Diffusion Model for Financial Time Series

Introduction

Abstract

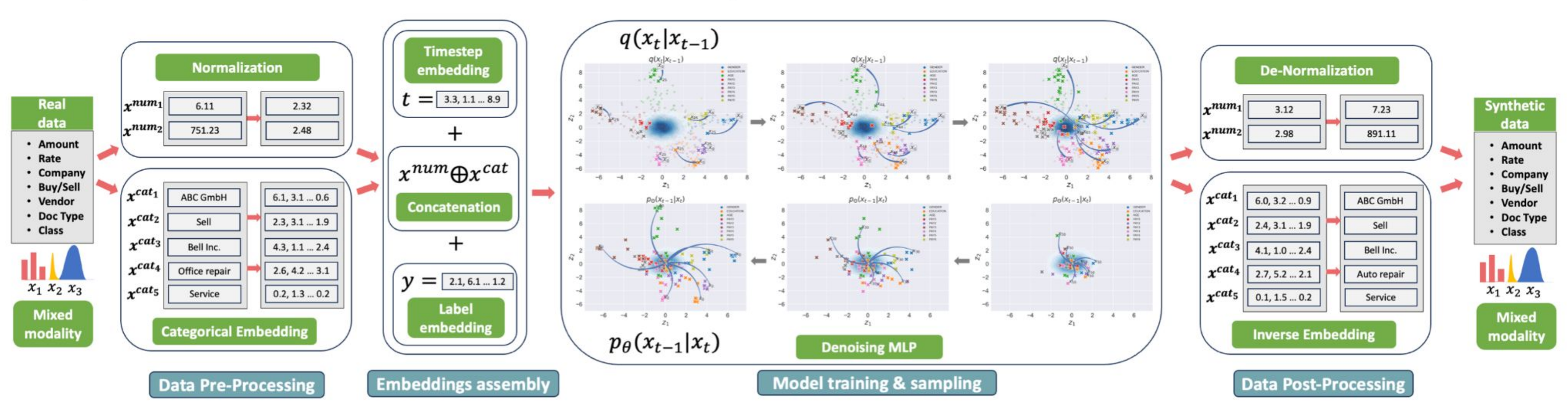
This project employs a diffusion model framework to generate synthetic financial time series data, enhanced by a **classifier-free conditional diffusion** approach that incorporates market regimes. Alongside **Transformer**, **BVAE**, and **Perceiver** architectures, we also integrated time series prediction models such as **DLinear**, **PatchTST**, **iTransformer**, and **UniTST**. The **seq2seq-style** models developed for this purpose showed significant improvement over previous tabular models, particularly in replicating the temporal patterns and statistical moments of real financial data. Our framework was validated using benchmark datasets, including indices for equities, fixed income, and real estate, resulting in synthetic data that closely aligns with the characteristics of actual market data.

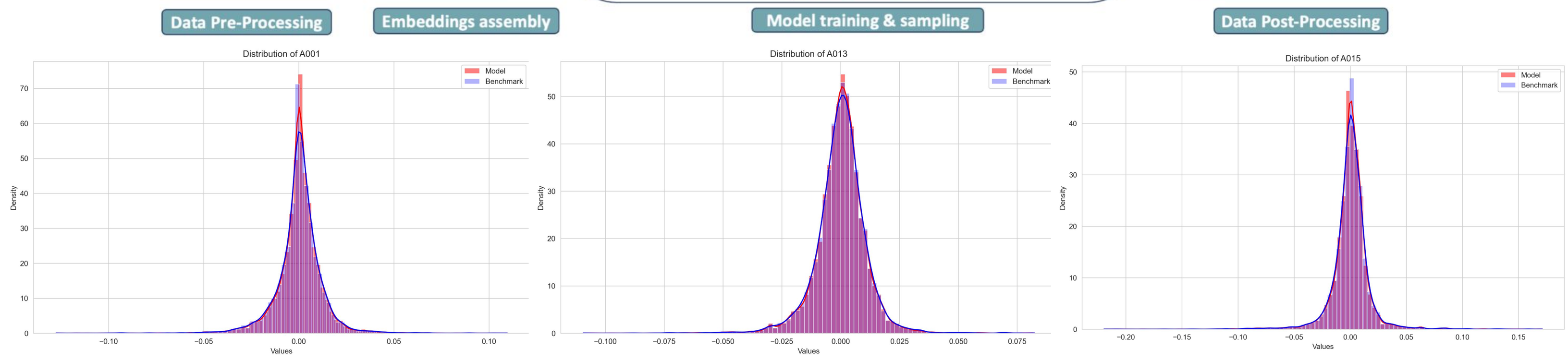
The Diffusion Model Framework

Our financial time series diffusion framework supports two types of diffusion model:

1. **Denoising Diffusion Probabilistic Model (DDPM)**: In this approach, the model gradually adds Gaussian noise to the data and then learns to predict and remove this noise at each step to recover the original data distribution.
2. **DiffusionTS**: This approach directly predicts the original data from the perturbed version without explicitly modeling the noise, allowing for a more straightforward reconstruction of the data.

In our experiments, we observed that the DDPM approach consistently performed well, particularly in accurately recovering the original data distribution, making it a robust choice for synthetic financial time series generation.

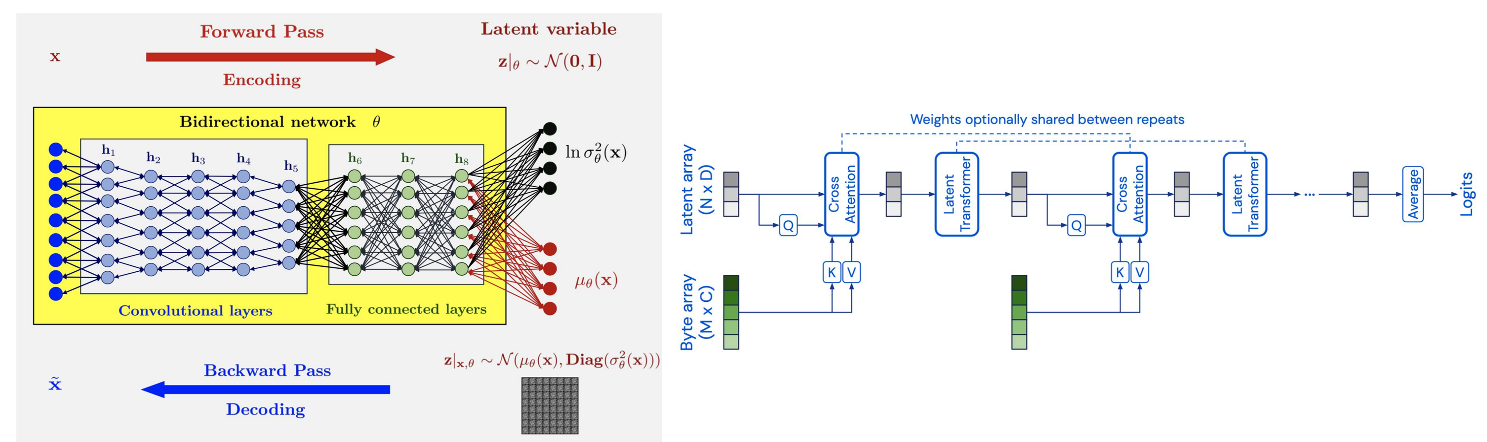




Model Exploration

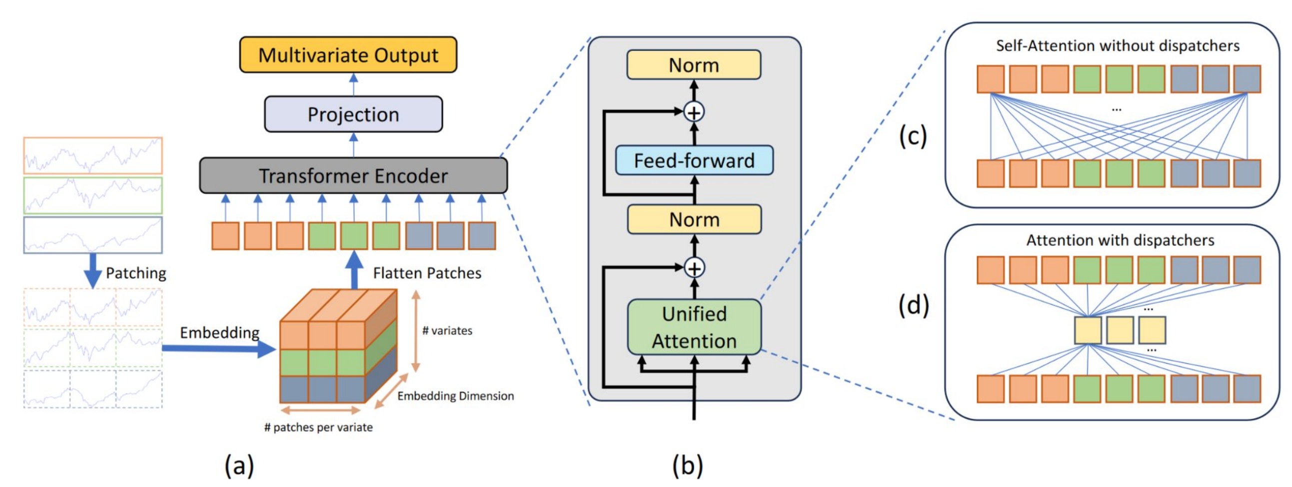
BVAE and Perceiver Architectures

Both BVAE and Perceiver architectures transform data into **latent space** to extract key attributes. The Perceiver architecture enhances this process by introducing a cross-attention module, serving as an **attention bottleneck**. It then iteratively applies cross-attention and self-attention to efficiently preserve temporal trends.



Time Series UniTST Architecture

UniTST slices the time series data into **small patches** and applies **attention** across all patches from all variates, capturing dependencies within the data. An optional intermediate layer, called **"dispatchers"**, can be introduced to reduce complexity and enforce regularization, helping to streamline the model and improve performance.

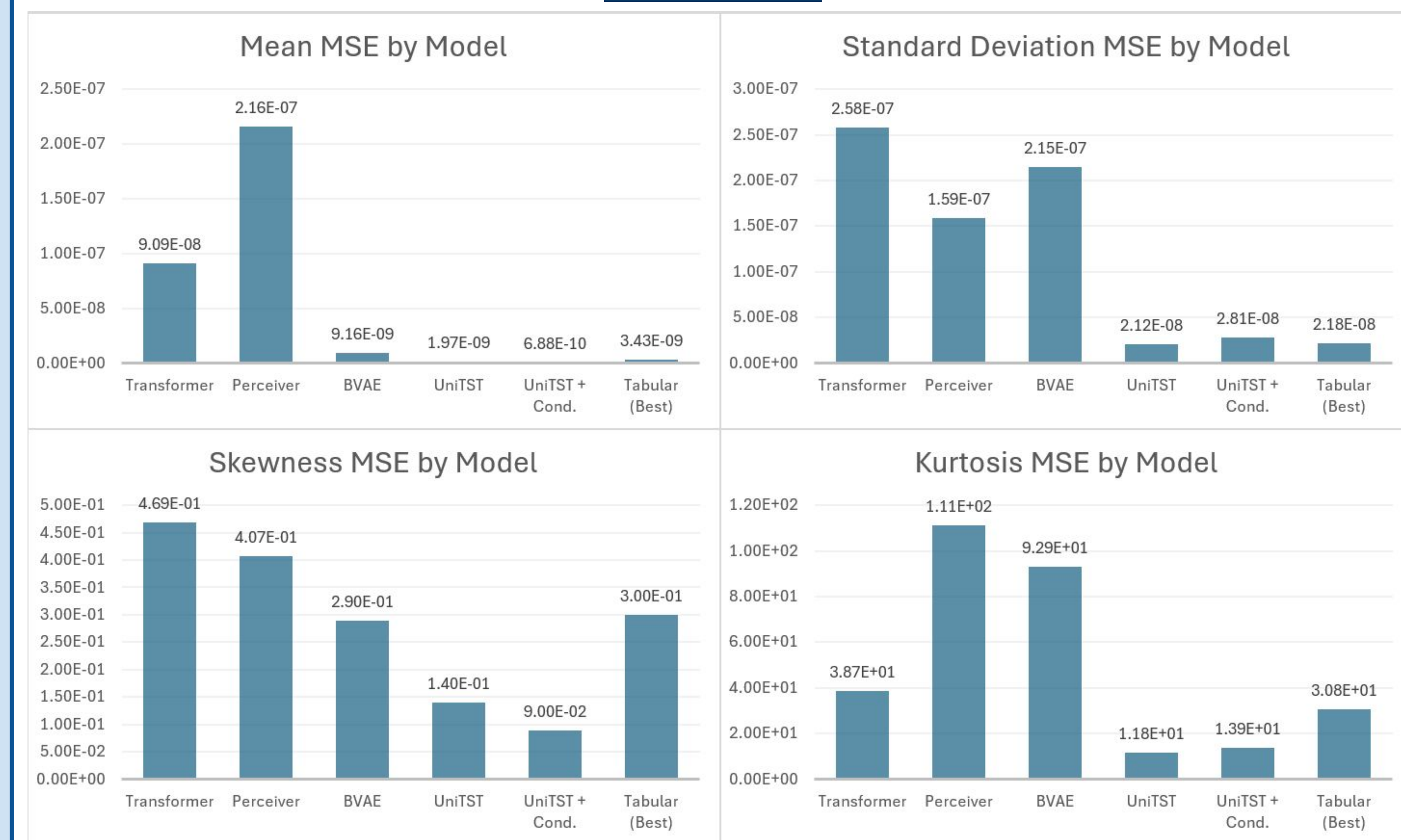


Classifier-Free Conditional Diffusion

In our classifier-free conditional diffusion approach, we extend the standard process of converting timesteps into embeddings by also converting regime labels into embeddings. These **regime embeddings** are summed with data embeddings and timestep embeddings, and then passed through **Adaptive Layer Normalization** at the beginning of each encoder layer. This enables the model to incorporate regime information effectively, enhancing its ability to generate synthetic data that reflects the underlying market conditions.

Results & Further Work

Results



The graphs display the MSE results for each statistical moment across different models, comparing real and synthetic data. The **UniTST** model, both with and without conditional diffusion, outperforms all other models, including the BVAE, which is a more generic architecture designed for various modalities. This superior performance in the UniTST models, originally developed for time series prediction tasks, suggests that time series data may possess **unique characteristics** that are better captured by specialized architectures.

Compared to the best tabular method, our best seq2seq model shows significant improvements: a **79.94% reduction** in MSE for mean, a **2.75% reduction** for standard deviation, a **70% reduction** for skewness, and a **61.86% reduction** for kurtosis.

Conclusion & Further Direction

Our current seq2seq diffusion model successfully captures the four key statistical moments, demonstrating its effectiveness in generating accurate synthetic financial data. However, challenges remain in accurately capturing **autocorrelation** and **covariance** among assets. As a next step, we plan to explore different methods for incorporating **market regimes** and to investigate explicit **loss penalty** terms that encourage the model to learn these complex characteristics.

Advisors: Professor Ali Hirsa, Miao Wang
Group Members: Chenkai Li, Jason Manuel

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science