University of Science and Technology of China

TimeDART: A Diffusion Autoregressive Transformer for Self-Supervised Time Series Representation

Daoyu Wang, Mingyue Cheng*, Zhiding Liu, Qi Liu

The 42nd International Conference on Machine Learning

June 14th, 2025

Overview

- Statement of the Problem
- Motivation
- The Proposed TimeDART
- Experiments
- Analysis
- Conclusion

Statement of the Problem

Questions

- What is Time Series data?
- What is Self-Supervised Time Series Representation Learning?
- How to evaluate this task?

Answers

- Time series data is a sequence of data points recorded in chronological order, defined by its sequential and temporal characteristics.
- This pre-training approach learns transferable representations from unlabeled time series data by generating supervision from the data's own structure.
- This task is evaluated by fine-tuning the pre-trained model on downstream tasks, such as forecasting and classification.

Motivation

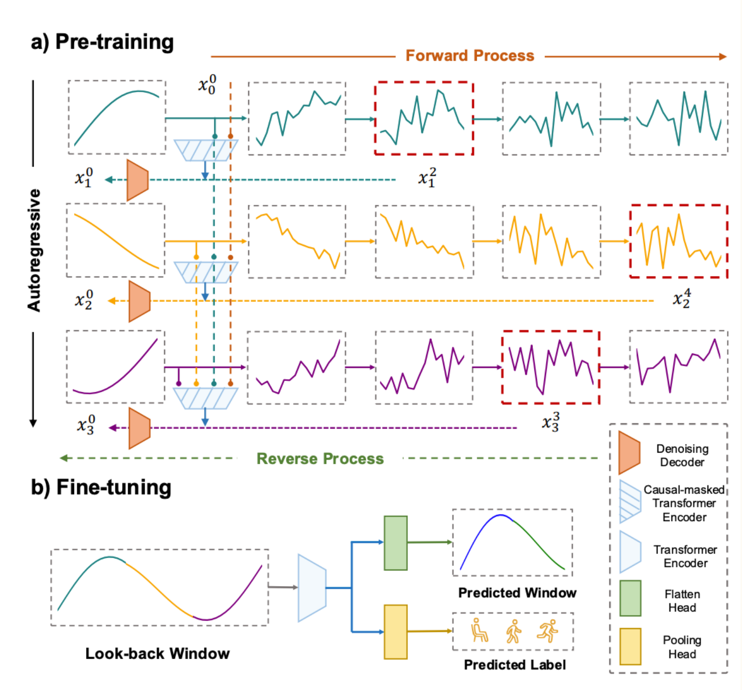
The Problem:
Existing self-supervised methods have limitations:

- **Masked Reconstruction**: Excel at learning patterns but can have inconsistencies between pre-training and fine-tuning.
- **Contrastive Discrimination**: Are great for sequence-level distinctions but may miss fine-grained temporal details.
- **Autoregressive Prediction**: Naturally model time flow but tend to overfit noise and make an overly simplistic Gaussian distribution assumption.

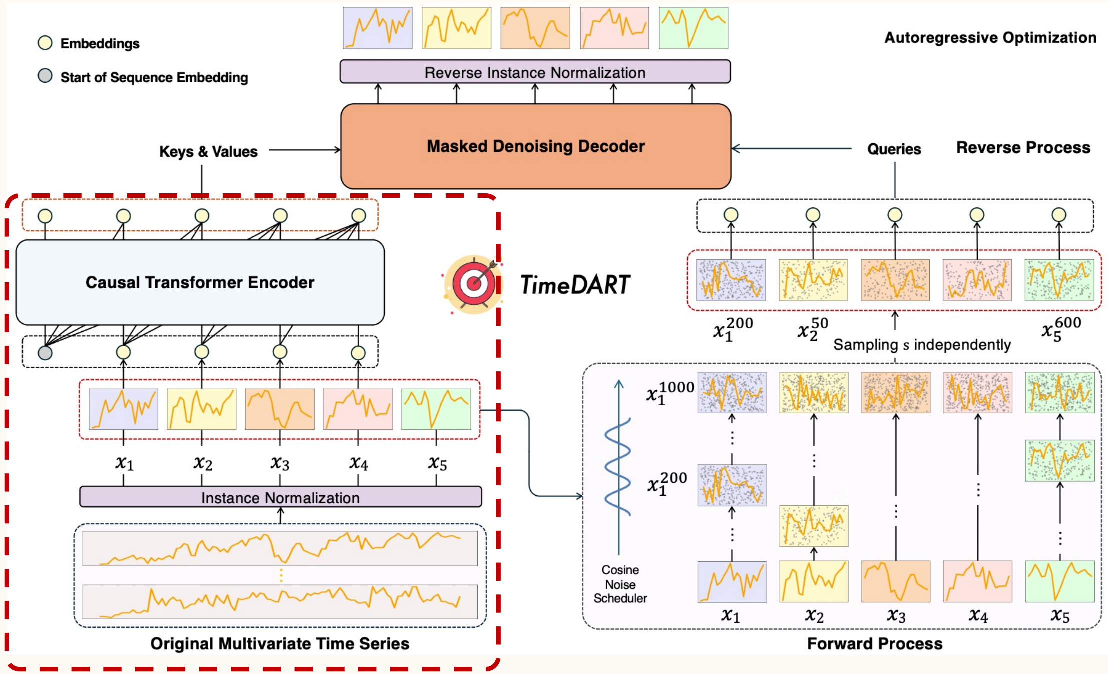
Core Insight

Our Core Idea:
We unify two powerful generative model to learn more transferable representations, TimeDART combines:

- **Autoregressive Modeling**: To capture long-term global dynamic evolution.
- **Denoising Diffusion Process**: To capture subtle, fine-grained local evolution.



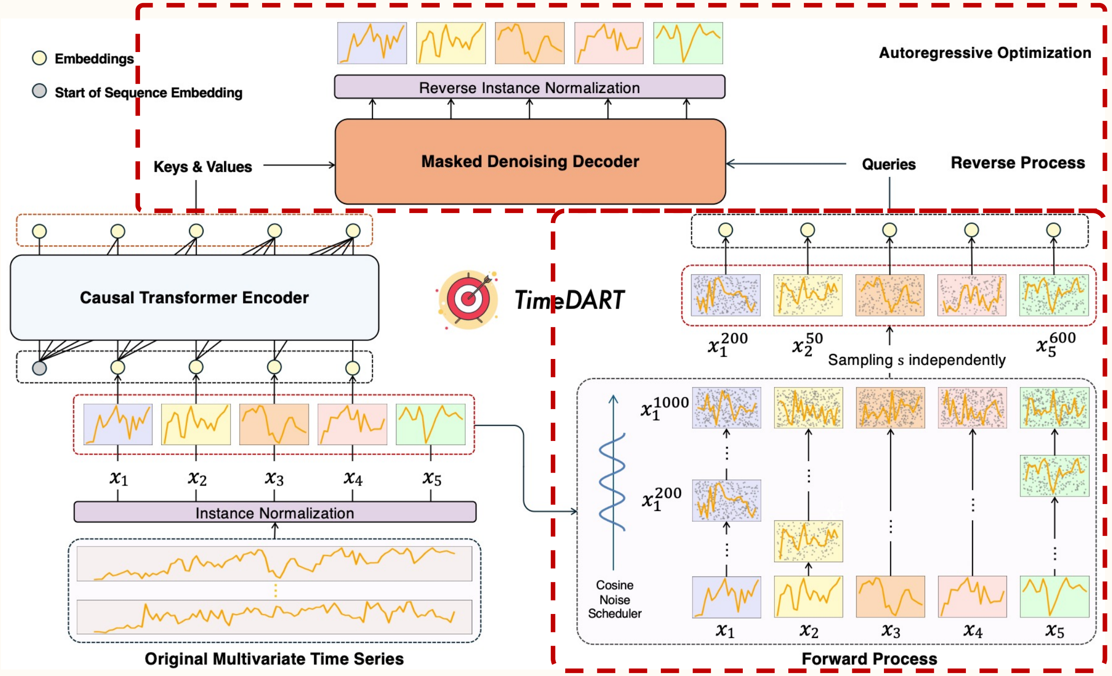
TimeDART (I)



Causal Transformer Encoder (Global Trend)

- The input time series is divided into non-overlapping patches.
- A **Causal Transformer** processes these patches, ensuring it only attaches past information.

TimeDART (II)



Patch-level Diffusion and Denoising (Local Patterns)

- We independently add noise to each patch.
- A **Denoising Decoder** then uses the contextualized output from the encoder to reconstruct the original, clean patch from its noisy version.

TimeDART (III)

$$\mathcal{L}_{\text{mse}} \propto \sum_{j=1}^{N} ||x_j^0 - \text{Projector}(f(\boldsymbol{z}_{1:j-1}^{in}))||^2.$$

$$\frac{1}{2\sigma^2}||x_j^0 - \text{Projector}(f(\boldsymbol{z}_{1:j-1}^{in}))||^2 = -\log \mathcal{N}(x_j^0; \text{Projector}(f(\boldsymbol{z}_{1:j-1}^{in})), \sigma^2) + C,$$

$$\boxed{\mathcal{L}_{\text{diff}} = \sum_{j=1}^{N} \mathbb{E}_{\epsilon, q(x_j^0)} \left[||x_j^0 - g(\hat{z}_j^{in},\ f(\boldsymbol{z}_{1:j-1}^{in}))||^2 \right].}$$

The Self-Supervised Objective

- Our diffusion loss trains the model to denoising each patch, guided by the autoregressive history.
- We avoid to making an overly simplistic Gaussian distribution assumption that the pure autoregressive objective has.

Experiments

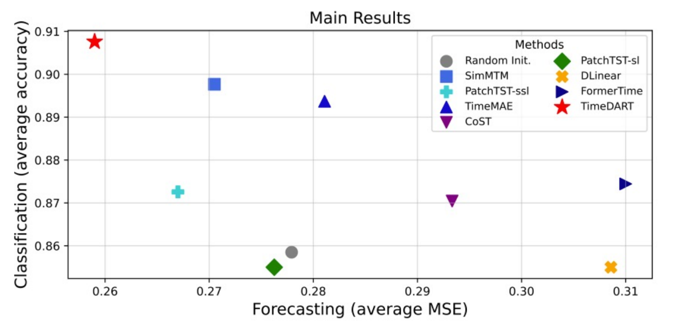


Figure 3: Comparison between TimeDART and baselines for the forecasting task (MSE ↓) across forecasting datasets on the x-axis and the classification task (Accuracy ↑) across classification datasets on the y-axis.

Experimental Setup

- Evaluated on 9 public datasets for forecasting and classification tasks.
- Compared against strong self-supervised and supervised baselines.

Experiments

Table 2: Multivariate time series forecasting results. All results are averaged MSE and Mean Absolute Error (MAE) from 4 different predicted windows of $\{12, 24, 36, 48\}$ for PEMS datasets and $\{96, 192, 336, 720\}$ for others. The best results are in **bold** and the second best are underlined. Full results are detailed in Appendix D.

Methods	Ours				Self-supervised								Supervised			
	TimeDART		Random Init.		SimMTM		PatchTST		TimeMAE		CoST		PatchTST		DLinear	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.411	**0.426**	0.439	0.444	**0.409**	0.428	0.433	0.437	0.434	0.445	0.465	0.464	0.427	0.435	0.439	0.449
ETTh2	**0.346**	**0.387**	0.358	0.396	0.353	0.390	0.354	0.393	0.402	0.431	0.399	0.427	0.357	0.395	0.458	0.459
ETTm1	0.344	**0.379**	0.351	0.383	0.348	0.385	**0.342**	0.380	0.350	0.383	0.356	0.385	0.362	0.388	0.361	0.383
ETTm2	**0.257**	**0.316**	0.269	0.323	0.263	0.320	0.272	0.327	0.270	0.326	0.282	0.343	0.270	0.329	0.281	0.343
Electricity	0.163	**0.254**	0.177	0.277	**0.162**	0.256	0.163	0.255	0.196	0.309	0.215	0.295	0.167	0.260	0.168	0.265
Traffic	**0.388**	**0.263**	0.410	0.277	0.392	0.264	0.404	0.272	0.410	0.275	0.435	0.362	0.421	0.284	0.435	0.297
Weather	**0.226**	0.263	0.231	0.268	0.230	0.271	0.227	**0.262**	0.227	0.265	0.242	0.282	**0.226**	0.263	0.246	0.298
Exchange	**0.359**	**0.405**	0.440	0.450	0.451	0.455	0.376	0.413	0.427	0.446	0.456	0.455	0.379	0.414	0.393	0.425
PEMS03	**0.152**	**0.257**	0.164	0.266	0.158	0.260	0.156	0.261	0.165	0.269	0.169	0.273	0.178	0.288	0.277	0.373
PEMS04	**0.133**	**0.245**	0.145	0.255	0.143	0.253	0.139	0.249	0.144	0.256	0.147	0.262	0.149	0.266	0.290	0.381
PEMS07	**0.128**	**0.232**	0.138	0.243	0.131	0.236	0.132	0.237	0.137	0.241	0.139	0.245	0.149	0.253	0.322	0.387
PEMS08	**0.201**	**0.282**	0.213	0.293	0.206	0.286	0.206	0.287	0.211	0.292	0.215	0.295	0.230	0.295	0.359	0.402

Table 4: Multivariate time series classification results. Results are are reported as Accuracy (Acc.) and Macro-F1 (F1). The best results are in **bold** and the second best are underlined.

Methods	Ours				Self-supervised								Supervised	
	TimeDART		Random Init.		SimMTM		PatchTST		TimeMAE		CoST		FormerTime	
Metric	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
HAR	**0.9247**	**0.9286**	0.8738	0.8723	0.9200	0.9220	0.8789	0.8773	0.9204	0.9248	0.8997	0.8927	0.8816	0.8878
Epilepsy	**0.9712**	**0.9698**	0.9265	0.9237	0.9565	0.9543	0.9312	0.9234	0.9459	0.9584	0.9198	0.9156	0.9315	0.9341
EEG	**0.8269**	0.5983	0.7752	0.5138	0.8165	**0.6123**	0.8076	0.5460	0.8148	0.5787	0.7918	0.5314	0.8102	0.5658

Key Results:

- **Forecasting**: Achieving state-of-the-art results on 83.3% of the metrics, with a 6.8% MSE reduction over random initialization.
- **Classification**: Surpassed all baselines, including specialized supervised methods, improving accuracy by 5.7%.

Analysis

Why does it work?

- **Ablation Study**: Removing either the autoregressive part or the diffusion part causes a major drop in performance, proving both are essential.
- **Different Backbone**: TCN as backbone also works!

Table 5: Performance of TCN as backbone. Average MSE and MAE from 4 different predicted windows for forecasting while Accuracy and Macro-F1 for classification task.

Method	TCN		Random Init.		Transformer	
Forecasting	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	0.349	0.396	0.357	0.403	**0.346**	**0.387**
ETTm2	0.263	0.323	0.269	0.326	**0.257**	**0.316**
Electricity	0.165	**0.254**	0.177	0.278	**0.163**	**0.254**
PEMS04	0.134	0.246	0.145	0.256	**0.133**	**0.245**
Classification	Acc.	F1	Acc.	F1	Acc.	F1
HAR	**0.9252**	**0.9250**	0.8842	0.8901	0.9247	0.9249
Epilepsy	**0.9723**	0.9689	0.9525	0.9513	0.9712	**0.9698**

Table 6: The results of ablation study. Average MSE and MAE from 4 different predicted windows for forecasting while Accuracy and Macro-F1 for classification task.

Method	TimeDART		*w/o* AR		*w/o* Diff		*w/o* AR-Diff	
Forecasting	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	**0.346**	**0.387**	0.365	0.399	0.352	0.391	0.364	0.398
ETTm2	**0.257**	**0.316**	0.281	0.338	0.265	0.322	0.285	0.346
Electricity	**0.163**	**0.254**	0.193	0.304	0.164	0.255	0.190	0.299
PEMS04	**0.133**	**0.245**	0.144	0.255	0.145	0.256	0.149	0.260
Classification	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
HAR	**0.9247**	**0.9286**	0.8966	0.8994	0.9002	0.9028	0.8785	0.8756
Epilepsy	**0.9712**	**0.9698**	0.9505	0.9518	0.9598	0.9586	0.9486	0.9472

Analysis

Deeper analysis:

- **Few-Shot:** Fine-tuned on only 10% of data, TimeDART beats supervised models trained on 100% of the data.
- **Linear Probing:** Just training a linear head on top of the frozen pre-trained encoder also yields strong results, confirming the high quality of the learned representations.
- **Handles Extended-Length Inputs:** TimeDART is pre-trained to handle noise, so its performance consistently improves with longer look-back windows, unlike methods that struggle with the noise in longer series.

Conclusion

- We introduce TimeDART. A novel SSL framework that unifies autoregressive modeling and denoising diffusion process.
- It effectively captures both **global trends** and **local patterns.**
- It establishes a new state-of-the-art and learns highly data-efficient representations.

Q&A Session

Thank you for listening!

My Github

SCAN ME

TimeDART Code