

# Auto-Regressive Moving Diffusion Models for Time Series Forecasting

Jiaxin Gao<sup>12\*</sup>, Qinglong Cao<sup>12\*</sup>, Yuntian Chen<sup>2†</sup>

<sup>1</sup>Shanghai Jiao Tong University, Shanghai, China

<sup>2</sup>Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang, China  
jiaxingao@sjtu.edu.cn; caoql2022@sjtu.edu.cn; ychen@eitech.edu.cn

## Abstract

Time series forecasting (TSF) is essential in various domains, and recent advancements in diffusion-based TSF models have shown considerable promise. However, these models typically adopt traditional diffusion patterns, treating TSF as a noise-based conditional generation task. This approach neglects the inherent continuous sequential nature of time series, leading to a fundamental misalignment between diffusion mechanisms and the TSF objective, thereby severely impairing performance. To bridge this misalignment, and inspired by the classic Auto-Regressive Moving Average (ARMA) theory, which views time series as continuous sequential progressions evolving from previous data points, we propose a novel Auto-Regressive Moving Diffusion (ARMD) model to first achieve the continuous sequential diffusion-based TSF. Unlike previous methods that start from white Gaussian noise, our model employs chain-based diffusion with priors, accurately modeling the evolution of time series and leveraging intermediate state information to improve forecasting accuracy and stability. Specifically, our approach reinterprets the diffusion process by considering future series as the initial state and historical series as the final state, with intermediate series generated using a sliding-based technique during the forward process. This design aligns the diffusion model’s sampling procedure with the forecasting objective, resulting in an unconditional, continuous sequential diffusion TSF model. Extensive experiments conducted on seven widely used datasets demonstrate that our model achieves state-of-the-art performance, significantly outperforming existing diffusion-based TSF models.

**Code** — <https://github.com/daxin007/ARMD>

## Introduction

Time series forecasting (TSF) plays a pivotal role in various real-world applications, including transportation planning (Fang et al. 2022), energy management (Gao et al. 2023a, 2024), and financial market analysis (Lopez-Lira and Tang 2023). With the development of deep-learning technology, numerous advanced deep-learning TSF models (Piao et al. 2024; Wu et al. 2023; Zhou et al. 2021) have been

\*These authors contributed equally.

†Corresponding author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

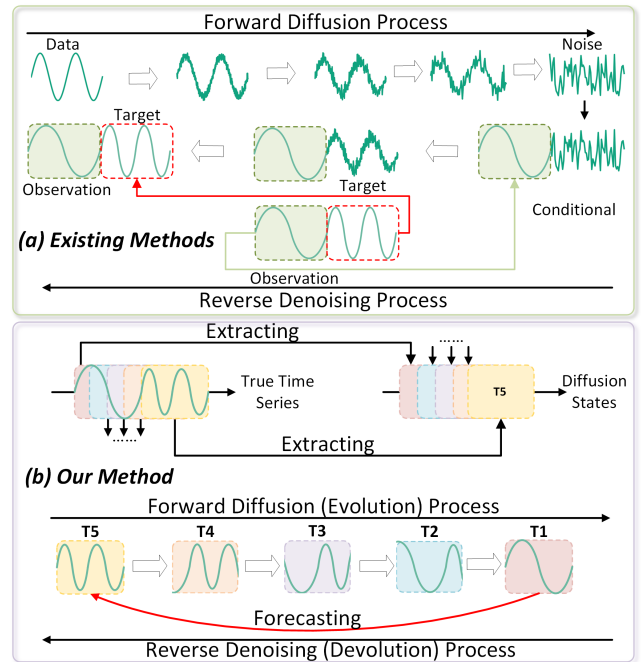


Figure 1: The comparison between (a) existing diffusion-based time series forecasting methods and (b) our methods.

proposed. Notably, diffusion-based models have emerged as some of the most promising methods in the realm of TSF tasks (Meijer and Chen 2024; Yang et al. 2024).

Diffusion models have garnered significant attention in computer vision due to their capability to generate high-quality images (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Cao et al. 2024). These methods typically consist of two main phases: the forward diffusion process, where noise is progressively added to an image until it transforms into white Gaussian noise, and the reverse denoising process, where a denoising network gradually removes noise to reconstruct the original image. While diffusion models are predominantly used in computer vision, recent research has begun exploring their applications in the TSF task.

Current diffusion-based TSF methods (Tashiro et al. 2021; Alcaraz and Strodthoff 2022; Yuan and Qiao 2024)

typically follow traditional diffusion patterns, treating TSF as a noise-based conditional generation task. As illustrated in Fig. 1 (a), the forward diffusion process gradually adds noise to transform the original time series into white Gaussian noise. During forecasting, the historical series serves as a condition, and the reverse denoising process converts the white Gaussian noise into the predicted future time series. While these noise-based conditional diffusion methods have shown some success, they often overlook the continuous sequential nature of time series, leading to a misalignment between the diffusion mechanism and the TSF objective, which significantly suppresses performance. Additionally, by directly diffusing the true series into white Gaussian noise, these methods fail to capture and utilize the valuable intermediate information within the evolution of time series.

To address these limitations, we propose a new continuous sequential diffusion-based model, Auto-Regressive Moving Diffusion (ARMD), for TSF. Drawing inspiration from the principles of the classic TSF theory, Auto-Regressive Moving Average (ARMA), the time series can be viewed as a continuous progression of data points, where each new point is influenced by past data points and includes some random noise. This relationship can be expressed as:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t,$$

where  $x_t$  represents the value at time step  $t$ , and  $\epsilon_t$  represents the random noise. Our model fully leverages the inherent features of time series, which exhibit continuous sequential evolution and form a chain of dependencies characterized by prior information. Unlike previous diffusion methods that focus solely on denoising, ARMD is designed to simulate and learn the underlying evolution of the series.

As illustrated in Fig. 1 (b), our proposed ARMD introduces a novel approach where the future series is progressively diffused into the historical series during the forward diffusion/evolution process. In contrast, the reverse denoising/devolution process leverages the historical series to forecast the future series. This approach conceptualizes the evolution of a time series as a diffusion process, where each intermediate series at a given time step represents a specific state within this process. Consequently, the forward diffusion process can also be described as a forward evolution process of the time series, and the traditional reverse denoising process is reinterpreted as a reverse devolution process. Unlike conventional methods that introduce noise to generate intermediate states from the original series, ARMD derives intermediate states by sliding the series according to the diffusion steps. This results in a deterministic process that eliminates the uncertainty typically associated with noise. These deterministic intermediate series are then employed for model training. In the reverse devolution process, a linear-based devolution (denoising) network is utilized to devolve the series through a distance-based method. Specifically, the devolution network, equipped with a linear module, estimates the distance between an intermediate state (series) and the target series. It then adaptively adjusts the weighting of the intermediate state and the estimated distance based on the time step, ultimately generating the pre-

dition for the target series and its evolution trend.

During the final sampling/forecasting phase, the model begins with the historical time series and gradually transforms it into the future time series. This approach aligns the sampling process with the ultimate prediction objective, removing the need for conditional generation. Consequently, our method improves stability during both training and sampling, resulting in enhanced forecasting performance.

- Inspired by classical time series theory, specifically ARMA, and viewing the evolution of time series as an entire diffusion process, we propose the Auto-Regressive Moving Diffusion (ARMD) model as the first continuous sequential diffusion-based TSF model.
- By employing a simple sliding operation, our method leverages the valuable intermediate information of the time series as intermediate states within the diffusion process. This approach successfully aligns the diffusion mechanism with the true evolution of time series, naturally leading to improved performance.
- The proposed ARMD is extensively evaluated on the common TSF benchmarks, and the experimental results demonstrate that our method achieves state-of-the-art forecasting performance.

## Related Work

Traditional TSF models like ARMA and ARIMA (Box et al. 2015) rely on statistical methods that assume linear relationships between past and present observations to identify patterns. Recently, deep-learning models have gained prominence in TSF. Temporal Convolution Networks (TCNs) (Wu et al. 2023; Gao et al. 2023b) use convolution layers to learn temporal dependencies. Transformer-based models (Xu et al. 2024) are also widely used in TSF. Notably, PatchTST (Nie et al. 2022) segments the time series into multiple tokens and uses attention to learn relationships between them. Client (Gao, Hu, and Chen 2023) and iTransformer (Liu et al. 2024) apply the attention mechanism on the inverted dimensions to capture multivariate correlations. Additionally, DLinear (Zeng et al. 2023) demonstrates the efficacy of linear models in TSF tasks.

In the realm of diffusion-based TSF models, TimeGrad (Rasul et al. 2021) generates future values autoregressively, suitable for short-term forecasting but suffers from error accumulation and slow inference for long-term forecasting. CSDI (Tashiro et al. 2021) avoids autoregressive inference by incorporating a self-supervised strategy with input masking. SSSD (Alcaraz and Strodthoff 2022) improves on CSDI by using a structured state space model instead of Transformers to address quadratic complexity issues, though both CSDI and SSSD face boundary disharmony problems (Lugmayr et al. 2022). TimeDiff (Shen and Kwok 2023) mitigates these limitations with inductive biases tailored for time series data. D3VAE (Li et al. 2022) uses a coupled diffusion probabilistic model for data augmentation, integrating multi-scale denoising score matching and disentangling multivariate latent variables for enhanced accuracy and stability. TS-Diff (Kollovieh et al. 2024) employs a self-guidance mechanism during inference without changing the training process,

excelling in forecasting and synthetic data generation. mr-Diff (Shen, Chen, and Kwok 2024) applies seasonal-trend decomposition for trend extraction and non-autoregressive denoising. Similarly, MG-TSD (Fan et al. 2024) leverages data granularity levels as intermediate diffusion targets for better forecasting. Diffusion-TS (Yuan and Qiao 2024) introduces an encoder-decoder Transformer framework with disentangled temporal representations to generate high-quality multivariate time series samples from noisy inputs.

## Preliminary

**Time Series Forecasting.** Given a historical time series  $X_{-L+1:0}$ , where  $L$  denotes the series length. The objective of TSF is to forecast the future values of the same series, denoted as  $X_{1:T}$ , where  $T$  represents the number of time steps to forecast.  $X_{-L+1:T}$  can be either a univariate time series or a multivariate time series.

**Diffusion Models.** The traditional diffusion models first progressively add noise (Ho, Jain, and Abbeel 2020) to the original data  $X^0$  through a forward diffusion process, producing noised data  $X^T$ . Then the diffusion process is reversed to reconstruct the original data or generate the new data. The intermediate state in each step of diffusion can be calculated based on the original data  $X^0$  and the random noise applied at each step. Trained to predict the added noise or the intermediate state at each step, diffusion models generate new instances by iterative sampling from white Gaussian noise. Diffusion models have been widely applied in diverse fields and have demonstrated notable advantages in generation quality (Croitoru et al. 2023).

**Diffusion Models for TSF in ARMD.** In our proposed ARMD, TSF is aligned with diffusion models by associating the time series with states in the diffusion process. The future series  $X_{1:T}$  would serve as the initial state of the diffusion process. Following the notation convention in (Shen and Kwok 2023), the initial state is further indicated as  $X_{1:T}^0$ . Conversely, the historical time series  $X_{-L+1:0}$  is the final state of the diffusion process, which is further denoted as  $X_{-L+1:0}^T$ . Here, the upper index of  $X$  indicates its state in the diffusion process, while the lower index represents the time steps it encompasses. In ARMD, the length of the historical series matches that of the future series, allowing the historical series to be defined as  $X_{-T+1:0}^T$ . The intermediate state corresponds to a series transitioning from the future series to the historical series, denoted as  $X_{1-t:T-t}^t$ .

## Proposed Model

In this section, we introduce Auto-Regressive Moving Diffusion (ARMD), a novel continuous sequential diffusion-based TSF model. ARMD’s overview is depicted in Fig. 2.

### Forward Diffusion (Evolution) of ARMD

In ARMD, the evolution of a time series is conceptualized as a diffusion process. Here, the future series  $X_{1:T}^0$  serves as the initial state of the forward diffusion (evolution) process. In contrast, the historical series  $X_{-T+1:0}^T$  represents the final state, as opposed to white Gaussian noise commonly

used in traditional diffusion models. Unlike conventional approaches that progressively add noise, the intermediate state  $X_{1-t:T-t}^t$  in ARMD corresponds to sliding the  $X_{1:T}^0$  by  $t$  steps, approaching the historical series  $X_{-T+1:0}^T$ . This approach leverages the properties of time series, where each intermediate state in the diffusion process reflects an intermediate series in the time series evolution. The process where  $X_{2-t:T-t+1}^{t-1}$  is diffused to  $X_{1-t:T-t}^t$  represents a single-step movement towards the historical series, as illustrated in Fig. 2, akin to the  $q$  process in denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020). The process can be formally expressed as:

$$X_{1-t:T-t}^t = \text{Slide}(X_{2-t:T-t+1}^{t-1}, 1), \quad (1)$$

where  $\text{Slide}(X, k)$  denotes the  $k$ -step movement of the series window  $X$  towards the historical series. Furthermore, at any time step  $t$ ,  $X_{1-t:T-t}^t$  can be directly obtained using  $X_{1:T}^0$ , similar to the process in DDPM. Building upon DDPM’s foundational assumption, we extend this principle to our sliding sampling process. This extension enables us to calculate the evolution trends and train the devolution network effectively. Specifically, the  $t$ -step forward process can be rewritten as:

$$X_{1-t:T-t}^t = \text{Slide}(X_{1:T}^0, t) = \sqrt{\bar{\alpha}_t} X_{1:T}^0 + \sqrt{1 - \bar{\alpha}_t} z^t, \quad (2)$$

where  $z^t$  represents the evolution trend from the series  $X_{1:T}^0$  to  $X_{1-t:T-t}^t$ , functionally similar to the noise added in the original DDPM. Given that each time step of  $X_{1-t:T-t}^t$  is deterministic,  $z^t$  can be calculated as:

$$z^t = \left( \sqrt{\frac{1}{\bar{\alpha}_t}} X_{1-t:T-t}^t - X_{1:T}^0 \right) / \sqrt{\frac{1}{\bar{\alpha}_t} - 1}. \quad (3)$$

Here,  $z^t$  serves as the ground truth for the optimization objective at each time step. In this diffusion scheme, the maximum number of diffusion steps  $T$  equals the length of the series to be predicted.

### Reverse Denoising (Devolution) of ARMD

The reverse process in ARMD utilizes the historical series  $X_{-T+1:0}^T$  to iteratively generate (forecast) the future series  $X_{1:T}^0$ . At each devolution step, the linear-based devolution network  $R(\cdot)$  predicts the evolution trend  $z^t$ , which is used to devolve  $X_{1-t:T-t}^t$  to  $X_{2-t:T-t+1}^{t-1}$ . Given an intermediate state (series)  $X_{1-t:T-t}^t$  and the diffusion step  $t$ ,  $R(\cdot)$  predicts  $\hat{X}^0(X^t, t, \theta)$  for the future series  $X_{1:T}^0$ . More specifically, within the devolution network  $R(\cdot)$ , a linear module first provides a prediction of the distance  $D$  from the input  $X_{1-t:T-t}^t$  to  $X_{1:T}^0$ :

$$D = \text{Linear}(X_{1-t:T-t}^t). \quad (4)$$

Then, the diffusion step  $t$  is used to adaptively balance the interaction between the distance prediction  $D$  and the input  $X_{1-t:T-t}^t$ , enabling the network to produce more accurate predictions. Particularly, as  $t$  decreases, the input  $X_{1-t:T-t}^t$

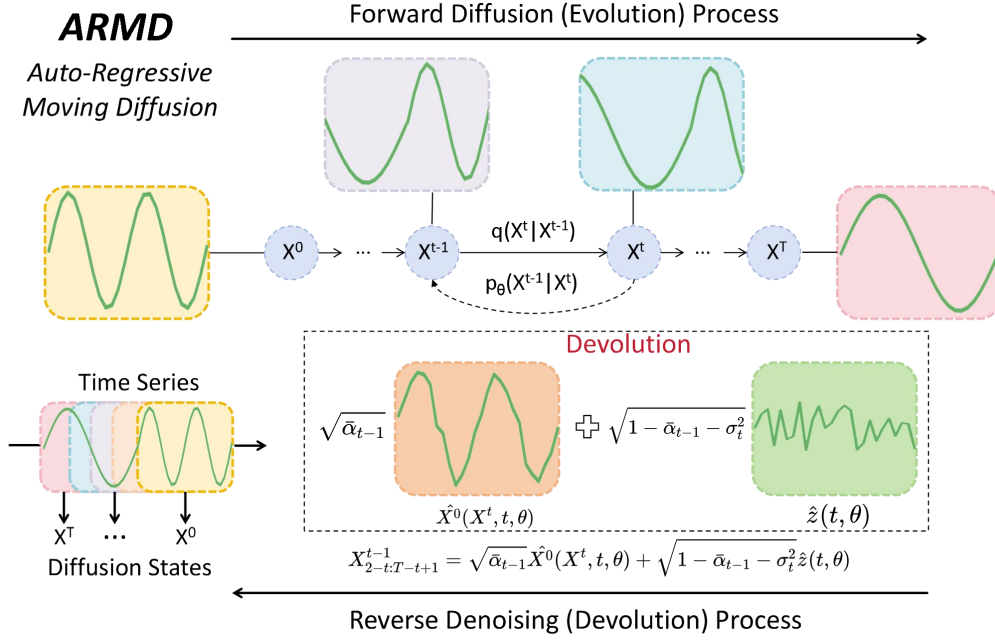


Figure 2: Illustration of the diffusion process in ARMD. During the forward diffusion process, the future series is progressively diffused into the historical series. Conversely, the reverse denoising process utilizes the historical series to iteratively generate/forecast the future series. For clarity, the series  $X$  is only annotated with an upper index to indicate its state.

becomes closer to the target  $X_{1:T}^0$ , so the model's output should increasingly resemble the input, placing less emphasis on  $D$ . Conversely, as  $t$  increases, the model places greater reliance on  $D$ . This adaptive balancing can be mathematically expressed as:

$$\hat{X}^0(X^t, t, \theta) = \frac{W(t) * X_{1-t:T-t}^t + (1 - bW(t)) * D}{(1 + cW(t))^d}, \quad (5)$$

where  $W(t)$  represents a weight coefficient that decreases as  $t$  increases, ranging from 0 to 1. We initialize  $W(t)$  with the predefined coefficients  $\bar{\alpha}_t$  of DDPM, and it is updated along with the linear module's parameters during training. The hyper-parameters  $b$ ,  $c$ , and  $d$  are leveraged to balance the interaction between the distance prediction  $D$  and the input  $X_{1-t:T-t}^t$ . To increase sample diversity and prevent over-fitting, a small deviation is added to the input of  $R(\cdot)$  during the training process.

After obtaining the prediction  $\hat{X}^0(X^t, t, \theta)$ , the predicted evolution trend  $\hat{z}(t, \theta)$  can be calculated as:

$$\hat{z}(t, \theta) = \left( \sqrt{\frac{1}{\bar{\alpha}_t}} X_{1-t:T-t}^t - \hat{X}^0(X^t, t, \theta) \right) / \sqrt{\frac{1}{\bar{\alpha}_t} - 1}. \quad (6)$$

With the ground truth  $z^t$  calculated in Equation (3) and the prediction  $\hat{z}(t, \theta)$ , the training objective is formulated as:

$$L_\theta = \mathbb{E}_t[|z^t - \hat{z}(t, \theta)|]. \quad (7)$$

### Sampling/Forecasting of ARMD

In the sampling phase, starting from the historical series  $X_{-T+1:0}^T$ , ARMD iteratively generates the future series

$X_{1:T}^0$ , which successfully aligns the sampling process with the TSF objective, making the model an unconditional diffusion TSF model. The method follows the sampling approach from DDIM (Song, Meng, and Ermon 2020), replacing the predicted noise  $\epsilon_\theta(x_t, t)$  with the predicted evolution trend  $\hat{z}(t, \theta)$ . The sampling process from  $t$  to  $t-1$  (akin to the  $p$  process in DDPM/DDIM) can be expressed as:

$$X_{2-t:T-t+1}^{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left( \frac{X_{1-t:T-t}^t - \sqrt{1 - \bar{\alpha}_t} \hat{z}(t, \theta)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \hat{z}(t, \theta) + \sigma_t \epsilon_t, \quad (8)$$

where  $\hat{z}(t, \theta)$  is the predicted evolution trend at the time step  $t$ , and  $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Since the series evolution in ARMD is deterministic, we remove the noise term  $\sigma_t \epsilon_t$ . The content within the parentheses in the first term is actually  $\hat{X}^0(X^t, t, \theta)$ . Thus, the simplified equation is:

$$X_{2-t:T-t+1}^{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{X}^0(X^t, t, \theta) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \hat{z}(t, \theta). \quad (9)$$

To accelerate sampling,  $k$  steps can be skipped at each iteration, and the process can be further inferred as:

$$X_{1-t+k:T-t+k}^{t-k} = \sqrt{\bar{\alpha}_{t-k}} \hat{X}^0(X^t, t, \theta) + \sqrt{1 - \bar{\alpha}_{t-k} - \sigma_t^2} \hat{z}(t, \theta). \quad (10)$$

For better understanding, the training and sampling procedures are detailed in Algorithm 1 and Algorithm 2.

---

**Algorithm 1: Training.**

---

**Require:** Maximum number of diffusion steps  $T$ , which also represents the length of the historical/future series; Predefined coefficients  $\bar{\alpha}_{0:T}$ .

- 1: **repeat**
  - 2: Sample  $X_{1:T}^0$  from the training set;
  - 3: Sample  $t \sim \text{Uniform}(\{1, 2, \dots, T\})$ ;
  - 4: Generate the diffused sample  $X_{1-t:T-t}^t$  using Equation (2), and calculate the evolution trend  $z^t$  using Equation (3);
  - 5: Use the devolution network  $R(\cdot)$  to generate the predicted sample  $\hat{X}^0(X^t, t, \theta)$  using Equation (5), and obtain the predicted evolution trend  $\hat{z}(t, \theta)$  using Equation (6);
  - 6: Calculate the loss  $L_\theta$  using Equation (7);
  - 7: Update the devolution network  $R(\cdot)$  of ARMD by taking a gradient descent step on  $\nabla_\theta L$ ;
  - 8: **until** converged.
- 

---

**Algorithm 2: Sampling/Forecasting.**

---

**Require:** Historical series  $X_{-T+1:0}^T$ ; Trained devolution network  $R(\cdot)$ ; Sampling interval  $\Delta t$ ; Predefined coefficients  $\bar{\alpha}_{0:T}$ .

- 1: **for**  $t = T$  **to** 0 **by**  $\Delta t$  **do**
  - 2: Obtain  $\hat{X}^0(X^t, t, \theta)$  using  $X_{1-t:T-t}^t$  and  $t$  with the devolution network  $R(\cdot)$ , and calculate the corresponding evolution trend  $\hat{z}(t, \theta)$  using Equation (6);
  - 3: Update  $X_{1-t:T-t}^t$  using Equation (10);
  - 4: **end for**
  - 5: Output the prediction of  $X_{1:T}^0$ .
- 

## Experiments

### Experimental Settings

ARMD is evaluated on seven widely used benchmark datasets, including Solar Energy (Lai et al. 2018), Exchange (Lai et al. 2018), Stock (Yoon, Jarrett, and Van der Schaar 2019), and four ETT datasets (Zhou et al. 2021). We compare ARMD with five advanced diffusion-based TSF models: Diffusion-TS (Yuan and Qiao 2024), MG-TSD (Fan et al. 2024), TSDiff (Kollovieh et al. 2024), D3VAE (Li et al. 2022), TimeGrad (Rasul et al. 2021). Additionally, some other advanced TSF models, including iTransformer (Liu et al. 2024), TimesNet (Wu et al. 2023), DLinear (Zeng et al. 2023), PatchTST (Nie et al. 2022), and Client (Gao, Hu, and Chen 2023) are also compared with ARMD.

For all datasets, the historical length and prediction length are both set to 96. Following the evaluation methodology employed in a previous study (Zhou et al. 2021), we calculate the mean squared error (MSE) and mean absolute error (MAE) on z-score normalized data, enabling a consistent assessment of various variables.

### Comparison with Diffusion-Based Models

The multivariate forecasting results, comparing our proposed ARMD with various advanced diffusion-based TSF

models, are summarized in Table 1, with each model’s results averaged over 10 sampling runs. Notably, ARMD demonstrates superior forecasting performance, achieving optimal results in 12 out of the 14 experimental settings. Our extensive experiments reveal that while other diffusion-based TSF models can achieve effective results on specific datasets, they often suffer from instability, performing well in certain scenarios but deviating significantly from the true values in others, which indicates a lack of generalization across diverse data. In contrast, ARMD consistently exhibits robust performance across all datasets, suggesting a higher degree of reliability and adaptability to various multivariate forecasting challenges. Specifically, on the ETTm1 dataset, ARMD achieves a substantial 47.7% reduction in MSE and a 30.1% reduction in MAE compared to the second-best model, D3VAE. On the Stock dataset, ARMD surpasses TS-Diff, the second-best model in this setting, with a 28.8% reduction in MSE and a 26.3% reduction in MAE. Furthermore, ARMD attains more than a 10% reduction in both MSE and MAE on the Solar Energy, ETTh1, and Exchange datasets compared to the nearest competitor. These results indicate that ARMD is the best diffusion-based TSF model.

### Comparison with Other TSF Models

As demonstrated in Table 2, ARMD consistently outperforms other advanced models in multivariate TSF tasks. The comparative results for other TSF models are primarily sourced from iTransformer (Liu et al. 2024), with the Stock dataset being an exception, where additional evaluations are conducted using official implementations to ensure a thorough analysis. ARMD achieves the highest number of best counts, indicating its consistent ability to surpass other TSF methods across multiple benchmarks. These results not only establish ARMD as the leading diffusion-based TSF model but also underscore its superiority over existing models. Overall, these findings highlight ARMD’s promise as a robust and effective solution for multivariate TSF tasks.

### Qualitative Analysis

As illustrated in Fig. 3, we qualitatively analyze the performance of ARMD by comparing it with the advanced diffusion-based model, Diffusion-TS. Given the same historical series, both models make 10 independent predictions for the future series. Our analysis reveals that ARMD consistently produces more stable and accurate predictions, particularly in scenarios involving high peak values. This is in contrast to Diffusion-TS, which shows greater variability and less precision in similar settings. We attribute ARMD’s superior performance to its deterministic training process and the reduced sampling steps required during inference. These features of ARMD not only enhance the model’s ability to generate predictions that closely align with actual series but also result in narrower uncertainty distributions.

### Efficiency Comparison

We compare the training and inference efficiency of ARMD with the diffusion-based TSF models Diffusion-TS, MG-TSD, and TimeGrad on ETTm1, as shown in Table 3. Due to

Methods	Metric	Solar Energy	ETTh1	ETTh2	ETTM1	ETTM2	Exchange	Stock	Best Count
<b>ARMD (Ours)</b>	MSE	<b>0.167</b>	<b>0.445</b>	0.311	<b>0.337</b>	<b>0.181</b>	<b>0.093</b>	<b>0.235</b>	<b>12</b>
	MAE	<b>0.236</b>	<b>0.459</b>	<b>0.338</b>	<b>0.376</b>	0.255	<b>0.203</b>	<b>0.269</b>	
Diffusion-TS (Yuan and Qiao 2024)	MSE	0.181	0.643	0.544	0.678	0.497	0.275	0.416	0
	MAE	0.252	0.586	0.494	0.613	0.459	0.382	0.533	
MG-TSD (Fan et al. 2024)	MSE	0.443	1.096	<b>0.295</b>	0.690	0.202	0.396	0.365	1
	MAE	0.529	0.765	0.345	0.631	0.278	0.460	0.453	
TSDiff (Kollovich et al. 2024)	MSE	0.352	0.614	0.470	0.686	0.242	0.125	0.330	0
	MAE	0.432	0.521	0.418	0.603	0.311	0.240	0.365	
D3VAE (Li et al. 2022)	MSE	0.416	1.123	0.389	0.644	0.394	0.240	0.345	0
	MAE	0.492	0.728	0.373	0.538	0.410	0.371	0.390	
TimeGrad (Rasul et al. 2021)	MSE	0.359	0.884	0.297	0.661	0.182	0.508	0.333	1
	MAE	0.449	0.725	0.349	0.639	<b>0.254</b>	0.554	0.376	

Table 1: Result comparisons of multivariate series forecasting with diffusion-based TSF models. The best results are highlighted in bold. The ‘‘Best Count’’ column indicates the times of achieving the best result.

Methods	Metric	Solar Energy	ETTh1	ETTh2	ETTM1	ETTM2	Exchange	Stock	Best Count
<b>ARMD (Ours)</b>	MSE	<b>0.167</b>	0.445	0.311	0.337	0.181	0.093	<b>0.235</b>	<b>7</b>
	MAE	<b>0.236</b>	0.459	<b>0.338</b>	0.376	<b>0.255</b>	<b>0.203</b>	<b>0.269</b>	
iTransformer (Liu et al. 2024)	MSE	0.203	0.386	<b>0.297</b>	0.334	0.180	<b>0.086</b>	0.342	2
	MAE	0.237	0.405	0.349	0.368	0.264	0.206	0.413	
TimesNet (Wu et al. 2023)	MSE	0.250	<b>0.384</b>	0.340	0.338	0.187	0.107	0.427	1
	MAE	0.292	0.402	0.347	0.375	0.267	0.234	0.499	
DLinear (Zeng et al. 2023)	MSE	0.290	0.386	0.333	0.345	0.193	0.088	0.286	1
	MAE	0.378	<b>0.400</b>	0.387	0.372	0.292	0.218	0.325	
PatchTST (Nie et al. 2022)	MSE	0.234	0.414	0.302	<b>0.329</b>	<b>0.175</b>	0.088	0.516	3
	MAE	0.286	0.419	0.348	<b>0.367</b>	0.259	0.205	0.524	
Client (Gao, Hu, and Chen 2023)	MSE	0.199	0.392	0.305	0.336	0.184	<b>0.086</b>	0.352	1
	MAE	0.239	0.409	0.353	0.369	0.267	0.206	0.433	

Table 2: Result comparisons of multivariate series forecasting with other TSF models. The best results are highlighted in bold.

	Training time (s)	Inference time (s)
<b>ARMD (Ours)</b>	<b>11.872</b>	<b>31.650</b>
Diffusion-TS	149.974	1183.593
MG-TSD	834.815	3299.918
TimeGrad	449.792	1749.896

Table 3: Training/inference time (s) of ARMD, Diffusion-TS, MG-TSD and TimeGrad on the ETTm1 dataset.

the optimized model structure and the reduction in the number of sampling steps, ARMD achieves more than a tenfold reduction in both training and inference times compared to other diffusion models.

## Ablation Studies

In this section, we set a range of ablation experiments to testify diverse components of the proposed ARMD.

**Intermediate State Generation Method.** ARMD generates intermediate states using a sliding-based method to maintain series continuity. An alternative method to generate intermediate states involves interpolating between the initial series  $X_{1:T}^0$  and final series  $X_{-T+1:0}^T$ , as described by:

$$X_{1-t:T-t}^t = X_{1:T}^0 + (X_{-T+1:0}^T - X_{1:T}^0) * t/T.$$

The results of using these interpolation-based intermediate states are presented in the second row of Table 4. The results indicate that our sliding-based method consistently outperforms the interpolation-based approach across all settings, underscoring its superiority in preserving series continuity.

Methods	Metric	Solar Energy	ETTh1	ETTh2	ETTm1	ETTm2	Exchange	Stock	Best Count
<b>ARMD (Ours)</b>	MSE	<b>0.167</b>	<b>0.445</b>	0.311	<b>0.337</b>	0.181	<b>0.093</b>	<b>0.235</b>	<b>11</b>
	MAE	<b>0.236</b>	<b>0.459</b>	<b>0.338</b>	<b>0.376</b>	0.255	<b>0.203</b>	<b>0.269</b>	
Interpolation Method	MSE	0.184	0.509	0.407	0.359	0.210	0.131	0.249	0
	MAE	0.250	0.481	0.373	0.385	0.265	0.249	0.278	
T-embedding Method	MSE	0.308	0.707	0.376	0.598	0.235	0.233	0.350	0
	MAE	0.380	0.589	0.380	0.527	0.331	0.330	0.423	
Transformer Backbone	MSE	1.343	0.631	<b>0.285</b>	0.578	<b>0.168</b>	0.137	0.239	3
	MAE	0.713	0.531	0.341	0.499	<b>0.253</b>	0.242	0.271	
Adding Noise	MSE	0.199	0.482	0.334	0.552	0.199	0.103	0.328	0
	MAE	0.315	0.479	0.353	0.526	0.295	0.228	0.347	

Table 4: Ablation experimental results. **Interpolation Method**: replacing the intermediate state generation method with an interpolation method. **T-embedding Method**: using a  $t$ -embedding approach for denoising. **Transformer Backbone**: utilizing a Transformer-based backbone. **Adding Noise**: adding sampling noise. The best results are highlighted in bold.

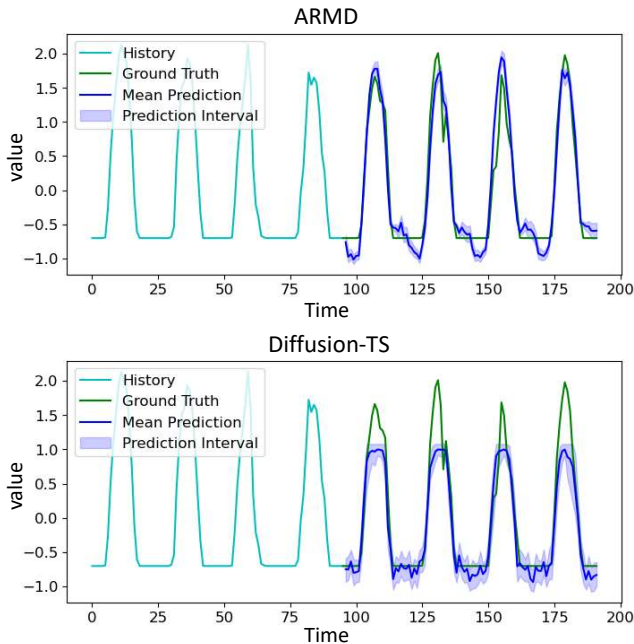


Figure 3: The distributions of 10 different predictions made by ARMD and Diffusion-TS given the same historical series. ARMD achieves more stable and accurate predictions.

**Devolution/Denosing Learning Method.** The devolution network of ARMD takes a distance-based method to provide the prediction  $\hat{X}^0(X^t, t, \theta)$  and  $\hat{z}(t, \theta)$ , differing from traditional denoising (devolution) networks that utilize a  $t$ -embedding-based method (Yuan and Qiao 2024; Shen and Kwok 2023; Shen, Chen, and Kwok 2024). In this method, the model initially generates a prediction based on the input intermediate state, then embeds the time step  $t$  to generate a  $t$ -embedding, and finally combines the  $t$ -embedding with the initial prediction to produce the final output. The re-

sults of applying this  $t$ -embedding-based method within the devolution network are shown in the third row of Table 4, demonstrating that our distance-based method outperforms the  $t$ -embedding-based method across all settings.

**Backbone of the Devolution Network.** Many existing diffusion-based TSF models use a Transformer-based backbone in the denoising network (Yuan and Qiao 2024; Li et al. 2024; Tashiro et al. 2021). Nevertheless, the devolution network in ARMD uses a linear-based backbone, which accelerates training and sampling. The impact of replacing the linear-based backbone with a Transformer-based backbone is presented in the fourth row of Table 4. The results show that the linear-based backbone outperforms the Transformer-based backbone in most settings (11 out of 14). **Noise in the Sampling Process.** Given that the series sampling/forecasting process is deterministic, the noise term  $\sigma_{t\epsilon_t}$  in Equation (8) is set to 0 during sampling/forecasting. The results of adding sampling noise are shown in the last row of Table 4, showing that introducing randomness leads to a decline in model performance.

## Conclusion

Inspired by the classic ARMA theory, we creatively introduce an Auto-Regressive Moving Diffusion (ARMD) model for TSF, which reinterprets the evolution of time series as a diffusion process. Particularly, regarding the target future series as the initial state, and the historical series as the final state, the intermediate series are generated with the series slide operations to complete the diffusion process. In this manner, the sampling procedure of the diffusion model becomes the series forecasting, which denotes that the diffusion mechanism is successfully aligned with the TSF objective, and further resulting in an unconditional, continuous sequential diffusion TSF model. The proposed ARMD is validated across seven widely used datasets, and the experimental results show that our method effectively suits the unique characteristics of time series data, and achieves superior forecasting performance.

## Acknowledgements

This work is financially supported by the The Major Science and Technology Projects of Ningbo (No. 2022Z236), Natural Science Foundation of Ningbo of China (No. 2023J027), China Meteorological Administration under Grant QBZ202316 as well as by the High Performance Computing Centers at Eastern Institute of Technology, Ningbo, and Ningbo Institute of Digital Twin. We would like to express our gratitude to Associate Professor Wenbo Hu from Hefei University of Technology, Associate Professor Hao Sun from Renmin University of China, and Professor Cewu Lu from Shanghai Jiao Tong University for their assistance with this work.

## CRedit Authorship Contribution Statement

**Jiaxin Gao:** Conceptualization, Investigation, Methodology, Software, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing.

**Qinglong Cao:** Conceptualization, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing.

**Yuntian Chen:** Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing – review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- Alcaraz, J. M. L.; and Strodthoff, N. 2022. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399*.
- Box, G. E.; Jenkins, G. M.; Reinsel, G. C.; and Ljung, G. M. 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.
- Cao, Q.; Wang, D.; Li, X.; Chen, Y.; Ma, C.; and Yang, X. 2024. Teaching Video Diffusion Model with Latent Physical Phenomenon Knowledge. *arXiv:2411.11343*.
- Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869.
- Fan, X.; Wu, Y.; Xu, C.; Huang, Y.; Liu, W.; and Bian, J. 2024. MG-TSD: Multi-granularity time series diffusion models with guided learning process. *arXiv preprint arXiv:2403.05751*.
- Fang, W.; Zhuo, W.; Yan, J.; Song, Y.; Jiang, D.; and Zhou, T. 2022. Attention meets long short-term memory: A deep learning network for traffic flow forecasting. *Physica A: Statistical Mechanics and its Applications*, 587: 126485.
- Gao, J.; Cao, Q.; Chen, Y.; and Zhang, D. 2024. Cross-variable Linear Integrated ENhanced Transformer for Photovoltaic power forecasting. *arXiv:2406.03808*.
- Gao, J.; Chen, Y.; Hu, W.; and Zhang, D. 2023a. An adaptive deep-learning load forecasting framework by integrating Transformer and domain knowledge. *Advances in Applied Energy*, 100142.
- Gao, J.; Hu, W.; and Chen, Y. 2023. Client: Cross-variable linear integrated enhanced transformer for multivariate long-term time series forecasting. *arXiv preprint arXiv:2305.18838*.
- Gao, J.; Hu, Y.; Cao, Q.; Dai, S.; and Chen, Y. 2023b. CLearForecast: Contrastive Learning of High-Purity Representations for Time Series Forecasting. *arXiv:2312.05758*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Neural Information Processing Systems*.
- Kollovich, M.; Ansari, A. F.; Bohlke-Schneider, M.; Zschiegner, J.; Wang, H.; and Wang, Y. B. 2024. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. *Advances in Neural Information Processing Systems*, 36.
- Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 95–104.
- Li, Y.; Chen, W.; Hu, X.; Chen, B.; and Zhou, M. 2024. Transformer-Modulated Diffusion Models for Probabilistic Multivariate Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.
- Li, Y.; Lu, X.; Wang, Y.; and Dou, D. 2022. Generative time series forecasting with diffusion, denoise, and disentanglement. *Advances in Neural Information Processing Systems*, 35: 23009–23022.
- Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.
- Lopez-Lira, A.; and Tang, Y. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *arXiv preprint arXiv:2304.07619*.
- Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Meijer, C.; and Chen, L. Y. 2024. The Rise of Diffusion Models in Time-Series Forecasting. *arXiv preprint arXiv:2401.03006*.
- Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2022. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. *arXiv preprint arXiv:2211.14730*.
- Piao, X.; Chen, Z.; Murayama, T.; Matsubara, Y.; and Sakurai, Y. 2024. Fredformer: Frequency Debiased Transformer for Time Series Forecasting. *arXiv preprint arXiv:2406.09009*.
- Rasul, K.; Seward, C.; Schuster, I.; and Vollgraf, R. 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*.

- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Shen, L.; Chen, W.; and Kwok, J. 2024. Multi-Resolution Diffusion Models for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.
- Shen, L.; and Kwok, J. 2023. Non-autoregressive conditional diffusion models for time series prediction. In *International Conference on Machine Learning*, 31016–31029. PMLR.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Tashiro, Y.; Song, J.; Song, Y.; and Ermon, S. 2021. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Neural Information Processing Systems*.
- Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.
- Xu, J.; Wu, C.; Li, Y.-F.; Danoy, G.; and Bouvry, P. 2024. Survey and Taxonomy: The Role of Data-Centric AI in Transformer-Based Time Series Forecasting. *arXiv preprint arXiv:2407.19784*.
- Yang, Y.; Jin, M.; Wen, H.; Zhang, C.; Liang, Y.; Ma, L.; Wang, Y.; Liu, C.; Yang, B.; Xu, Z.; et al. 2024. A survey on diffusion models for time series and spatio-temporal data. *arXiv preprint arXiv:2404.18886*.
- Yoon, J.; Jarrett, D.; and Van der Schaar, M. 2019. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32.
- Yuan, X.; and Qiao, Y. 2024. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*.
- Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11121–11128.
- Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.