chen, L., Li, T., Chen, J., Wang, J., Li, F., and Gao, Y. Simplets: An efficient and universal model selection framework for time series forecasting. *Proc. VLDB Endow.*, 16(12):3741–3753, 2023.

Yao, Y., Jie, H., Chen, L., Li, T., Gao, Y., and Wen, S. Tsec: An efficient and effective framework for time series classification. In *ICDE*, pp. 1394–1406, 2024.

Yi, Q., He, Y., Wang, J., Song, X., Qian, S., Zhang, M., Sun, L., and Shi, T. Score: Story coherence and retrieval enhancement for ai narratives. *arXiv preprint arXiv:2503.23512*, 2025.

Yu, C., Wang, F., Shao, Z., Sun, T., Wu, L., and Xu, Y. Dsformer: A double sampling transformer for multivariate time series long-term prediction. In *CIKM*, pp. 3062–3072, 2023.

Yu, C., Wang, F., Shao, Z., Qian, T., Zhang, Z., Wei, W., and Xu, Y. Ginar: An end-to-end multivariate time series forecasting model suitable for variable missing. In *SIGKDD*, pp. 3989–4000, 2024a.

Yu, C., Wang, F., Shao, Z., Qian, T., Zhang, Z., Wei, W., An, Z., Wang, Q., and Xu, Y. Ginar+: A robust end-to-end framework for multivariate time series forecasting with missing values. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–14, 2025a.

Yu, X., Elazab, A., Ge, R., Jin, H., Jiang, X., Jia, G., Wu, Q., Shi, Q., and Wang, C. Ich-scnet: Intracerebral hemorrhage segmentation and prognosis classification network using clip-guided sam mechanism. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2795–2800, 2024b.

Yu, X., Li, X., Ge, R., Wu, S., Elazab, A., Zhu, J., Zhang, L., Jia, G., Xu, T., Wan, X., et al. Ichpro: Intracerebral hemorrhage prognosis classification via joint-attention fusion-based 3d cross-modal network. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2024c.

Yu, X., Elazab, A., Ge, R., Zhu, J., Zhang, L., Jia, G., Wu, Q., Wan, X., Li, L., and Wang, C. Ich-prnet: a cross-modal intracerebral haemorrhage prognostic prediction method using joint-attention interaction mechanism. *Neural Networks*, 184:107096, 2025b.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. Are transformers effective for time series forecasting? In *AAAI*, volume 37, pp. 11121–11128, 2023.

Zhang, J., Wen, X., Zhang, Z., Zheng, S., Li, J., and Bian, J. ProbTS: Benchmarking point and distributional forecasting across diverse prediction horizons. In *NeurIPS*, 2024a.

Zhang, Q. and Qi, Y. Can mllms guide weakly-supervised temporal action localization tasks? *arXiv preprint arXiv:2411.08466*, 2024.

Zhang, Q., Liu, X., Li, W., Chen, H., Liu, J., Hu, J., Xiong, Z., Yuan, C., and Wang, Y. Distilling semantic priors from sam to efficient image restoration models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25409–25419, 2024b.

Zhang, Q., Qi, Y., Tang, X., Fang, J., Lin, X., Zhang, K., and Yuan, C. Imdprompter: Adapting sam to image manipulation detection by cross-view automated prompt learning. *arXiv preprint arXiv:2502.02454*, 2025a.

Zhang, Q., Qi, Y., Tang, X., Yuan, R., Lin, X., Zhang, K., and Yuan, C. Rethinking pseudo-label guided learning for weakly supervised temporal action localization from the perspective of noise correction. *arXiv preprint arXiv:2501.11124*, 2025b.

Zhao, K., Guo, C., Cheng, Y., Han, P., Zhang, M., and Yang, B. Multiple time series forecasting with dynamic graph modeling. *Proc. VLDB Endow.*, 17(4):753–765, 2023.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, volume 35, pp. 11106–11115, 2021.

A. Theoretical Analyses

A.1. The Stability of KalmanNet

Since the proposed KalmanNet works in a data-driven manner, the floating-point operation error may cause the covariance matrix P losing positive definiteness, which often occurs in the step (21). To mitigate this, we utilize a numerically stable form for this step.

Theorem A.1. *The positive-definiteness of covariance matrix P_k during the update step $\mathrm{P}_k = (I - K_k H_k)\hat{\mathrm{P}}_k$ can be retained through a numerically stable form:*

$$\mathrm{P}_k = \frac{1}{2}(\mathrm{P}_k + \mathrm{P}_k^T), \tag{30}$$

$$\mathrm{P}_k^{dual} = (I - K_k H_k)\hat{\mathrm{P}}_k(I - K_k H_k)^T + K_k R_k K_k^T \tag{31}$$

Proof. The goal is to demonstrate the equivalence of the numerically stable form and original form: $\mathrm{P}_k^{dual} = \mathrm{P}_k$.

$$\begin{aligned}
\mathrm{P}_k^{dual} &= (I - K_k H_k)\hat{\mathrm{P}}_k(I - K_k H_k)^T + K_k R_k K_k^T, \\
&= (I - K_k H_k)\hat{\mathrm{P}}_k - (I - K_k H_k)\hat{\mathrm{P}}_k H_k^T K_k^T + K_k R_k K_k^T, \\
&= (I - K_k H_k)\hat{\mathrm{P}}_k - \hat{\mathrm{P}}_k H_k^T K_k^T + K_k H_k \hat{\mathrm{P}}_k H_k^T K_k^T + K_k R_k K_k^T, \\
&= (I - K_k H_k)\hat{\mathrm{P}}_k - \hat{\mathrm{P}}_k H_k^T K_k^T + K_k (H_k \hat{\mathrm{P}}_k H_k^T + R_k) K_k^T, \\
K_k &= \hat{\mathrm{P}}_k H_k^T (H_k \hat{\mathrm{P}}_k H_k^T + R)^{-1}, \\
\mathrm{P}_k^{dual} &= (I - K_k H_k)\hat{\mathrm{P}}_k - \hat{\mathrm{P}}_k H_k^T K_k^T + \hat{\mathrm{P}}_k H_k^T K_k^T, \\
&= (I - K_k H_k)\hat{\mathrm{P}}_k = \mathrm{P}_k,
\end{aligned}$$

P_k^{dual} is numerically equivalent to the original P_k. □

where (30) ensures the symmetry, (31) stabilizes the positive-definiteness by decomposing the formula into the sum of two positive definite terms, which better ensures positive definiteness during floating operation.

A.2. The Convergence of K^2VAE

Since K^2VAE models a linear dynamical system in the measurement space where the Koopman Operator serves as the state transition equation, we hope that the convergence state of the KalmanNet does not violate the assumptions of Koopman Theory. In K^2VAE, we meticulously design the KalmanNet by making it gradually converge to the Koopman Operator in the forecasting horizon.

Theorem A.2. *When $U \to \mathbf{0}$, the state transition equation of the KalmanNet gradually converges to the Koopman Operator.*

Proof. Under the assumptions of Koopman Theory, $u_k \to \mathbf{0}$ means the linear system constructed by Koopman Operator has little bias in the current measurement space, which leads to high performance in prediction. Meanwhile, the Predict and Update Steps of z_t are converted to:

$$\textit{Predict:} \quad \hat{z}_k = A z_{k-1} \tag{32}$$

$$\textit{Update:} \quad K_k = \hat{\mathrm{P}}_k H^T (H \hat{\mathrm{P}}_k H^T + R)^{-1}, \tag{33}$$

$$z_k = \hat{z}_k + K_k(\hat{x}_k^H - H\hat{z}_k) \tag{34}$$

In this basic case, the state transition equation obeys the basic assumptions of Koopman Theory and A can be treated as a "fine-tuned" Koopman Operator $\mathcal{K}$ which is enhanced by the Kalman gain and has stronger generalization ability.

We then consider the special case that KalmanNet fully relys on the observation $\hat{x}_k^H$ from the linear system constructed by Koopman Operator $\mathcal{K}$, thus $H \to I, A \to \mathbf{0}, R \to \mathbf{0}$, the Predict and Update Steps are converted to:

$$\textit{Predict:} \quad \hat{z}_k = \mathbf{0} \tag{35}$$

$$\textit{Update:} \quad z_k = \hat{x}_k^H \tag{36}$$

The system constructed by KalmanNet can be treated as $z_t = \mathcal{K} z_{t-1}$ equivalent to the original Koopman Operator. □

B. Related Works

B.1. Time Series Forecasting

Time series forecasting (TSF) predicts future observations based on historical observations. TSF methods are mainly categorized into three distinct approaches: (1) statistical learning-based methods, (2) machine learning-based methods, and (3) deep learning-based methods. Early TSF methods primarily rely on statistical learning approaches such as ARIMA (Box & Pierce, 1970), ETS (Hyndman et al., 2008), and VAR (Godahewa et al., 2021). With advancements in machine learning, methods like XGBoost (Chen & Guestrin, 2016), Random Forests (Breiman, 2001), and LightGBM (Ke et al., 2017) gain popularity for handling nonlinear patterns. However, these methods still require manual feature engineering and model design. Recently, deep learning has made impressive progress in natural language processing (Chen et al., 2024; Zhang & Qi, 2024; Wang et al., 2024e; Wu et al., 2024a; 2025a), computer vision (Zhang et al., 2025b; 2024b; Wu et al., 2025b; Cui et al., 2024b; Yu et al., 2024b; Li et al., 2025a;b; Yu et al., 2024c), multimodal (Zhang et al., 2025a; Cui et al., 2024a; Jing et al., 2023; 2024), and other aspects (Wang et al., 2025a; Yu et al., 2025b; Cui et al., 2025; Yi et al., 2025; Li et al., 2023b; Miao et al., 2024b; Wu et al., 2023b; Zhao et al., 2023; Chen et al., 2025). Studies have shown that learned features may perform better than human-designed features (Qiu et al., 2025b;a; Liu et al., 2025b; Yu et al., 2024a). Leveraging the representation learning of deep neural networks (DNNs), many deep learning-based methods emerge. TimesNet (Wu et al., 2023a) and SegRNN (Lin et al., 2023) model time series as vector sequences, using CNNs or RNNs to capture temporal dependencies. Additionally, Transformer architectures, including Informer (Zhou et al., 2021), Dsformer (Yu et al., 2023), TimeFilter (Hu et al., 2025b), TimeBridge (Liu et al., 2025c), PDF (Dai et al., 2024), Triformer (Cirstea et al., 2022a), PatchTST (Nie et al., 2023), ROSE (Wang et al., 2025c), LightGTS (Wang et al., 2025b), and MagicScaler (Pan et al., 2023b) capture complex relationships between time points more accurately, significantly improving forecasting performance. MLP-based methods, including DUET (Qiu et al., 2025d), AMD (Hu et al., 2025a), SparseTSF (Lin et al., 2024b), CycleNet (Lin et al., 2024c), NLinear (Zeng et al., 2023), and DLinear (Zeng et al., 2023), adopt simpler architectures with fewer parameters but still achieve highly competitive forecasting accuracy.

C. Experimental Details

C.1. Datasets

In order to comprehensively evaluate the performance of K^2VAE, we conduct experiments on 8 datasets of short-term forecasting and 9 datasets of long-term forecasting under the framework of ProbTS (Zhang et al., 2024a), a comprehensive benchmark used to evaluate probabilistic forecasting tasks. Specifically, we use the datasets ETTh1-S, ETTh2-S, ETTm1-S, ETTm2-S, Electricity-S, Solar-S, Traffic-S, and Exchange-S for short-term forecasting, of which the context length is equivalent to forecasting horizon with $T = L = 30$ for Exchange-S and $T = L = 24$ for the others. For long-term forecasting, we use the datasets ETTh1-L, ETTh2-L, ETTm1-L, ETTm2-L, Electricity-L, Traffic-L, Exchange-L, Weather-L, and ILI-L with prediction length $L \in \{24, 36, 48, 60\}$ for ILI-L and $L \in \{96, 192, 336, 720\}$ for the others. Note that we fix the context length of all the models with $T = 36$ for ILI-L and $T = 96$ for the others to ensure a fair comparison. *Please note that although datasets with the same prefix may appear similar, they are not necessarily the same. For example, Electricity-L and Electricity-S are not the same dataset, despite both having the prefix "Electricity." The datasets we use are all derived from the authoritative probabilistic forecasting benchmark, ProbTS. Furthermore, due to the differences in long-term and short-term tasks, the datasets used for long-term and short-term forecasting in ProbTS and K^2VAE are also different.* Table 7 lists statistics of the multivariate time series datasets.

C.2. Baselines

In the realm of probabilistic time series forecasting, numerous models have surfaced in recent years. Following the experimental setting in ProbTS, we compare K^2VAE with 11 strong baselines including 4 point forecasting models: FITS, PatchTST, iTransformer, Koopa, and 7 generative models: TSDiff, D^3VAE, GRU NVP, GRU MAF, Trans MAF, TimeGrad, CSDI on both short-term and long-term probabilistic forecasting scenarios. The specific code repositories for each of these models–see Table 8.

C.3. Evaluation Metrics

We use two commonly-used metrics NMAE (Normalized Mean Absolute Error) and CPRS (Continuous Ranked Probability Score) in ProbTS (Zhang et al., 2024a) to evaluate the probabilistic forecasts.

Table 7. Dataset Summary.

Horizon	Dataset	#var.	range	freq.	timesteps	Description
Long-term	ETTh1/h2-L	7	$\mathbb{R}^+$	H	17,420	Electricity transformer temperature per hour
	ETTm1/m2-L	7	$\mathbb{R}^+$	15min	69,680	Electricity transformer temperature every 15 min
	Electricity-L	321	$\mathbb{R}^+$	H	26,304	Electricity consumption (Kwh)
	Traffic-L	862	(0,1)	H	17,544	Road occupancy rates
	Exchange-L	8	$\mathbb{R}^+$	Busi. Day	7,588	Daily exchange rates of 8 countries
	ILI-L	7	(0,1)	W	966	Ratio of patients seen with influenza-like illness
	Weather-L	21	$\mathbb{R}^+$	10min	52,696	Local climatological data
Short-term	ETTh1/h2-S	7	$\mathbb{R}^+$	H	17,420	Electricity transformer temperature per hour
	ETTm1/m2-S	7	$\mathbb{R}^+$	15min	69,680	Electricity transformer temperature every 15 min
	Exchange-S	8	$\mathbb{R}^+$	Busi. Day	6,071	Daily exchange rates of 8 countries
	Solar-S	137	$\mathbb{R}^+$	H	7,009	Solar power production records
	Electricity-S	370	$\mathbb{R}^+$	H	5,833	Electricity consumption
	Traffic-S	963	(0,1)	H	4,001	Road occupancy rates

Table 8. Code repositories for baselines.

Baselines	Code repositories
Koopa	https://github.com/thuml/koopa
iTransformer	https://github.com/thuml/iTransformer
FITS	https://github.com/VEWOXIC/FITS
PatchTST	https://github.com/yuqinie98/PatchTST
TSDiff	https://github.com/amazon-science/unconditional-time-series-diffusion
D^3VAE	https://github.com/PaddlePaddle/PaddleSpatial/tree/main/research/D3VAE
GRU NVP	https://github.com/zalandoresearch/pytorch-ts
GRU MAF	https://github.com/zalandoresearch/pytorch-ts
Trans MAF	https://github.com/zalandoresearch/pytorch-ts
TimeGrad	https://github.com/yuqinie98/PatchTST
CSDI	https://github.com/ermongroup/CSDI
K^2VAE (ours)	https://github.com/decisionintelligence/K2VAE

Normalized Mean Absolute Error (NMAE) The Normalized Mean Absolute Error (NMAE) is a normalized version of the MAE, which is dimensionless and facilitates the comparability of the error magnitude across different datasets or scales. The mathematical representation of NMAE is given by:

$$\text{NMAE} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T}|x_t^k - \hat{x}_t^k|}{\sum_{k=1}^{K}\sum_{t=1}^{T}|x_t^k|}. \tag{37}$$

Continuous Ranked Probability Score (CRPS) The Continuous Ranked Probability Score (CRPS) (Matheson & Winkler, 1976) quantifies the agreement between a cumulative distribution function (CDF) F and an observation x, represented as:

$$\text{CRPS} = \int_{\mathbb{R}} (F(z) - \mathbb{I}\{x \leq z\})^2 dz, \tag{38}$$

where $\mathbb{I}\{x \leq z\}$ denotes the indicator function, equating to one if $x \leq z$ and zero otherwise.

Being a proper scoring function, CRPS reaches its minimum when the predictive distribution F coincides with the data distribution. When using the empirical CDF of F, denoted as $\hat{F}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq z\}$, where n represents the number of samples $X_i \sim F$, CRPS can be precisely calculated from the simulated samples of the conditional distribution $p_\theta(\mathbf{x}_t|\mathbf{h}_t)$. In our practice, 100 samples are employed to estimate the empirical CDF.

C.4. Implementation Details

For each method, we adhere to the hyper-parameter as specified in their original papers. Additionally, we perform hyper-parameter searches across multiple sets, with a limit of 8 sets. The optimal result is then selected from these evaluations, contributing to a comprehensive and unbiased assessment of each method's performance.

The *"Drop Last"* issue is reported by several researchers (Qiu et al., 2024; 2025c; Li et al., 2025c). That is, in some previous works evaluating the model on test set with drop-last=True setting may cause additional errors related to test batch size. In our experiment, to ensure fair comparison in the future, we set the drop last to False for all baselines to avoid this issue.

All experiments are conducted using PyTorch (Paszke et al., 2019) in Python 3.10 and execute on an NVIDIA Tesla-A800 GPU. The training process is guided by the $\mathcal{L}_{ELBO}$ and $\mathcal{L}_{Rec}$, employing the ADAM optimizer. Initially, the batch size is set to 32, with the option to reduce it by half (to a minimum of 8) in case of an Out-Of-Memory (OOM) situation. To ensure reproducibility and facilitate experimentation, datasets and code are available at: https://github.com/decisionintelligence/K2VAE.

C.5. Full Results

We provide all the main results of LPTSF in Table 9 and Table 10, covering all four horizons ($L \in \{96, 192, 336, 720\}$) on 9 real world datasets. The results show that K^2VAE achieves a comprehensive lead in long-term prediction tasks, not only outperforming generative models specialized for probabilistic prediction but also demonstrating significant advantages compared to long-term point-based prediction models.

We provide the complete results of ablation studies in Table 12–13. We compare the different variants under various tasks across different horizons, empirical results demonstrate that K^2VAE adopts the most appropriate design.

We also provide the complete efficiency analyses under different forecasting scenarios, which demonstrates that our proposed K^2VAE exhibits low memory overhead, fast inference speed, and high accuracy across various tasks. Compared to generative models such as those diffusion-based or flow-based models, K^2VAE is both more precise and lightweight.

Table 9. Results of CRPS (mean$_{\text{std}}$) on long-term forecasting scenarios, each containing five independent runs with different seeds. The context length is set to 36 for the ILI-L dataset and 96 for the others. Lower CRPS values indicate better predictions. The means and standard errors are based on 5 independent runs of retraining and evaluation. **Red**: the best, Blue: the 2nd best.

Dataset	Horizon	Koopa	iTransformer	FITS	PatchTST	GRU MAF	Trans MAF	TSDiff	CSDI	TimeGrad	GRU NVP	K^2VAE
ETTm1-L	96	0.285±0.018	0.301±0.033	0.267±0.023	0.261±0.051	0.295±0.055	0.313±0.045	0.344±0.050	0.236±0.006	0.522±0.105	0.383±0.053	**0.232**±0.010
	192	0.289±0.024	0.314±0.023	0.261±0.022	0.275±0.030	0.389±0.033	0.424±0.029	0.345±0.035	0.291±0.025	0.603±0.092	0.396±0.030	**0.259**±0.013
	336	0.286±0.035	0.311±0.029	0.275±0.030	0.285±0.028	0.429±0.021	0.481±0.019	0.462±0.043	0.322±0.033	0.601±0.028	0.486±0.032	**0.262**±0.030
	720	0.295±0.027	0.455±0.021	0.305±0.024	0.304±0.029	0.536±0.033	0.688±0.043	0.478±0.027	0.448±0.038	0.621±0.037	0.546±0.036	**0.294**±0.026
ETTm2-L	96	0.178±0.023	0.181±0.031	0.162±0.053	0.142±0.034	0.177±0.024	0.227±0.013	0.175±0.019	**0.115**±0.009	0.427±0.042	0.319±0.044	0.126±0.007
	192	0.185±0.014	0.190±0.010	0.185±0.053	0.172±0.023	0.411±0.026	0.253±0.037	0.255±0.029	**0.147**±0.008	0.424±0.061	0.326±0.025	0.148±0.009
	336	0.198±0.015	0.206±0.055	0.218±0.053	0.195±0.042	0.377±0.023	0.253±0.013	0.328±0.047	0.190±0.018	0.469±0.049	0.449±0.145	**0.164**±0.010
	720	0.233±0.025	0.311±0.024	0.449±0.034	0.229±0.036	0.272±0.029	0.355±0.043	0.344±0.046	0.239±0.035	0.470±0.054	0.561±0.273	**0.221**±0.023
ETTh1-L	96	0.307±0.033	0.292±0.032	0.294±0.023	0.312±0.036	0.293±0.037	0.333±0.045	0.395±0.052	0.437±0.018	0.455±0.046	0.379±0.030	**0.264**±0.020
	192	0.301±0.014	0.298±0.020	0.304±0.028	0.313±0.034	0.348±0.075	0.351±0.063	0.467±0.044	0.496±0.051	0.516±0.038	0.425±0.019	**0.290**±0.016
	336	0.312±0.019	0.327±0.043	0.318±0.023	0.319±0.035	0.377±0.026	0.371±0.031	0.450±0.027	0.454±0.025	0.512±0.026	0.458±0.054	**0.308**±0.021
	720	0.318±0.009	0.350±0.019	0.348±0.025	0.323±0.020	0.393±0.043	0.363±0.053	0.516±0.027	0.528±0.012	0.523±0.027	0.502±0.039	**0.314**±0.011
ETTh2-L	96	0.199±0.012	0.185±0.013	0.187±0.011	0.197±0.021	0.239±0.019	0.263±0.020	0.336±0.021	0.164±0.013	0.358±0.026	0.432±0.141	**0.162**±0.009
	192	0.198±0.022	0.199±0.019	0.195±0.022	0.204±0.055	0.313±0.034	0.273±0.024	0.265±0.043	0.226±0.018	0.457±0.081	0.625±0.170	**0.186**±0.018
	336	0.262±0.019	0.271±0.033	**0.246**±0.044	0.277±0.054	0.376±0.034	0.265±0.042	0.350±0.031	0.274±0.022	0.481±0.078	0.793±0.319	0.257±0.023
	720	0.293±0.026	0.542±0.015	0.314±0.022	0.304±0.018	0.990±0.023	0.327±0.033	0.406±0.056	0.302±0.040	0.445±0.016	0.539±0.090	**0.280**±0.014
Electricity-L	96	0.110±0.004	0.102±0.004	0.105±0.006	0.126±0.005	0.083±0.009	0.088±0.014	0.344±0.006	0.153±0.137	0.096±0.002	0.094±0.003	**0.073**±0.002
	192	0.109±0.011	0.104±0.014	0.112±0.104	0.123±0.032	0.093±0.024	0.097±0.009	0.345±0.006	0.200±0.094	0.100±0.004	0.097±0.002	**0.080**±0.004
	336	0.121±0.011	0.104±0.010	0.111±0.014	0.131±0.024	0.095±0.001	-	0.462±0.054	-	0.102±0.007	0.099±0.001	**0.054**±0.001
	720	0.113±0.018	0.109±0.044	0.115±0.024	0.127±0.015	0.106±0.007	-	0.478±0.005	-	0.108±0.003	0.114±0.013	**0.057**±0.005
Traffic-L	96	0.297±0.019	0.256±0.004	0.258±0.004	0.194±0.002	0.215±0.003	0.208±0.004	0.294±0.003	-	0.202±0.004	0.187±0.002	**0.086**±0.001
	192	0.308±0.009	0.250±0.002	0.275±0.003	0.198±0.004	-	-	0.306±0.004	-	0.208±0.003	0.192±0.001	**0.088**±0.002
	336	0.334±0.017	0.261±0.001	0.327±0.001	0.204±0.002	-	-	0.317±0.006	-	0.213±0.003	0.201±0.004	**0.195**±0.003
	720	0.358±0.022	0.284±0.004	0.374±0.004	0.214±0.001	-	-	0.391±0.002	-	0.220±0.002	0.211±0.004	**0.200**±0.001
Weather-L	96	0.132±0.008	0.131±0.011	0.210±0.013	0.131±0.007	0.139±0.008	0.105±0.011	0.104±0.020	**0.068**±0.008	0.130±0.017	0.116±0.013	0.080±0.007
	192	0.133±0.017	0.132±0.018	0.205±0.019	0.131±0.014	0.143±0.020	0.142±0.022	0.134±0.012	**0.068**±0.006	0.127±0.019	0.122±0.021	0.079±0.009
	336	0.136±0.021	0.132±0.010	0.221±0.005	0.137±0.008	0.129±0.012	0.133±0.014	0.137±0.010	0.083±0.002	0.130±0.006	0.128±0.011	**0.082**±0.010
	720	0.140±0.007	0.133±0.004	0.267±0.003	0.142±0.005	0.122±0.006	0.113±0.004	0.152±0.003	0.087±0.003	0.113±0.011	0.110±0.004	**0.084**±0.003
Exchange-L	96	0.063±0.006	0.061±0.003	0.048±0.004	0.063±0.006	0.026±0.010	0.028±0.002	0.079±0.007	**0.028**±0.003	0.068±0.003	0.071±0.006	0.031±0.002
	192	0.065±0.020	0.062±0.010	0.049±0.011	0.067±0.008	0.034±0.009	0.046±0.017	0.093±0.011	0.045±0.003	0.087±0.013	0.068±0.004	**0.032**±0.010
	336	0.072±0.008	0.067±0.008	0.052±0.013	0.071±0.017	0.058±0.023	**0.045**±0.010	0.081±0.007	0.060±0.004	0.074±0.009	0.072±0.002	0.048±0.004
	720	0.091±0.012	0.087±0.023	0.074±0.011	0.097±0.007	0.160±0.019	0.148±0.017	0.082±0.010	0.143±0.020	0.099±0.015	0.079±0.009	**0.069**±0.005
ILI-L	24	0.245±0.018	0.212±0.013	0.233±0.015	0.312±0.014	0.097±0.010	0.092±0.019	0.228±0.024	0.250±0.013	0.275±0.047	0.257±0.003	**0.087**±0.003
	36	0.214±0.008	0.182±0.016	0.217±0.023	0.241±0.021	0.117±0.017	0.115±0.011	0.235±0.010	0.285±0.010	0.272±0.057	0.281±0.004	**0.113**±0.005
	48	0.271±0.021	0.213±0.012	0.185±0.026	0.242±0.018	0.128±0.019	0.133±0.022	0.265±0.039	0.285±0.036	0.295±0.033	0.288±0.008	**0.124**±0.010
	60	0.228±0.022	0.222±0.020	0.211±0.011	0.233±0.019	0.172±0.034	0.155±0.018	0.263±0.022	0.283±0.012	0.295±0.083	0.307±0.005	**0.142**±0.008

Due to the excessive time and memory consumption, some results are unavailable in our implementation and denoted as -.

Table 10. Results of NMAE (mean$_{\text{std}}$) on long-term forecasting scenarios, each containing five independent runs with different seeds. The context length is set to 36 for the ILI-L dataset and 96 for the others. Lower NMAE values indicate better predictions. The means and standard errors are based on 5 independent runs of retraining and evaluation. **Red**: the best, Blue: the 2nd best.

Dataset	Horizon	Koopa	iTransformer	FITS	PatchTST	GRU MAF	Trans MAF	TSDiff	CSDI	TimeGrad	GRU NVP	K^2VAE
ETTm1-L	96	0.362±0.022	0.369±0.029	0.349±0.032	0.329±0.100	0.402±0.087	0.456±0.042	0.441±0.021	0.308±0.005	0.645±0.129	0.488±0.058	**0.284**±0.011
	192	0.365±0.032	0.384±0.041	0.341±0.032	0.338±0.022	0.476±0.046	0.553±0.012	0.441±0.019	0.377±0.026	0.748±0.084	0.514±0.042	**0.323**±0.020
	336	0.364±0.026	0.380±0.020	0.356±0.022	0.344±0.013	0.522±0.019	0.590±0.047	0.571±0.033	0.419±0.042	0.759±0.015	0.630±0.029	**0.330**±0.014
	720	0.377±0.037	0.490±0.038	0.406±0.072	0.382±0.066	0.711±0.081	0.822±0.034	0.622±0.045	0.578±0.051	0.793±0.034	0.707±0.050	**0.373**±0.032
ETTm2-L	96	0.225±0.039	0.221±0.039	0.210±0.040	0.216±0.035	0.212±0.082	0.279±0.031	0.224±0.033	0.146±0.012	0.525±0.047	0.413±0.059	**0.144**±0.011
	192	0.233±0.026	0.229±0.031	0.234±0.038	0.215±0.022	0.535±0.029	0.292±0.041	0.316±0.040	0.189±0.012	0.530±0.060	0.427±0.033	**0.170**±0.009
	336	0.267±0.023	0.245±0.049	0.276±0.019	0.234±0.024	0.407±0.043	0.309±0.032	0.397±0.051	0.248±0.024	0.566±0.047	0.580±0.169	**0.187**±0.021
	720	0.290±0.033	0.385±0.042	0.540±0.052	0.288±0.034	0.355±0.048	0.475±0.029	0.416±0.065	0.306±0.040	0.561±0.044	0.749±0.385	**0.275**±0.035
ETTh1-L	96	0.407±0.052	0.386±0.092	0.393±0.142	0.407±0.022	0.371±0.034	0.423±0.047	0.510±0.029	0.557±0.022	0.585±0.058	0.481±0.037	**0.336**±0.041
	192	0.396±0.022	0.388±0.041	0.406±0.079	0.405±0.088	0.430±0.022	0.451±0.012	0.596±0.056	0.625±0.065	0.680±0.058	0.531±0.018	**0.372**±0.023
	336	0.406±0.028	0.415±0.022	0.410±0.063	0.412±0.024	0.462±0.049	0.481±0.041	0.581±0.035	0.574±0.026	0.666±0.047	0.580±0.064	**0.394**±0.022
	720	0.412±0.008	0.449±0.022	0.468±0.012	0.428±0.024	0.496±0.019	0.455±0.025	0.657±0.017	0.657±0.014	0.672±0.015	0.643±0.046	**0.396**±0.012
ETTh2-L	96	0.249±0.015	0.234±0.011	0.243±0.009	0.247±0.028	0.292±0.012	0.345±0.042	0.421±0.033	0.214±0.018	0.448±0.031	0.548±0.158	**0.189**±0.010
	192	0.249±0.032	0.247±0.040	0.252±0.022	0.265±0.091	0.376±0.112	0.343±0.044	0.339±0.033	0.294±0.027	0.575±0.089	0.766±0.223	**0.213**±0.021
	336	0.274±0.027	0.297±0.029	0.291±0.032	0.314±0.045	0.454±0.057	0.333±0.078	0.427±0.041	0.353±0.028	0.606±0.095	0.942±0.408	**0.263**±0.039
	720	0.286±0.042	0.667±0.012	0.401±0.022	0.371±0.021	1.092±0.019	0.412±0.020	0.482±0.022	0.382±0.030	0.550±0.018	0.688±0.161	**0.278**±0.020
Electricity-L	96	0.146±0.015	0.134±0.002	0.137±0.002	0.168±0.012	0.108±0.009	0.114±0.010	0.441±0.013	0.203±0.189	0.119±0.003	0.118±0.003	**0.093**±0.002
	192	0.143±0.023	0.137±0.022	0.143±0.112	0.163±0.032	0.120±0.033	0.131±0.008	0.441±0.005	0.264±0.129	0.124±0.005	0.121±0.003	**0.102**±0.010
	336	0.151±0.017	0.136±0.002	0.139±0.002	0.168±0.010	0.122±0.018	-	0.571±0.022	-	0.126±0.008	0.123±0.001	**0.107**±0.002
	720	0.149±0.025	0.140±0.009	0.149±0.012	0.164±0.024	0.136±0.098	-	0.622±0.142	-	0.134±0.004	0.144±0.017	**0.117**±0.019
Traffic-L	96	0.377±0.024	0.332±0.008	0.332±0.007	**0.228**±0.010	0.274±0.012	0.265±0.007	0.342±0.042	-	0.234±0.006	0.231±0.003	0.230±0.010
	192	0.388±0.011	0.326±0.009	0.350±0.010	**0.225**±0.012	-	-	0.354±0.012	-	0.239±0.004	0.236±0.002	0.234±0.003
	336	0.416±0.028	0.335±0.010	0.405±0.011	0.242±0.022	-	-	0.392±0.006	-	0.246±0.003	0.248±0.006	**0.242**±0.007
	720	0.432±0.032	0.361±0.030	0.453±0.022	0.253±0.012	-	-	0.478±0.006	-	0.263±0.001	0.264±0.006	**0.248**±0.010
Weather-L	96	0.146±0.019	0.144±0.017	0.279±0.027	0.145±0.016	0.176±0.011	0.139±0.010	0.113±0.022	0.087±0.012	0.164±0.023	0.145±0.017	**0.086**±0.011
	192	0.148±0.022	0.145±0.015	0.264±0.013	0.144±0.012	0.166±0.022	0.160±0.037	0.144±0.020	0.086±0.007	0.158±0.024	0.147±0.025	**0.083**±0.011
	336	0.152±0.032	0.146±0.011	0.283±0.021	0.149±0.023	0.168±0.014	0.170±0.027	0.138±0.033	0.098±0.002	0.162±0.006	0.160±0.012	**0.093**±0.010
	720	0.162±0.009	0.147±0.019	0.317±0.021	0.152±0.029	0.149±0.034	0.148±0.040	0.141±0.026	0.102±0.005	0.136±0.020	0.135±0.008	**0.099**±0.009
Exchange-L	96	0.079±0.005	0.077±0.001	0.069±0.007	0.079±0.002	0.033±0.003	0.036±0.009	0.090±0.010	0.036±0.005	0.079±0.002	0.091±0.009	**0.032**±0.002
	192	0.081±0.015	0.078±0.008	0.069±0.007	0.081±0.002	0.044±0.004	0.058±0.007	0.106±0.010	0.058±0.005	0.100±0.019	0.087±0.005	**0.040**±0.005
	336	0.086±0.003	0.083±0.005	0.071±0.005	0.085±0.010	0.074±0.017	0.058±0.009	0.106±0.010	0.076±0.006	0.086±0.008	0.091±0.002	**0.054**±0.001
	720	0.116±0.022	0.113±0.015	0.097±0.011	0.126±0.001	0.182±0.010	0.191±0.006	0.142±0.009	0.173±0.020	0.113±0.016	0.103±0.009	**0.084**±0.017
ILI-L	24	0.303±0.021	0.265±0.027	0.271±0.032	0.382±0.018	0.124±0.019	0.118±0.033	0.242±0.086	0.263±0.012	0.296±0.044	0.283±0.001	**0.116**±0.011
	36	0.262±0.013	0.222±0.047	0.258±0.058	0.286±0.037	0.144±0.011	0.143±0.089	0.246±0.117	0.298±0.011	0.298±0.048	0.307±0.007	**0.142**±0.008
	48	0.334±0.028	0.262±0.023	0.225±0.043	0.291±0.032	0.159±0.020	0.160±0.039	0.275±0.044	0.301±0.034	0.320±0.025	0.314±0.009	**0.152**±0.017
	60	0.288±0.031	0.278±0.017	0.245±0.017	0.287±0.023	0.216±0.014	0.183±0.019	0.272±0.020	0.299±0.013	0.325±0.068	0.333±0.005	**0.167**±0.007

Due to the excessive time and memory consumption, some results are unavailable in our implementation and denoted as -.

C.6. Model Analysis

Table 11. Comparison on model efficiency. Lower values of Inference Speed (sec/sample) or Memory (GB) indicate higher model efficiency. L: the forecasting horizon. The results are obtained with batch size equals 1. **Red**: the best, Blue: the 2nd best.

Model	Metric	Electricity-L (L = 720)	ETTm1-L (L = 720)	Exchange-S (L = 30)	Solar-S (L = 24)	Electricity-L (L = 96)	ETTm1-L (L = 96)	Electricity-L (L = 192)	ETTm1-L (L = 192)	Electricity-L (L = 336)	ETTm1-L (L = 336)
TSDiff	Inference Speed	43.068	2.841	1.269	1.858	4.770	1.200	12.339	1.314	20.582	1.676
	Memory	0.896	0.332	0.033	0.040	0.255	0.084	0.330	0.122	0.479	0.184
GRU NVP	Inference Speed	**26.296**	3.460	0.405	0.665	3.450	0.580	**7.414**	1.115	**12.441**	1.856
	Memory	0.427	**0.023**	0.014	0.040	0.145	0.014	0.173	0.015	0.244	**0.018**
GRU MAF	Inference Speed	435.105	18.635	0.817	9.120	49.442	1.631	177.86	4.853	290.000	9.088
	Memory	0.372	0.028	0.013	0.040	0.129	0.023	0.175	0.024	0.246	0.025
Trans MAF	Inference Speed	532.151	19.401	0.883	9.275	45.367	1.900	169.368	5.336	311.954	10.130
	Memory	**0.368**	0.081	0.011	0.037	0.147	0.073	0.201	0.076	0.272	0.075
TimeGrad	Inference Speed	-	-	24.896	19.641	113.103	94.888	142.104	155.013	-	284.951
	Memory	-	-	0.016	0.041	0.128	0.016	**0.149**	0.022	-	0.034
CSDI	Inference Speed	-	86.182	19.251	29.251	388.315	16.328	659.428	25.838	-	39.883
	Memory	-	0.133	0.182	0.723	1.411	0.027	3.024	0.033	-	0.051
K2VAE	Inference Speed	39.834	**0.998**	**0.309**	**0.483**	**3.310**	**0.257**	8.836	**0.374**	17.961	**0.475**
	Memory	0.474	0.028	**0.011**	**0.017**	**0.094**	**0.013**	0.154	**0.015**	**0.240**	0.019

Due to the excessive time and memory consumption, some results are unavailable in our implementation and denoted as -.

Table 12. Comparison on different Koopman Operators. Lower CRPS or NMAE values indicate better performance. **Red**: the best. L: the forecasting horizon.

Koopman Operator	Metrics	Electricity-L (L = 720)	ETTm1-L (L = 720)	Exchange-S (L = 30)	Solar-S (L = 24)	Electricity-L (L = 96)	ETTm1-L (L = 96)	Electricity-L (L = 192)	ETTm1-L (L = 192)	Electricity-L (L = 336)	ETTm1-L (L = 336)
$\mathcal{K}_{loc}$	CRPS	-	-	0.012±0.002	0.450±0.012	0.077±0.003	0.298±0.022	0.114±0.008	-	-	-
	NMAE	-	-	0.014±0.002	0.566±0.015	0.101±0.004	0.387±0.014	0.134±0.011	-	-	-
$\mathcal{K}_{glo}$	CRPS	0.065±0.007	0.311±0.024	0.011±0.001	0.374±0.004	0.079±0.004	0.248±0.016	0.082±0.004	0.263±0.012	0.057±0.002	0.268±0.026
	NMAE	0.130±0.024	0.395±0.027	0.013±0.001	0.488±0.008	0.109±0.005	0.304±0.021	0.106±0.012	0.329±0.017	0.114±0.003	0.341±0.020
$\mathcal{K}_{loc} + \mathcal{K}_{glo}$	CRPS	**0.057**±0.005	**0.294**±0.026	**0.009**±0.001	**0.367**±0.005	**0.073**±0.002	**0.232**±0.010	**0.080**±0.004	**0.259**±0.013	**0.054**±0.001	**0.262**±0.030
	NMAE	**0.117**±0.019	**0.373**±0.032	**0.009**±0.001	**0.480**±0.008	**0.093**±0.002	**0.284**±0.011	**0.102**±0.010	**0.323**±0.020	**0.107**±0.002	**0.330**±0.014

Due to the numerical instability, some results are unavailable in our implementation and denoted as -.

Table 13. Comparison on different connections of KalmanNet. Lower CRPS or NMAE values indicate better performance. **Red**: the best. L: the forecasting horizon.

Connections in KalmanNet	Metrics	Electricity-L (L = 720)	ETTm1-L (L = 720)	Exchange-S (L = 30)	Solar-S (L = 24)	Electricity-L (L = 96)	ETTm1-L (L = 96)	Electricity-L (L = 192)	ETTm1-L (L = 192)	Electricity-L (L = 336)	ETTm1-L (L = 336)
w/o AuxiliaryNet	CRPS	0.082±0.011	0.359±0.024	0.015±0.002	0.398±0.005	0.083±0.001	0.268±0.014	0.107±0.008	0.263±0.014	0.074±0.004	0.267±0.025
	NMAE	0.188±0.028	0.442±0.029	0.022±0.001	0.531±0.010	0.099±0.003	0.348±0.011	0.142±0.012	0.336±0.016	0.147±0.003	0.346±0.018
w/o skip connection	CRPS	0.063±0.007	0.315±0.016	0.011±0.001	0.388±0.006	0.092±0.004	0.243±0.012	0.087±0.002	0.277±0.013	0.058±0.001	0.266±0.019
	NMAE	0.131±0.015	0.402±0.030	0.013±0.004	0.511±0.008	0.116±0.003	0.292±0.010	0.114±0.007	0.376±0.022	0.119±0.001	0.349±0.022
w/o control input	CRPS	0.069±0.005	0.322±0.017	0.013±0.006	0.423±0.005	0.079±0.002	0.242±0.011	0.084±0.003	0.269±0.017	0.064±0.001	0.271±0.026
	NMAE	0.142±0.018	0.418±0.022	0.017±0.003	0.560±0.009	0.104±0.001	0.295±0.012	0.108±0.007	0.341±0.019	0.128±0.002	0.356±0.015
Mixed	CRPS	**0.057**±0.005	**0.294**±0.026	**0.009**±0.001	**0.367**±0.005	**0.073**±0.002	**0.232**±0.010	**0.080**±0.004	**0.259**±0.013	**0.054**±0.001	**0.262**±0.030
	NMAE	**0.117**±0.019	**0.373**±0.032	**0.009**±0.001	**0.480**±0.008	**0.093**±0.002	**0.284**±0.011	**0.102**±0.010	**0.323**±0.020	**0.107**±0.002	**0.330**±0.014

Table 14. Ablations on KoopmanNet and KalmanNet. Lower CRPS or NMAE values indicate better performance. **Red**: the best. L: the forecasting horizon.

Variants	Metrics	Electricity-L (L = 720)	ETTm1-L (L = 720)	Exchange-S (L = 30)	Solar-S (L = 24)	Electricity-L (L = 96)	ETTm1-L (L = 96)	Electricity-L (L = 192)	ETTm1-L (L = 192)	Electricity-L (L = 336)	ETTm1-L (L = 336)
w/o KoopmanNet	CRPS	0.074±0.009	0.443±0.034	0.014±0.002	0.385±0.008	0.079±0.003	0.265±0.018	0.087±0.006	0.288±0.019	0.114±0.006	0.291±0.037
	NMAE	0.162±0.015	0.601±0.058	0.016±0.001	0.528±0.014	0.112±0.006	0.328±0.022	0.112±0.012	0.382±0.027	0.175±0.009	0.372±0.027
w/o KalmanNet	CRPS	0.089±0.011	0.398±0.038	0.011±0.001	0.375±0.005	0.098±0.004	0.278±0.023	0.091±0.003	0.375±0.025	0.266±0.012	0.301±0.035
	NMAE	0.192±0.023	0.539±0.044	0.012±0.001	0.499±0.009	0.133±0.008	0.338±0.012	0.122±0.011	0.443±0.033	0.359±0.017	0.394±0.018
K^2VAE	CRPS	**0.057**±0.005	**0.294**±0.026	**0.009**±0.001	**0.367**±0.005	**0.073**±0.002	**0.232**±0.010	**0.080**±0.004	**0.259**±0.013	**0.054**±0.001	**0.262**±0.030
	NMAE	**0.117**±0.019	**0.373**±0.032	**0.009**±0.001	**0.480**±0.008	**0.093**±0.002	**0.284**±0.011	**0.102**±0.010	**0.323**±0.020	**0.107**±0.002	**0.330**±0.014

C.7. Showcases

We provide some showcases of K^2VAE in Figure 5, 6, and 7, which demonstrates the strong interval estimation capabilities of K^2VAE. We observe that K^2VAE achieves good performance in 95% confidence interval, which means the forecasting horizon of the time series is well modeled and estimated.

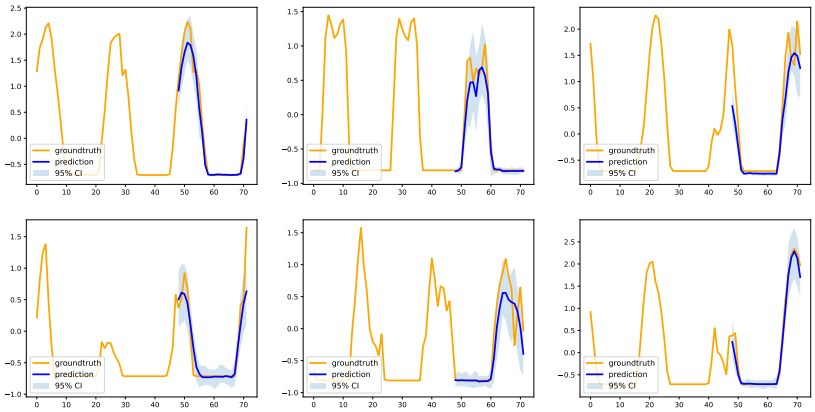

Figure 5. Visualization of input-24-predict-24 results on the Solar-S dataset.

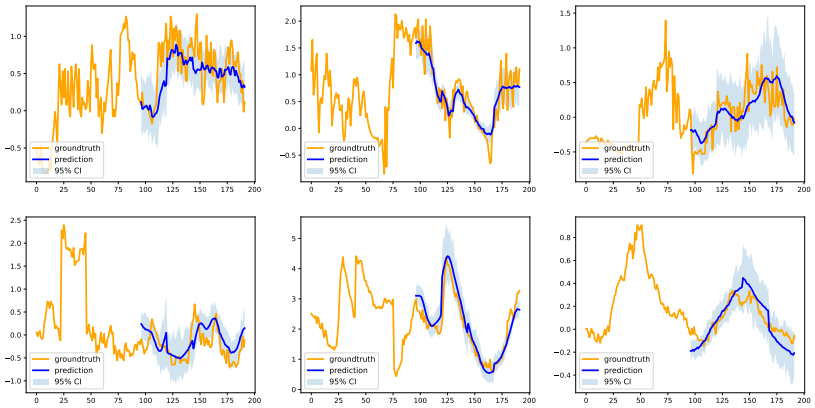

Figure 6. Visualization of input-96-predict-96 results on the ETTm1-L dataset.

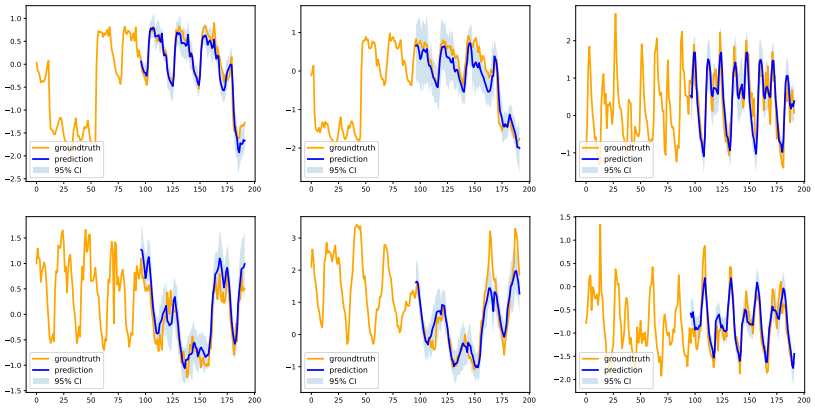

Figure 7. Visualization of input-96-predict-96 results on the Electricity-L dataset.