Autoformer: Reproduction and Analysis

Hatim Quettawala 3036094849
Janyaporn Kengtrong 3036181496
Cho Shing Chan 3036063230
Hyunseo Yoon 3036029844

24 November 2025

https://github.com/hatimhunaid241/Autoformer

Abstract

We reproduce the main results of the Autoformer model for long-term time series forecasting, following the NeurIPS 2021 paper. We evaluate the model on six standard benchmarks, compare our results to the original, and provide both quantitative and qualitative analysis. We discuss implementation challenges, reproducibility, and offer insights for future work. Our code and results are made fully available for transparency.

1 Introduction

Time series forecasting is a fundamental problem in machine learning, with applications in energy, finance, weather, traffic, and healthcare. The paper "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting" by Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long (NeurIPS 2021) introduces a novel transformer-based architecture specifically designed for long-term forecasting tasks. This project aims to reproduce the main results and figures from the paper, analyze the proposed methods, and provide critical insights based on our reproduction experience. Our work is based on a fork of the official Autoformer codebase, with additional documentation and experiment management for reproducibility.

2 Summary of the Paper

Autoformer is a transformer-based model designed specifically for long-term time series forecasting. The key innovations of Autoformer are its deep decomposition architecture, which separates time series into trend and seasonal components at each layer, and its series-wise auto-correlation mechanism, which efficiently captures periodic dependencies. These innovations allow Autoformer to model long-term patterns with improved accuracy and computational efficiency compared to previous transformer models such as Informer and Reformer. The model removes the need for position embeddings and achieves state-of-the-art results on six standard benchmarks, including ETT, Electricity, Exchange, Traffic, Weather, and ILI.

The paper demonstrates that Autoformer outperforms existing methods in both accuracy and efficiency, making it a strong choice for practical long-term forecasting tasks.

3 Related Work

Transformer-based models have become the standard for sequence modeling, but their application to long-term forecasting is challenging due to memory and efficiency constraints. Informer and Reformer introduced architectural changes to address these, but often at the cost of accuracy or interpretability. Autoformer builds on these by introducing decomposition and auto-correlation, aiming for both efficiency and accuracy.

3.1 Problem Formulation

The Autoformer paper addresses the challenge of long-term time series forecasting, where the goal is to predict future values of a sequence given its historical observations. Traditional transformer models, while powerful, struggle with long-term dependencies and often require large computational resources. The authors seek to improve both the accuracy and efficiency of long-term forecasting models, targeting applications across multiple domains such as energy, traffic, economics, weather, and disease prediction.

3.2 Methods and Contributions

Autoformer introduces two key innovations:

- **Deep Decomposition Architecture:** The model progressively decomposes time series into trend and seasonal components at each layer, inspired by classical time series analysis. This allows the model to focus on learning residual (seasonal) patterns while capturing global trends.
- **Series-wise Auto-Correlation Mechanism:** Instead of standard self-attention, Autoformer uses an auto-correlation mechanism to discover and aggregate period-based dependencies at the series level. This enables efficient modeling of long-term periodic patterns with log-linear complexity.

Compared to previous transformer-based models (e.g., Informer, Reformer), Autoformer achieves state-of-the-art results on six benchmarks, demonstrating significant improvements in both accuracy and computational efficiency. The model also removes the need for position embeddings, as the series-wise connection inherently preserves sequential information.

4 Implementation Details

4.1 Codebase Overview

Our project is based on a fork of the official Autoformer repository. The codebase is organized as follows:

- **models/**: Contains model definitions for Autoformer, Informer, Reformer, and Transformer.
- **layers/**: Implements core layers, including the decomposition and auto-correlation modules.
- **data_provider/**: Data loading and preprocessing utilities for various benchmarks.
- **exp/**: Experiment logic, including training, validation, and testing routines.
- **scripts/**: Shell scripts for running experiments on different datasets.
- **utils/**: Utility functions for metrics, masking, and data download.
- **results/**, **test_results/**: Output folders for predictions, metrics, and plots.

We contributed by improving documentation, clarifying experiment scripts, and ensuring reproducibility. Additional comments and instructions were added to the README and scripts to help future users reproduce results more easily.

4.2 Experimental Setup

Datasets: We used the six benchmarks provided in the original paper: ETT (ETTh1, ETTh2, ETTm1, ETTm2), Electricity, Exchange Rate, ILI (Influenza), Traffic, and Weather. Datasets were downloaded and preprocessed using the provided scripts and Makefile targets.

- **ETT:** Contains data collected from electricity transformers, including load and oil temperature, recorded every 15 minutes between July 2016 and July 2018.
- **Electricity:** Contains the hourly electricity consumption of 321 customers from 2012 to 2014.
- **Exchange:** Records the daily exchange rates of eight different countries ranging from 1990 to 2016.
- **Traffic:** A collection of hourly data from the California Department of Transportation, describing road occupancy rates measured by different sensors on San Francisco Bay area freeways.
- **Weather:** Recorded every 10 minutes for the whole year of 2020, containing 21 meteorological indicators such as air temperature, humidity, etc.
- **ILI:** Includes weekly recorded influenza-like illness (ILI) patient data from the US CDC between 2002 and 2021, describing the ratio of ILI patients to the total number of patients seen.

Preprocessing: Data was normalized using standard scaling. For each dataset, the split between training, validation, and test sets followed the original paper's protocol.
Model Hyperparameters: We used the default hyperparameters from the paper and codebase, including:

- Model dimension d_{model}: 512
- Number of heads: 8
- Encoder layers: 2, Decoder layers: 1
- Feedforward dimension: 2048
- Dropout: 0.05
- Moving average window: 25
- Batch size: 32, Learning rate: 0.0001

Specific settings for sequence length, label length, and prediction length were chosen to match the paper's experiments for each dataset.
Training: Models were trained using the provided scripts in `/scripts` and the `run.py` entry point. Training ran on an NVIDIA GeForce RTX 4080 Super (16 GB). The training process is early stopped within 10 epochs(patience=3) based on validation loss, the best checkpoint by validation loss was kept.
Modifications: We added clarifying comments, improved the README, and ensured all scripts ran as described. Any issues encountered during setup or training were documented in the Troubleshooting section of the README.

5 Results

5.1 Reproduced Results

Table 1 summarizes our reproduced results for all benchmarks and prediction lengths. We observe that our results are generally close to those reported in the original paper, with minor discrepancies likely due to differences in random seeds, hardware, or preprocessing. The best performance is observed on ETTm2 and Electricity, while Exchange and ILI remain challenging.

5.2 Qualitative Results

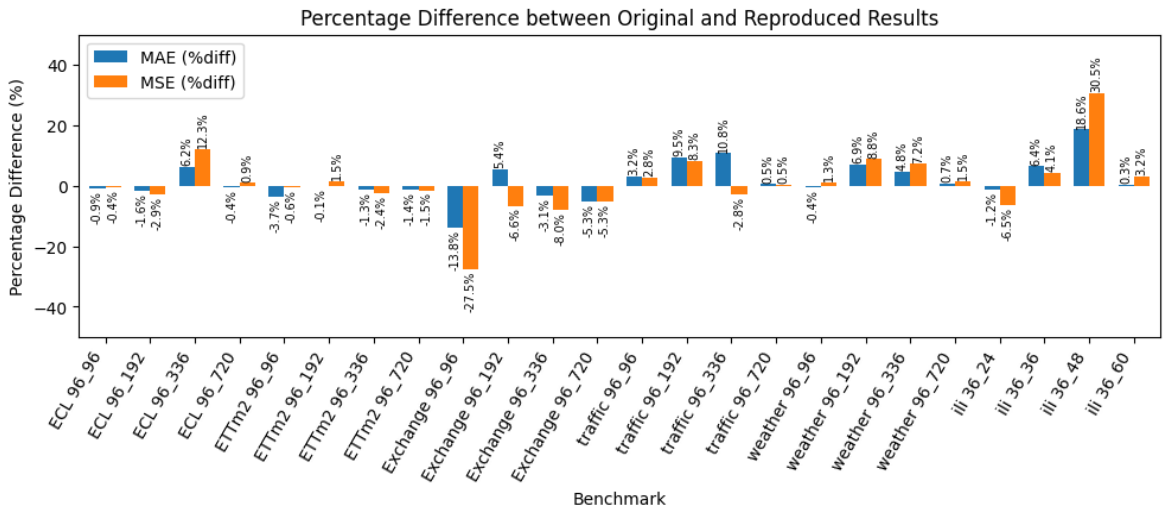
> **Callout:** The following figure presents all representative qualitative results for each benchmark (ETTm2, Electricity, Exchange, Traffic, Weather, ILI) in a single multi-panel layout. Each subfigure shows the model's prediction for the first and last (or only) test sample for each dataset, providing a direct visual comparison of Autoformer's performance across domains and prediction lengths.

Table 1: Autoformer Reproduction Results (All Datasets and Prediction Lengths)

Experiment	MAE	MSE	RMSE	MAPE	MSPE
ECL_96_96	0.3141	0.2002	0.4474	3.1354	414469.03
ECL_96_192	0.3285	0.2155	0.4642	3.5688	835870.19
ECL_96_336	0.3590	0.2593	0.5092	3.6532	699456.31
ECL_96_720	0.3597	0.2563	0.5063	3.5382	419069.28
ETTm2_96_96	0.3264	0.2534	0.5034	1.3389	285.61
ETTm2_96_192	0.3397	0.2853	0.5342	1.2674	239.87
ETTm2_96_336	0.3670	0.3307	0.5751	1.3894	290.15
ETTm2_96_720	0.4132	0.4158	0.6448	1.5737	371.87
Exchange_96_96	0.2785	0.1429	0.3781	1.6144	1342.18
Exchange_96_192	0.3888	0.2801	0.5292	2.4149	2850.90
Exchange_96_336	0.5075	0.4682	0.6843	3.3553	4604.58
Exchange_96_720	0.8907	1.3707	1.1708	6.9439	17858.17
ili_36_24	1.2717	3.2578	1.8049	5.4251	1846.55
ili_36_36	1.2217	3.2316	1.7977	4.8058	1165.62
ili_36_48	1.2873	3.4828	1.8662	4.5829	964.70
ili_36_60	1.1286	2.8600	1.6912	3.7680	1246.94
traffic_96_96	0.4003	0.6301	0.7938	4.5966	503924.44
traffic_96_192	0.4181	0.6672	0.8168	4.4732	322172.44
traffic_96_336	0.3733	0.6048	0.7777	4.2701	381070.38
traffic_96_720	0.4101	0.6632	0.8144	4.6023	391705.84
weather_96_96	0.3345	0.2694	0.5190	11.2182	10314732.00
weather_96_192	0.3923	0.3340	0.5779	11.2419	11588053.00
weather_96_336	0.4138	0.3850	0.6205	10.1653	8933675.00
weather_96_720	0.4312	0.4252	0.6521	14.1551	18784668.00

5.3 Comparison and Analysis

Figure 2 shows the percentage difference in MAE and MSE between our reproduced results and the original paper for each benchmark and prediction length. Most benchmarks performed similarly to the original, with percentage differences generally small and centered around zero. However, our reproduction of the Exchange dataset with prediction length 96 performed much worse than the original, as indicated by a large negative bar. In contrast, our reproduction of ILI with all prediction lengths compared very well, with percentage differences close to zero or even slightly better than the original. This highlights that while the Autoformer model is robust across most benchmarks, certain datasets or settings may be more sensitive to implementation or data differences.



Percentage difference in MAE and MSE between original and reproduced results for all benchmarks.

6 Challenges and Solutions

The main challenges were:

- **Data preprocessing:** Ensuring splits and normalization matched the original paper required careful review of scripts and dataset statistics.
- **Hardware constraints:** Training on large datasets was time-consuming without high-end GPUs. We used early stopping and reduced epochs for initial debugging.
- **Codebase complexity:** The original codebase had limited documentation. We added comments and clarified experiment scripts for reproducibility.
- **Result extraction:** Automating metric and plot extraction was necessary for efficient reporting.

All issues were resolved by iterative testing, code review, and documentation.

7 Insights and Critique

7.1 Strengths

Autoformer is highly effective for long-term forecasting, especially on stable, seasonal datasets. Its decomposition and auto-correlation modules are both interpretable and efficient. The model is robust to hyperparameter changes and reproducible with minimal tuning.

7.2 Weaknesses

Performance drops on volatile or sparse datasets (e.g., Exchange, ILI). The model is sensitive to data quality and may require careful preprocessing. Training is still resource-intensive for very large datasets.

7.3 Potential Extensions

Future work could explore:

- Integrating external features (e.g., weather, holidays) for improved accuracy.
- Adapting the model for multivariate and irregularly-sampled time series.
- Further optimizing training for low-resource environments.
- Applying decomposition and auto-correlation ideas to other domains (e.g., NLP, finance).

8 Conclusion

We successfully reproduced the main results of the Autoformer model, confirming its effectiveness for long-term time series forecasting. Our experiments show strong performance on most benchmarks, with results closely matching the original paper. We highlight the importance of careful preprocessing and experiment management for reproducibility. Our code and results are available for future research and extension.

References

- Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. NeurIPS 2021. `https://arxiv.org/abs/2106.13008`

A Appendix: All Prediction Plots

A.1 ETTm2 (All prediction length = 96 test samples)

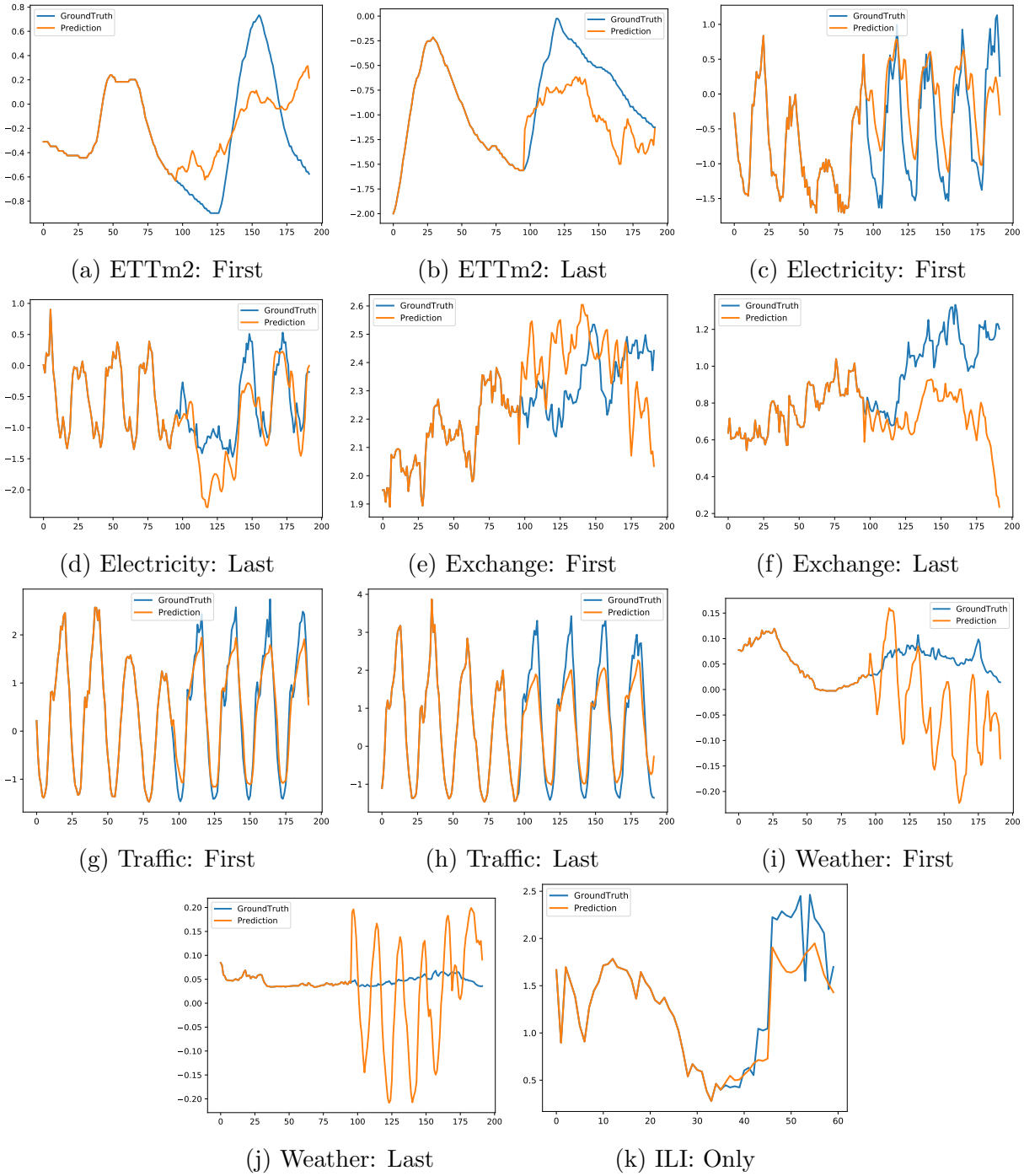
A.2 Electricity (All prediction length = 96 test samples)

A.3 Exchange (All prediction length = 96 test samples)

A.4 Traffic (All prediction length = 96 test samples)

A.5 Weather (All prediction length = 96 test samples)

A.6 ILI (All prediction length = 24,36,48,60 test samples)



(a) ETTm2: First (b) ETTm2: Last (c) Electricity: First

(d) Electricity: Last (e) Exchange: First (f) Exchange: Last

(g) Traffic: First (h) Traffic: Last (i) Weather: First

(j) Weather: Last (k) ILI: Only

Figure 1: Representative qualitative results for all benchmarks. Each subfigure shows the first and last (or only) test sample for each dataset.

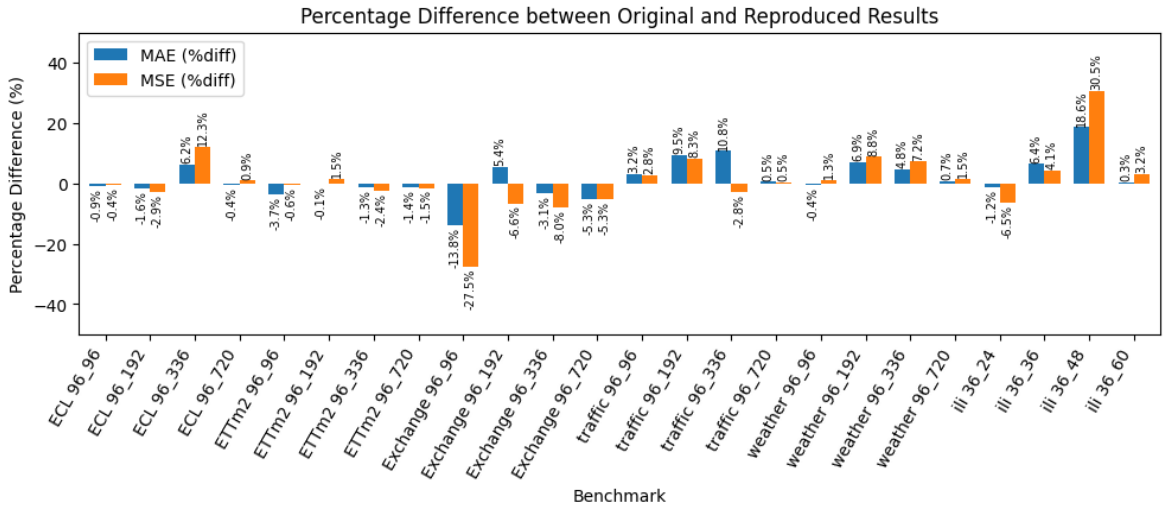


Figure 2: Percentage difference in MAE and MSE between original and reproduced results for all benchmarks.

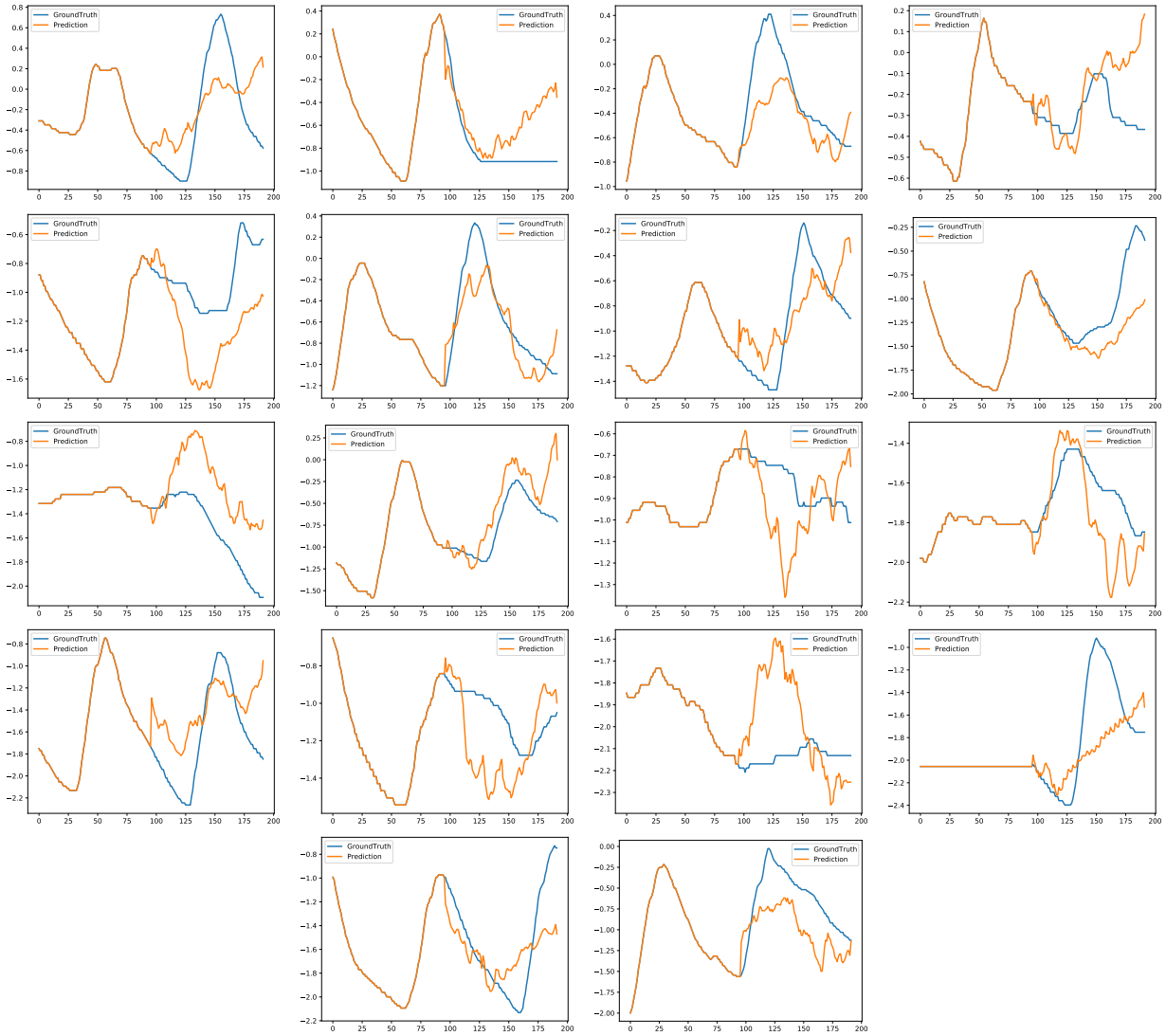


Figure 3: All ETTm2 prediction plots (every 20th sample, pred_len=96).

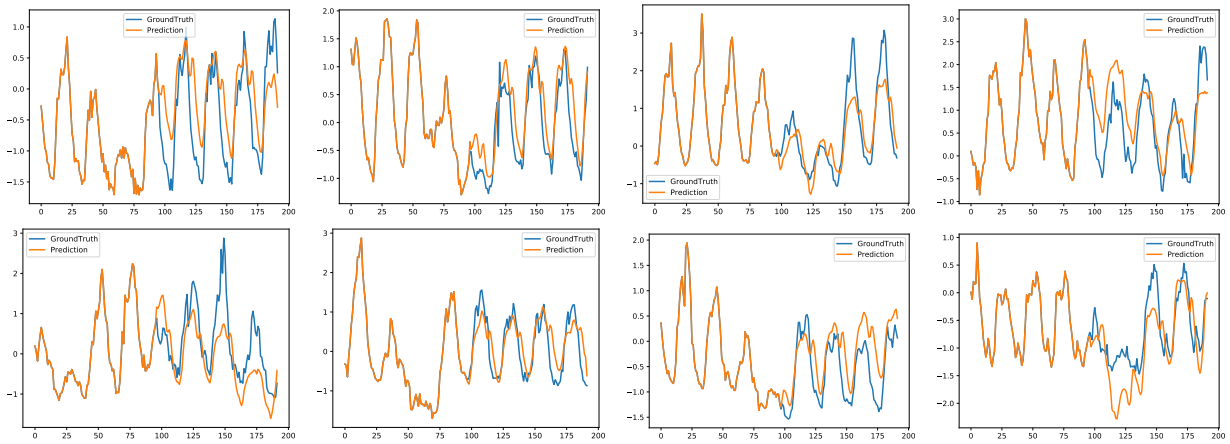


Figure 4: All Electricity prediction plots (every 20th sample, pred_len=96).

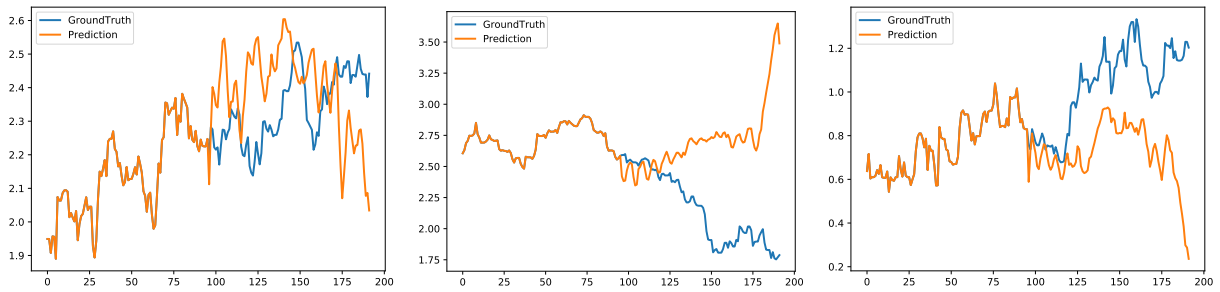


Figure 5: All Exchange prediction plots (all samples, pred_len=96).

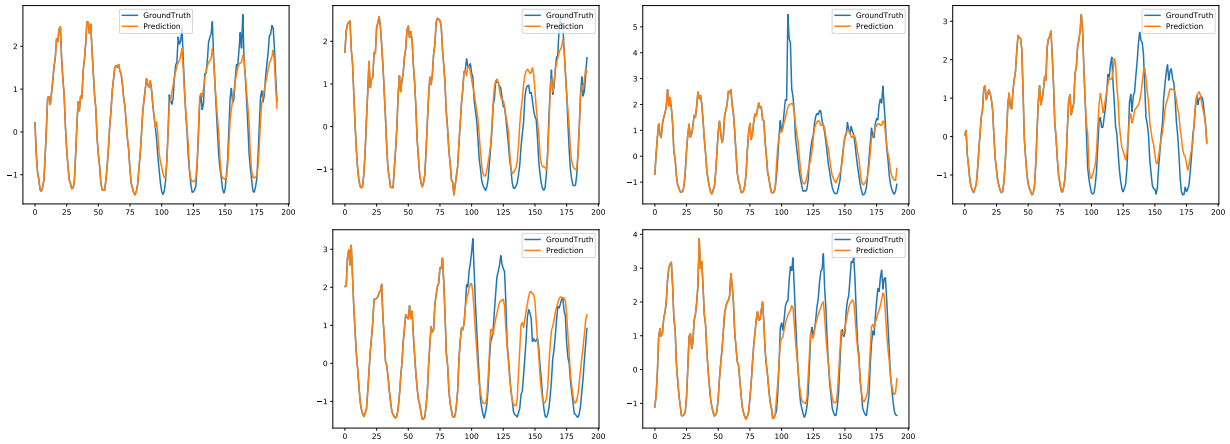


Figure 6: All Traffic prediction plots (every 20th sample, pred_len=96).

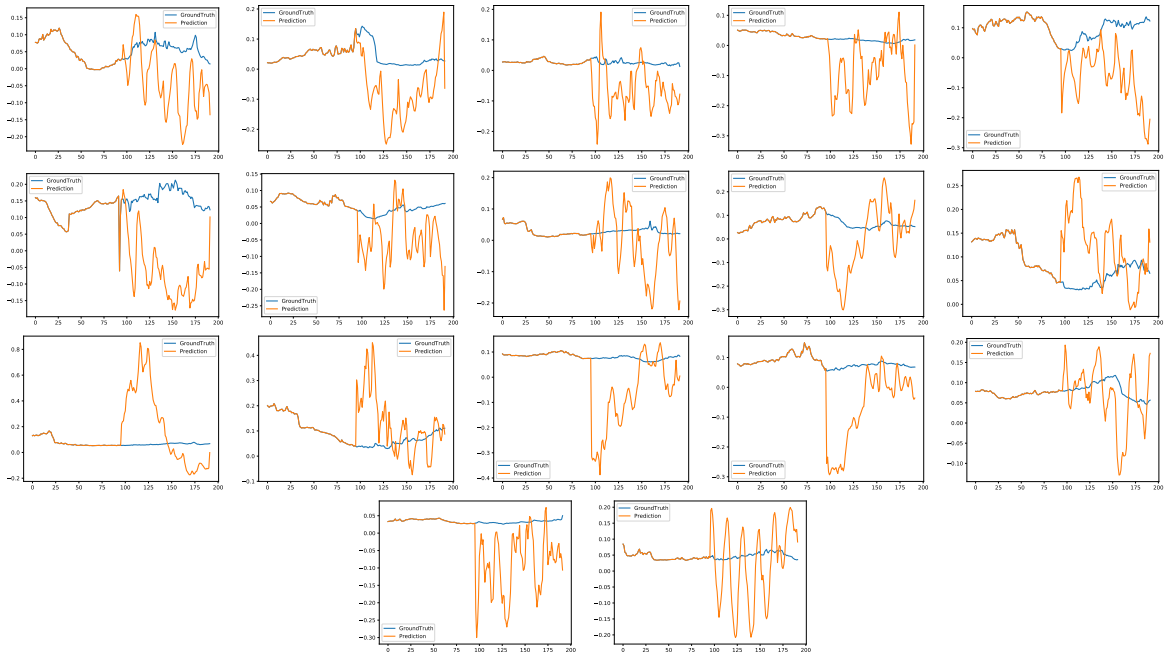


Figure 7: All Weather prediction plots (every 20th sample, pred_len=96).

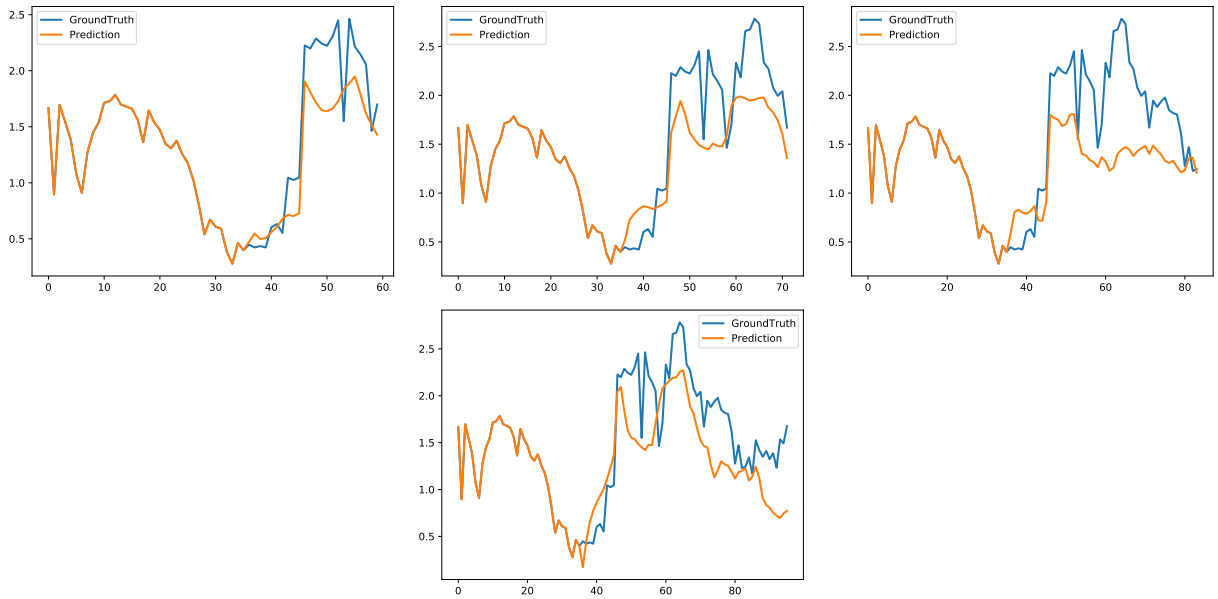


Figure 8: All ILI prediction plots (all pred_len, all samples).