MR-ImagenTime: Multi-Resolution Time Series Generation through Dual Image Representations

Xianyong Xu
State Grid Hunan Electric Power Company Limited Research Institute & Hunan Province Engineering Technology Research Center of Electric Power Multimodal Perception and Edge Intelligence
Changsha, China
93615073@qq.com

Yuanjun Zuo
State Grid Hunan Electric Power Company Limited Research Institute & Hunan Province Engineering Technology Research Center of Electric Power Multimodal Perception and Edge Intelligence
Changsha, China
hjzuoyuanjun@163.com

Zhihong Huang
State Grid Hunan Electric Power Company Limited Research Institute & Hunan Province Engineering Technology Research Center of Electric Power Multimodal Perception and Edge Intelligence
Changsha, China
zhihong_huang111@163.com

Yihan Qin
Hunan University
Changsha, China
yihan_qin@hnu.edu.cn

Haoxian Xu
Hunan University
Changsha, China
xuhaoxian13@163.com

Leilei Du
Hunan University
Changsha, China
leileidu@hnu.edu.cn

Haotian Wang
Hunan University
Changsha, China
wanghaotian@hnu.edu.cn

ABSTRACT

Time series forecasting is vital across many domains, yet existing models struggle with fixed-length inputs and inadequate multi-scale modeling. We propose MR-CDM, a framework combining hierarchical multi-resolution trend decomposition, an adaptive embedding mechanism for variable-length inputs, and a multi-scale conditional diffusion process. Evaluations on four real-world datasets demonstrate that MR-CDM significantly outperforms state-of-the-art baselines (e.g., CSDI, Informer), reducing MAE and RMSE by approximately 6–10 to a certain degree.

1 INTRODUCTION

Time series data is fundamental to decision-making in domains such as finance, healthcare, and transportation. While deep learning models like RNNs, CNNs, and Transformers have advanced time series forecasting, recent work has explored diffusion models for their ability to capture complex temporal distributions. However, real-world series often exhibit multi-scale patterns (trends, seasonality, noise) and variable lengths, posing challenges for standard diffusion approaches. Although trend-decomposition methods improve multi-scale modeling, they typically assume fixed input lengths. To address variable-length inputs, Naiman et al. [1] propose mapping time series to 2D image-like representations. While this enhances input flexibility, it risks distorting temporal continuity and long-range dependencies by treating sequential data as spatial grids.

Example 1. *Consider forecasting city-wide electricity demand with heterogeneous sampling rates (e.g., 5-minute vs. hourly sensors), as illustrated in Figure 1. High-frequency data captures local fluctuations, while low-frequency data reflects regional trends. However, most models assume fixed sampling rates. Forcing heterogeneous signals into a common resolution inevitably sacrifices fine-grained details or distorts global structures, limiting the ability to handle multi-scale dynamics.*

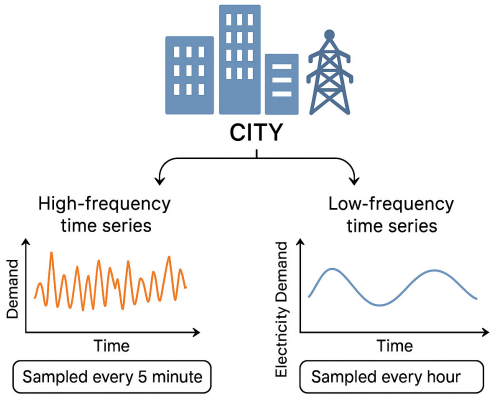

Figure 1: Multi-Resolution Time Series Example

To address these limitations, we propose a diffusion-based forecasting framework that (1) supports variable-length time series without fixed input windows, and (2) incorporates multi-scale trend decomposition to model temporal patterns at different resolutions. The method retains the generative strengths of diffusion models while improving robustness across heterogeneous sequence lengths and temporal scales. Experiments on multiple real-world datasets show that it consistently outperforms both diffusion-based and conventional baselines, achieving more accurate and reliable forecasts.

The remainder of this paper is organized as follows: Section 2 reviews related literature on time series forecasting and diffusion-based generative models. Section 3 introduces the problem statement. Section 4 presents the proposed MR-CDM framework in detail. Section 5 outlines the experimental setup and reports empirical the results. Finally, Section 6 concludes the paper and discusses potential directions for future research.

2 RELATED WORK

Time series forecasting is the process of generating a future segment of a time series by applying specific time series forecasting algorithms to historical time series data. For example, we use a historical wind speed series of 100 time units to generate a future wind speed series of 50 time units.Time series forecasting has undergone three evolutionary phases: statistical modeling, deep learning revolution, and the emerging diffusion paradigm.

A Statistical Modeling

Classical forecasting relied on explicit statistical assumptions, with linear models dominating the field. The Autoregressive Integrated Moving Average (ARIMA) [1] and its seasonal variant SARIMA provided interpretable formulations under stationarity assumptions, while Vector Autoregression (VAR) [2] extended these ideas to multivariate settings. Nonparametric methods like Holt–Winters exponential smoothing [3] and the Error–Trend–Seasonality (ETS) framework [4] modeled trend-seasonality interactions without strict distributional requirements. From a probabilistic perspective, Gaussian Processes (GPs) [5] introduced Bayesian inference for time series, providing principled uncertainty quantification via kernel functions.

B Deep Learning Revolution

Neural networks revolutionized forecasting by enabling hierarchical feature learning and nonlinear modeling. Early recurrent architectures like LSTM [6] and GRU [7] addressed gradient issues to capture long-term dependencies, with DeepAR [8] establishing a deep probabilistic framework. Convolution-based models such as WaveNet [9] and TCNs [10] overcame sequential bottlenecks using dilated causal convolutions for efficient parallel processing. More recently, attention-based architectures like the Transformer [11], Informer [12], and Autoformer [13] have become dominant by capturing global temporal dependencies through self-attention and frequency-domain analysis.

C Emerging Diffusion Paradigm

Diffusion models have recently emerged as powerful generative tools for time series. TimeGrad [14] introduced autoregressive denoising diffusion for probabilistic forecasting, while CSDI [15] extended this to conditional score-based diffusion for imputation tasks. To improve efficiency and handle long sequences, SSSD [16] incorporated state space models, and TSDiff [17] proposed 2D representations to accommodate variable-length sequences. Spatial-temporal extensions like DiffSTG [18] embedded graph structures to model complex spatio-temporal dynamics.

Our framework addresses these exsiting challenges through three core innovations: an adaptive delay embedding mechanism for handling variable-length sequences, a hierarchical multi-resolution decomposition strategy to capture multi-scale temporal patterns, and a conditionally guided diffusion process that leverages coarse-level trends as structural priors. Together, these components ensure robust learning of complex dynamics while enhancing computational efficiency and training stability.

3 PROBLEM STATEMENT

Problem Statement. We address the problem of time series forecasting, where the goal is to predict future values of a time series based on its past observations. Given a set of observed time series data $x \in \mathbb{R}^{L\times K}$, where L is the sequence length and K denotes the number of features, the challenge is to accurately model the complex temporal dependencies within the data, which span multiple time scales. This includes capturing the varying trends at different temporal resolutions and handling the high-dimensional and dynamic nature of the time series data for reliable forecasting.

4 METHOD

A Overview

The proposed MR-CDM framework utilizes a five-stage pipeline to address multi-scale time series characteristics,as illustrated in Figure 2. Initially, Multi-Scale Moving Average Decomposition (with windows of 5, 25, and 51) disentangles the input into short-term fluctuations, medium-term cycles, long-term trends, and high-frequency residuals [19]. Subsequently, these components undergo domain transformation: delay embedding converts short- and medium-term signals into 32×32 images to preserve temporal dynamics, while STFT maps the long-term trend to the time-frequency domain to highlight periodicity. These representations are then concatenated into a 35-channel fused image. A conditional diffusion model [20] then performs generation based on historical contexts to ensure temporal consistency. The process concludes with an Enhanced Hierarchical Reconstructor, which integrates hierarchical reconstruction, cross-scale attention, and adaptive weighting in parallel to recover high-fidelity predictions. This architecture effectively isolates multi-scale features, harnesses the generative power of diffusion models, and ensures precise detail recovery through multi-path fusion.

MR-CDM decomposes the input sequence into multi-scale components via a three-level moving average with window sizes of 5, 25, and 51, explicitly decoupling short-term fluctuations from seasonal periodicity. Trend1 and the residual, which carry local transient information, are mapped to image space [21, 22] via Delay Embedding to preserve fine-grained temporal dependencies. Trend3, representing long-term trends and seasonal cycles, is transformed via STFT to capture the amplitude and phase of periodic components in the time-frequency domain. Trend2 addresses medium-scale periodic structures between the two extremes. All four image branches are fused and modeled jointly by a diffusion model, followed by a hierarchical reconstructor to recover the predicted sequence. This branch-wise design prevents low-frequency components from masking high-frequency details, enhancing the model's ability to represent complex temporal dynamics.

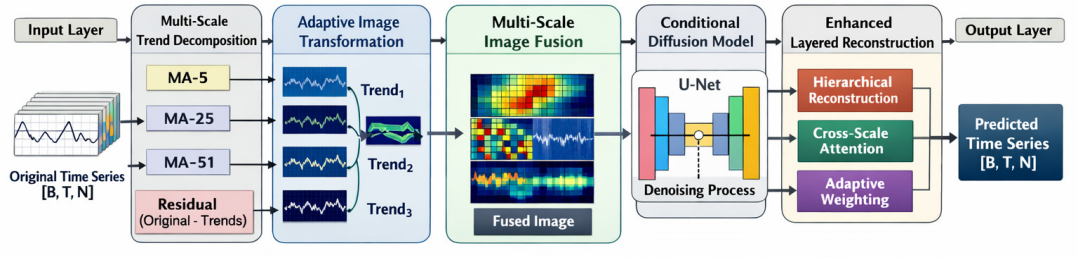

Figure 2: MR-CDM Model Architecture

B Proposed Algorithm

The proposed MR-CDM framework follows a multi-stage pipeline detailed in prior work .We briefly summarize the core components here:

Multi-Scale Decomposition: Input time series are hierarchically decomposed into short, medium, and long-term trends using moving averages, alongside high-frequency residuals [23]. **Time-Series-to-Image Mapping:** Temporal components are transformed into 2D image representations via adaptive delay embedding and STFT[24], enabling resolution-independent processing. **Conditional Diffusion:** A multi-resolution diffusion process is employed, where noise prediction is conditioned on historical context and cross-scale features to ensure temporal consistency.

5 EXPERIMENTS

A Experimental Setup

A.1 Research Objectives. The objective of this study is to evaluate the effectiveness of the proposed **MR-CDM** model for time series forecasting tasks. Specifically, the experiments are designed to: (1) verify the effectiveness of multi-scale trend decomposition in improving forecasting accuracy; (2) evaluate the performance of conditional diffusion models in time series forecasting; (3) analyze the effectiveness of transforming time series into image representations; and (4) compare traditional time series forecasting methods with diffusion-based approaches.

A.2 Experimental Environment. All experiments were conducted on a high-performance workstation configured with an Intel Xeon Silver 4110 CPU, 256 GB of DDR4 RAM, and two NVIDIA TITAN RTX GPUs, each equipped with 24 GB of dedicated memory. The software stack was built on Ubuntu 20.04 LTS and included Python 3.11, PyTorch 2.0.1, and CUDA 11.8 to enable efficient GPU-accelerated training and inference. This setup ensured consistent and reproducible experimental conditions across all evaluations.

A.3 Evaluation Metrics. We adopt three primary evaluation metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), which measure the squared error, absolute error, and scale-consistent error between predictions and ground truth, respectively.

B Datasets and Preprocessing

B.1 Dataset Description. Experiments are conducted on the ETTh1 (Electricity Transformer Temperature - Hourly) dataset. The dataset spans from July 2016 to July 2018 with an hourly sampling rate, containing 17,420 time points and seven features. Following common practice, we focus on the univariate forecasting task of the LUFL variable. The demonstration of our model's adaptability in other fields is provided in Appendix B .

B.2 Data Preprocessing. Data preprocessing consists of three main steps. First, missing values are filled using linear interpolation, and outliers are detected and handled based on the 3σ rule to ensure temporal continuity. Second, Z-score normalization is applied using statistics computed from the training set:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma}.$$

Finally, the dataset is split chronologically into training (70%), validation (10%), and testing (20%) sets, corresponding to 12,194, 1,742, and 3,484 time points, respectively.

B.3 Dataset Sythesis. To demonstrate the generalization capability in similar dataset of our model, we generated a synthetic dataset using a multi-component time series synthesis approach. The synthetic data preserves the statistical properties (mean, standard deviation, and value ranges) of the original ETTh1 dataset while incorporating realistic temporal patterns including daily cycles (24-hour periodicity), weekly cycles (7-day periodicity), long-term trends, and Gaussian noise. Inter-variable correlations were maintained through correlated signal generation, and daytime load enhancement was applied to simulate realistic power consumption patterns.

C Baseline Models

C.1 Traditional Time Series Model. We adopt ARIMA(2,1,2) as a standard univariate baseline, suitable for non-stationary series. It performs 96-step prediction via linear extrapolation of the recent trend, augmented with Gaussian noise to model uncertainty.

C.2 Deep Learning Baseline. A two-layer stacked LSTM with 128 hidden units and 0.2 dropout serves as our deep learning baseline. It generates 96-step forecasts from the final hidden state. We train for 200 epochs using AdamW (lr=0.0001, weight decay=0.00001), with cosine annealing and gradient clipping (max norm 1.0).

C.3 Advanced Diffusion Model. We utilize CSDI, a conditional score-based diffusion model for forecasting. It employs a masking mechanism for conditional generation and uses 6 residual blocks (causal convolutions + 4-head self-attention, hidden dim=64) to model both local and global temporal dependencies.

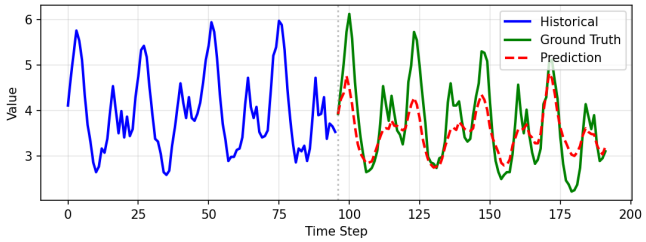

Figure 3: The Prediction Result Based on Our MR-CDM Model

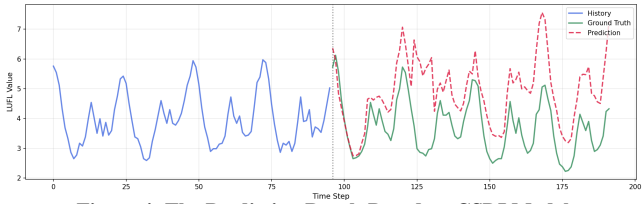

Figure 4: The Prediction Result Based on CSDI Model

D Implements Details of Feature Processing

Multi-scale Trend Decomposition. In the decomposition stage, we employ three moving average filters with window sizes of $[5, 25, 51]$, corresponding to temporal scales of approximately 5 hours, 1 day, and 2 days, respectively. This configuration enables the capture of multi-level features ranging from short-term fluctuations to long-term trends. Specifically, the filter with window size 5 (MA-5) extracts high-frequency short-term fluctuations, MA-25 identifies daily periodic patterns, and MA-51 captures long-term trends. This hierarchical design ensures that features at distinct temporal scales are disentangled and can be subjected to differentiated processing strategies.

Adaptive Image Transformation. For the components Trend_1, Trend_2, and the Residual, we utilize the delay embedding method to transform 1D time series into 32×32 2D images, setting the delay parameter $\tau = 3$ and the embedding dimension $d = 32$. Here, $\tau = 3$ effectively captures temporal dependencies over a 3-hour span, while $d = 32$ aligns with the base resolution of the U-Net architecture, thereby avoiding unnecessary upsampling or downsampling operations. Conversely, for Trend_3, we apply the Short-Time Fourier Transform (STFT) to convert the signal into the time-frequency domain, configured with n_fft = 64 and hop_length = 16. The choice of n_fft = 64 yields 32 frequency components sufficient for identifying periodic characteristics, while hop_length = 16 ensures a 75% window overlap, guaranteeing temporal continuity and information integrity.

Image Fusion. In the fusion stage, we adopt a channel concatenation strategy to merge the images of the four components into a single fused tensor with 35 channels (structured as $7 + 7 + 14 + 7$). This design preserves the complete information of all components, facilitating subsequent hierarchical reconstruction.

E Main Results

E.1 Overall Performance Comparison. Table 1 reports the forecasting performance on the ETTh1 LUFL task on these models above.

Table 1: Performance comparison on ETTh1

Model	MSE	MAE	RMSE
ARIMA	42.2121	6.2893	6.496
LSTM	13.4981	3.6270	3.6739
CSDI	1.5538	0.8992	1.2465
MR-CDM	1.4842	0.9650	1.2183

As shown in Table 1, we compare MR-CDM with two representative baselines and an advanced diffusion model on the ETTh1 feature prediction task. MR-CDM achieves superior performance, with a 96.5 percent MSE reduction compared to ARIMA and an 89.0 percent improvement over LSTM. Although CSDI outperforms ARIMA and LSTM, our model further surpasses CSDI. These results indicate that MR-CDM more effectively captures complex temporal patterns, highlighting the importance of explicitly modeling hierarchical trends.

As illustrated in Figure 3, we evaluate the prediction performance of MR-CDM on long-term LUFL forecasting using the ETTh1 dataset. The input sequence consists of the first 96 time steps (0–95), while the subsequent 96 steps (96–191) form the prediction horizon, covering a total duration of 192 hours (8 days) with hourly resolution.

Experimental results on ETTh1 further confirm that MR-CDM outperforms CSDI, as shown in Figure 4. This advantage arises from key architectural differences. MR-CDM explicitly decomposes time series into multi-scale components and applies tailored strategies such as delay embedding and STFT, whereas CSDI relies on a unified convolution that limits multi-scale representation. By transforming time series into the image domain, MR-CDM better leverages diffusion-based generation. In addition, its multi-path hierarchical reconstructor captures inter-scale dependencies more effectively than CSDI's single-path decoder, leading to improved feature learning and prediction accuracy.

Table 2: Performance comparison on Sythetic Dataset

Model	MSE	MAE	RMSE
ARIMA	39.8131	8.3046	6.3097
LSTM	12.7953	8.2361	3.5770
CSDI	1.6749	0.8741	1.2941
MR-CDM	1.2544	0.9080	1.1200

To further validate robustness and generalization, we conduct the same experiments on a synthetic dataset with similar statistical properties, multi-scale periodicity, and correlations as ETTh1, as shown in Table 2. The consistent improvements on both real and synthetic data demonstrate that MR-CDM does not overfit specific datasets but effectively captures underlying temporal dynamics. This validates the generalizability of combining multi-scale decomposition with conditional diffusion for time series forecasting.

A rigorous multiple-run validation protocol was employed to mitigate the impact of random variation and ensure fair comparison. Appendix C presents the comprehensive statistical summaries, confirming the reliability of our experimental setup and the consistent performance patterns across runs.

To evaluate the multi-step forecasting performance of MR-CDM on time series data, we conduct a comprehensive comparison experiment on the ETTh1 dataset. The historical input window is fixed at 96 time steps, while the prediction horizon is varied across four settings: 24, 48, 96, and 192 steps, enabling a systematic assessment of

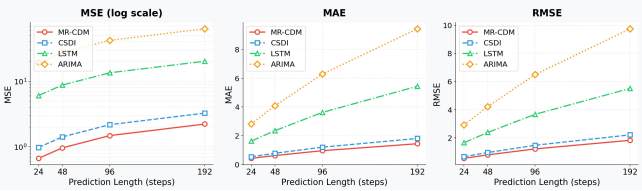

Figure 5: Multi-Step Prediction Performance Comparison

model accuracy and error degradation under increasing forecasting difficulty, as shown in Figure 5. Three baseline models are selected for comparison: ARIMA, LSTM, and CSDI. As illustrated in the figure, MR-CDM consistently achieves the lowest error across all prediction horizons, and its error growth rate remains significantly lower than that of the baselines as the prediction length increases. These results demonstrate that MR-CDM effectively suppresses error accumulation in long-range forecasting scenarios, confirming the superiority of the proposed approach for multi-step time series prediction tasks.

We conduct ablation studies to analyze the effectiveness of each component. Due to space limitations, detailed results are provided in Appendix A.1 and Appendix A.2.

6 CONCLUSION

This thesis presents a comprehensive study on time series forecasting through the development of **MR-CDM**, a novel framework that addresses key limitations of existing methods. The research makes several significant contributions: (1) a delay embedding technique that converts variable-length time series into structured 2D image representations while preserving temporal dependencies, thereby enabling the use of spatial inductive biases from computer vision; (2) a hierarchical trend decomposition module that explicitly captures multi-scale temporal patterns—including short-term fluctuations, seasonal cycles, and long-term trends; and (3) a hierarchical conditional diffusion model that performs denoising generation under multi-scale guidance, reducing stochasticity through coarse-to-fine constraints.

7 ACKNOWLEDGMENT

This research was funded by Science and Technology Project of State Grid Hunan Electric Power Company Limited, titled "Research on Key Technologies and Complete Equipment of Diffusion Super-Resolution Data Augmentation for Power Grid Model Identification and Situation Deduction", grant number 5216A5250009.

REFERENCES

[1] I. Naiman, N. Berman, I. Pemper, I. Arbiv, G. Fadlon, and O. Azencot, "Utilizing image transforms and diffusion models for generative modeling of short and long time series," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024* (A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, eds.), 2024.

[2] E. Zivot and J. Wang, *Vector Autoregressive Models for Multivariate Time Series*, pp. 369–413. New York, NY: Springer New York, 2003.

[3] A. B. Koehler, R. D. Snyder, and J. Ord, "Forecasting models and prediction intervals for the multiplicative holt–winters method," *International Journal of Forecasting*, vol. 17, no. 2, pp. 269–286, 2001.

[4] D. J. Hand, "Forecasting with exponential smoothing: The state space approach by rob j. hyndman, anne b. koehler, j. keith ord, ralph d. snyder," *International Statistical Review*, vol. 77, no. 2, pp. 315–316, 2009.

[5] C. E. Rasmussen and C. K. I. Williams, "Gaussian processes for machine learning (adaptive computation and machine learning)," *The MIT Press*, 2005.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] V. S. Ishwarya and M. Kothandaraman, "A novel feature-fusion-based sparse masked attention network for acoustic echo cancellation using wavelet and STFT synergies," *Circuits Syst. Signal Process.*, vol. 44, no. 4, pp. 2882–2901, 2025.

[8] V. Flunkert, D. Salinas, and J. Gasthaus, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, 2020.

[9] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop, SSW 2016, Sunnyvale, CA, USA, September 13-15, 2016* (A. W. Black, ed.), p. 125, ISCA, 2016.

[10] S. Bai, J. Z. Kolter, and V. Koltun, "Trellis networks for sequence modeling," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv*, 2017.

[12] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," *CoRR*, vol. abs/2012.07436, 2020.

[13] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," pp. 22419–22430, 2021.

[14] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," pp. 8857–8868, 2021.

[15] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "CSDI: conditional score-based diffusion models for probabilistic time series imputation," pp. 24804–24816, 2021.

[16] J. M. L. Alcaraz and N. Strodthoff, "Diffusion-based time series imputation and forecasting with structured state space models," *CoRR*, vol. abs/2208.09399, 2022.

[17] M. Kollovieh, K. Stelzner, J. Kossen, M. Lützenberger, and K. Kersting, "Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting," in *Advances in Neural Information Processing Systems*, 2023.

[18] H. Wen, Y. Lin, Y. Xia, H. Wan, Q. Wen, R. Zimmermann, and Y. Liang, "Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models," in *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, (New York, NY, USA), Association for Computing Machinery, 2023.

[19] L. Shen, W. Chen, and J. T. Kwok, "Multi-resolution diffusion models for time series forecasting," 2024.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.

[21] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*, pp. 366–381, 2006.

[22] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '83, Boston, Massachusetts, USA, April 14-16, 1983*, pp. 804–807, IEEE, 1983.

[23] L. Shen, W. Chen, and J. T. Kwok, "Multi-resolution diffusion models for time series forecasting," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024.

[24] V. S. Ishwarya and M. Kothandaraman, "A novel feature-fusion-based sparse masked attention network for acoustic echo cancellation using wavelet and STFT synergies," *Circuits Syst. Signal Process.*, vol. 44, no. 4, pp. 2882–2901, 2025.

A ABLATION STUDIES

A.1 First Ablation Study

To validate the effectiveness of key components in MR-CDM, we conduct ablation studies on the LUFL univariate forecasting task. We evaluate three variants: (1) **Baseline-NoDecomposition**, which

removes the multi-scale trend decomposition; (2) **UnconditionalDiffusion**, which disables historical conditioning in the diffusion process; and (3) **NoImageFusion**, which replaces our smart fusion module with simple feature concatenation. These components are selected because they each play a distinct and essential role: trend decomposition captures multi-resolution temporal patterns, conditional diffusion leverages historical context to guide generation, and image-inspired fusion enables effective integration of decomposed signals. The ablation results help isolate their individual contributions to the model's performance.

The first ablation experiment use consistent configurations: sequence length of 96 time steps for both input and prediction, batch size of 16, learning rate of 0.001 with Adam optimizer, and training for 50 epochs. The models are evaluated using MSE, MAE, and RMSE metrics on an 80/20 train-test split. This experimental design allows us to quantify the individual contribution of each component while maintaining computational efficiency for rapid iteration.

First Ablation Study Result. The results of the ablation experiments, summarized in Table 3, reveal the importance of each component in MR-CDM. To verify the effectiveness of our design, we conducted a systematic ablation study on the LUFL features of the ETTh1 dataset. The trend decomposition module proves essential: removing it (Baseline-NoDecomposition) leads to an 89.6% performance degradation compared to the FullModel, confirming that multi-scale decomposition is crucial for capturing complex temporal patterns. Even more dramatically, the UnconditionalDiffusion variant—where historical trends are not used to condition the diffusion process—suffers a 1280% drop in performance, strongly highlighting the necessity of conditioning on coarse-grained historical information for accurate and stable forecasting.

The NoImageFusion variant, which replaces our proposed image-inspired fusion with simple feature concatenation, performs better than the baseline but still falls significantly short of the full model. This demonstrates that naive concatenation fails to exploit the rich cross-scale correlations among decomposed trends, whereas our smart fusion mechanism effectively integrates multi-resolution features. Importantly, the removal of any single component causes a substantial performance decline that cannot be compensated by the remaining modules, validating the complementary nature of our MR-CDM architecture. Together, these components form a synergistic system, enabling the complete model to achieve the best results across all metrics (MSE: 1.4842, MAE: 0.9650, RMSE: 1.2183), thereby fully verifying the effectiveness of our approach.

Experiments demonstrate that MR-CDM significantly outperforms baselines. Ablation studies attribute this gain to three core mechanisms: multi-scale trend decomposition disentangles short-, mid-, and long-term components for independent feature learning, surpassing ARIMA's linearity,LSTM's monolithic encoding and CSDI's instability; conditional diffusion injects historical context to ensure temporal continuity, yielding higher accuracy than unconditional generation; and a multi-path hierarchical reconstructor leverages parallel pathways (hierarchical recovery, cross-scale attention, adaptive weighting) to precisely restore details, avoiding LSTM's gradient vanishing and information loss. Crucially, the complete framework exhibits strong synergistic effects, where the integrated performance exceeds the sum of individual contributions, validating the necessity of the proposed architecture.

Table 3: Ablation Studies on ETTh1

Model-Variant	MSE	MAE	RMSE
Baseline-NoDecomposition	14.2177	3.6753	3.7706
UnconditionalDiffusion	20.4822	4.4252	4.5257
NoImageFusion	13.0950	3.5406	3.6187
FullModel	1.4842	0.9650	1.2183

To systematically evaluate the role of the conditional information in the proposed diffusion model, we designed a set of ablation experiments, in which all external condition inputs (such as time features, covariates, or historical context guidance) were removed, and the diffusion process itself was solely relied upon for unconditional prediction of the time series. Figure 6 shows the prediction results of the model on the ETTh1 dataset under this ablation setting.

Compared with the MR-CDM with conditional information introduced in the main experiment, the unconditional diffusion model can roughly capture the overall trend of the sequence, but it is significantly deficient in detail modeling, phase alignment, and long-term dynamic evolution, manifested as overly smooth prediction curves, delayed peak responses, and distortion of local fluctuations.These results fully demonstrate that the introduced conditional mechanism plays a crucial role in enhancing the expression ability and temporal consistency of the diffusion model in complex time series prediction tasks.

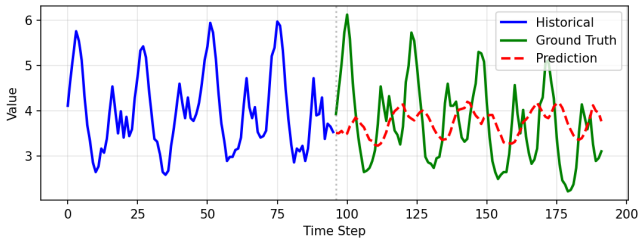

Figure 6: Prediction Result Based on MR- CDM model without conditional information

To visually demonstrate the crucial role of conditional information in the diffusion model, Figure 3 and Figure 6 respectively show the performance of the unconditional diffusion model and the proposed conditional diffusion model in the same time series prediction task. Here, the blue solid line represents the historical observed sequence (0–95 steps) as input, the green solid line represents the true values (Ground Truth, 96–191 steps), and the red dashed line represents the model's prediction results. From the figures, it can be seen that under the unconditional setting, although the model can capture the overall trend, there are significant deviations in terms of detail fluctuations, peak responses, and temporal alignment, especially in the rapidly changing areas, indicating its lack of the ability to precisely model the temporal dynamic structure.

In contrast, after introducing conditional information, the model can more accurately track the intense fluctuations and local features of the true signal, and the prediction curve closely matches the true values, significantly enhancing the ability to restore short-term fluctuations and maintain long-term consistency. This comparison fully

validates the effectiveness of the conditional mechanism in guiding the diffusion process and enhancing time-dependent modeling, further highlighting the superior performance of the proposed method in complex time series prediction tasks.

Table 4: Ablation Studies on Synthetic ETTh1

Model-Variant	MSE	MAE	RMSE
Baseline-NoDecomposition	108.5863	8.0619	10.4204
UnconditionalDiffusion	116.4437	8.3155	10.7909
NoImageFusion	111.8180	8.2349	10.5744
FullModel	1.2544	0.9080	1.1200

As shown in Table 4,we further conducted ablation studies on the synthetic dataset, and the results consistently demonstrate that each component of MR-CDM plays a vital role: multi-scale trend decomposition effectively captures hierarchical temporal patterns, conditional diffusion leverages historical context for accurate generation, and the image-based fusion mechanism enables meaningful integration of multi-scale features. Removing any one of these components leads to a significant performance drop, confirming that the full architecture is necessary and well-designed.

A.2 Second Ablation Study

To systematically evaluate the structural effectiveness of the proposed MR-CDM in multi-scale time-series modeling, we conduct an ablation study on the trend decomposition module. The experiments are performed on the LUFL variable of the ETTh1 dataset under a unified setting of prediction length, with MSE, MAE, and RMSE adopted as evaluation metrics. Specifically, three configurations are compared: the full MR-CDM model, MR-CDM w/o Trend1 (removing the low-frequency trend component), and MR-CDM w/o Trend3 (removing the high-frequency component). The reason for not removing Trend2 is that it is not the "highest frequency" nor the "smoothest", but somewhere in between, mainly carrying medium-term structural information.

The results Table 5 show that the full model achieves the best performance across all three metrics. Removing either trend component leads to a noticeable degradation in prediction accuracy, with a larger drop observed when Trend3 is removed. This indicates that the high-frequency component is critical for short-term fluctuation modeling, while the low-frequency component remains indispensable for preserving global trend dynamics. Overall, the proposed multi-scale trend decomposition mechanism enables collaborative modeling of short-term disturbances and long-term evolution within a unified framework, thereby improving model robustness and generalization in practical forecasting scenarios.

Table 5: Trend Decomposition Ablation Result

Model-Variant	MSE	MAE	RMSE
MR-CDM w/o Trend1	1.9342	1.1247	1.3907
MR-CDM w/o Trend3	2.2769	1.2678	1.5089
FullModel	1.4842	0.9650	1.2183

A.3 Input Length Sensitivity Analysis

To further investigate the effect of historical input length on the forecasting performance of MR-CDM, we conduct an input length sensitivity analysis on the LUFL variable of the ETTh1 dataset. The prediction length is fixed at 96, while the input sequence length Seq_Len is varied across 48, 96, and 192. As shown in the Table 6, all three error metrics decrease monotonically as the input length increases, indicating that longer historical sequences provide richer temporal context for the model.Longer input sequences contain more complete periodic structures and trend dynamics, enabling the multi-scale decomposition module to extract each frequency component more accurately and reducing decomposition errors caused by insufficient historical context. This enables the multi-scale trend decomposition module to extract more complete periodic and trend features, thereby improving prediction accuracy. These results confirm that MR-CDM is adaptive to varying input lengths, and suggest that increasing the historical window size in practical deployment can further enhance forecasting performance.

Table 6: Input Length Sensitivity Results

Seq_Len	MSE	MAE	RMSE
48	1.8998	1.1387	1.3783
96	1.4842	0.9650	1.2183
192	1.2319	0.8492	1.0965

B GENERALIZATION TO OTHER DOMAINS

To further validate the generalization capability and robustness of our model, we extended our evaluation to datasets from distinct domains: weather forecasting and traffic flow prediction. These datasets embody complex non-linear spatiotemporal dynamics characteristic of meteorological variations and urban traffic patterns, respectively, thereby serving as rigorous benchmarks for assessing cross-domain adaptability.In our experiments, the OT metric serves as the ground truth for forecasting tasks on both datasets. Experimental results demonstrate that our model not only retains its performance advantages but also achieves superior prediction accuracy and stability in these diverse scenarios, underscoring its significant potential for broad real-world applications.Table 7 and Table 8 show the prediction ability among these models.

Weather dataset comprises 21 multivariate time series collected from a meteorological station in Jena, Germany, with a sampling rate of 10 minutes. Characterized by complex seasonal patterns (daily and yearly cycles) and high-frequency noise, this dataset tests the model's robustness in handling non-stationary environmental data and capturing inter-variable dependencies.

Table 7: Model Performance on Weather Domain

Model	MSE	MAE	RMSE
ARIMA	191.2356	11.6258	13.8287
LSTM	124.7143	7.5514	11.1675
CSDI	96.3245	6.9438	9.8145
MR-CDM	79.4128	6.3487	8.9113

Traffic dataset records the road occupancy rates collected from 862 sensors in the San Francisco Bay Area. The data is sampled hourly and exhibits strong daily and weekly periodicity, as well as complex spatial dependencies among sensors. It serves as a rigorous benchmark for evaluating a model's ability to capture recurring patterns and handle high-dimensional multivariate series.

Table 8: Model Performance on Traffic Domain

Model	MSE	MAE	RMSE
ARIMA	0.0011	0.0237	0.0331
LSTM	0.0007	0.0152	0.0264
CSDI	0.0004	0.0143	0.0200
MR-CDM	0.0003	0.0139	0.0173

C MULTIPLE-RUN VALIDATION

To ensure the robustness and reproducibility of our results, each experiment is repeated three times using distinct random seeds (specifically, 42, 43, and 44). This practice helps account for the inherent stochasticity in model initialization and training dynamics. The corresponding statistical summaries—including mean and standard deviation of key performance metrics across the three runs—are reported in Table 9(ETTh1) and Table 10(Synthetic ETTH1). These aggregated results provide a more reliable basis for comparison and mitigate the influence of random variation on our conclusions.

Table 9: Multiple-Run Performance on ETTh1

Model	MSE	MAE	RMSE
ARIMA	30.0946	8.2296	5.4858
LSTM	13.2883	6.2361	3.6453
CSDI	2.7493	1.2478	1.6581

To ensure the reliability and reproducibility of our experimental results, we conducted multiple-run validation for all baseline models. Specifically, we trained and evaluated ARIMA,LSTM and CSDI models three times with different random seeds, and reported the averaged performance metrics across all runs. This rigorous validation protocol helps mitigate the impact of random initialization and stochastic training processes, providing more robust and statistically reliable comparisons. The results presented in Table 9 show the mean performance across three independent runs. These averaged results demonstrate consistent performance patterns across multiple runs, confirming the stability of baseline model performance and ensuring fair comparison with our proposed method. The relatively small variance across runs (not shown for brevity) further validates the reliability of our experimental setup and the robustness of the reported performance improvements.

Table 10: Multiple-Run Performance on Synsetic ETTh1

Model	MSE	MAE	RMSE
ARIMA	47.2826	7.4434	6.8762
LSTM	13.6146	3.6670	3.6897
CSDI	2.6159	1.9632	1.6173

Similarly, we also repeated the aforementioned baseline experiments on the synthetic dataset three times, using different random seeds to account for stochastic variations in model initialization and training. The results across all runs demonstrate consistent performance for ARIMA, LSTM and CSDI baseline models, confirming their stability on this controlled dataset. As shown in the corresponding evaluation metrics reported in the relevant Table 10, the low variance across repetitions further validates the reliability of these baselines under synthetic conditions. These findings not only reinforce the robustness of our experimental setup but also provide a solid foundation for comparing more advanced methods, thereby supporting our overall analysis and conclusions.

Experiments demonstrate that MR-CDM significantly outperforms baselines. Experiments studies attribute this gain to three core mechanisms: multi-scale trend decomposition disentangles short-, mid-, and long-term components for independent feature learning, surpassing ARIMA's linearity,LSTM's monolithic encoding and CSDI's instability; conditional diffusion injects historical context to ensure temporal continuity, yielding higher accuracy than unconditional generation; and a multi-path hierarchical reconstructor leverages parallel pathways (hierarchical recovery, cross-scale attention, adaptive weighting) to precisely restore details, avoiding LSTM's gradient vanishing and information loss. Crucially, the complete framework exhibits strong synergistic effects, where the integrated performance exceeds the sum of individual contributions, validating the necessity of the proposed architecture.

D BACKGROUND

Diffusion Models. Diffusion models [20] consist of a forward noising and a backward denoising process. *Forward Diffusion* gradually adds Gaussian noise over K steps. The closed-form expression for step k is:

$$\mathbf{x}^k = \sqrt{\bar{\alpha}_k}\mathbf{x}^0 + \sqrt{1-\bar{\alpha}_k}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_k = \prod_{s=1}^{k}(1-\beta_s)$. *Reverse Denoising* recovers clean data from noise, typically by predicting the noise ϵ or the clean data $\mathbf{x}^0$ via:

$$\mathcal{L}_\epsilon = \mathbb{E}\|\epsilon - \epsilon_\theta(\mathbf{x}^k, k)\|^2 \quad \text{or} \quad \mathcal{L}_\mathbf{x} = \mathbb{E}\|\mathbf{x}^0 - \mathbf{x}_\theta(\mathbf{x}^k, k)\|^2. \tag{2}$$

Conditional Diffusion Models. For time series prediction, the denoising process is conditioned on historical context $\mathbf{c} = \mathcal{F}(\mathbf{x}^0_{-L+1:0})$. The conditional distribution is defined as:

$$p_\theta(\mathbf{x}^{0:K}_{1:H} \mid \mathbf{c}) = p_\theta(\mathbf{x}^K_{1:H}) \prod_{k=1}^{K} p_\theta(\mathbf{x}^{k-1}_{1:H} \mid \mathbf{x}^k_{1:H}, \mathbf{c}), \tag{3}$$

where the mean $\mu_\theta(\mathbf{x}^k, k \mid \mathbf{c})$ is computed by leveraging both the noisy input and the conditioning context $\mathbf{c}$.

Hierarchical Trend Decomposition (HTD). HTD [19] decomposes time series into multi-scale trends. Given a series X_0, trend components are extracted sequentially:

$$X_s = \text{AvgPool}(\text{Padding}(X_{s-1}), \tau_s), \quad s = 1, \ldots, S-1, \tag{4}$$

where AvgPool is average pooling and τ_s (increasing with s) controls the smoothing kernel size, enabling extraction of progressively coarser trends.

Time Series to Image Transforms. Time series are mapped to images using invertible transforms to exploit spatial inductive biases [21, 22]:

- **Delay Embedding:** For a univariate series $x_{1:L}$, it constructs a matrix $X \in \mathbb{R}^{n\times q}$ where n is the embedding dimension and q is the number of time steps. Adaptive variants dynamically adjust m and n to capture complex dynamics.
- **Short Time Fourier Transform (STFT):** Maps the signal to the frequency domain using a sliding window. For input $\mathbf{x} \in \mathbb{R}^{L\times K}$, STFT outputs $\mathbf{x}_{\text{img}} \in \mathbb{R}^{2K\times H\times W}$ (storing real/imaginary parts). It is invertible with negligible information loss.