



ICML
International Conference
On Machine Learning



清华大学
Tsinghua University

Sundial: A Family of Highly Capable Time Series Foundation Models

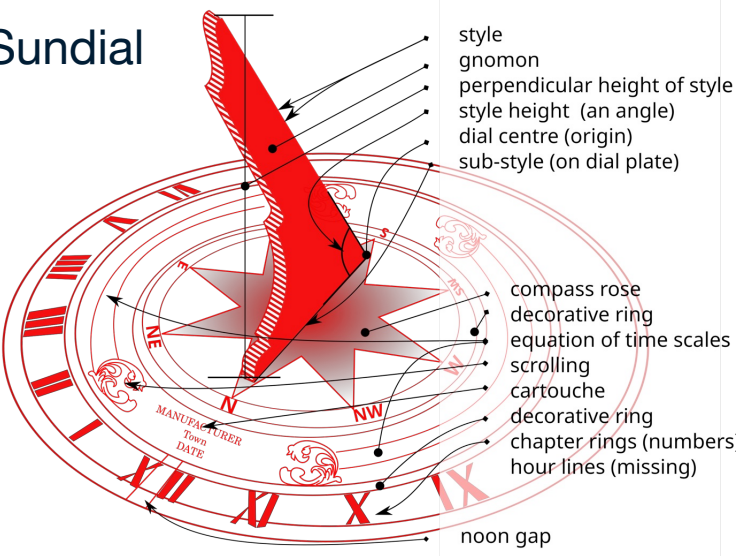
Yong Liu*, Guo Qin*, Zhiyuan Shi, Zhi Chen, Caiyin Yang,
Xiangdong Huang, Jianmin Wang, Mingsheng Long

School of Software, Tsinghua University

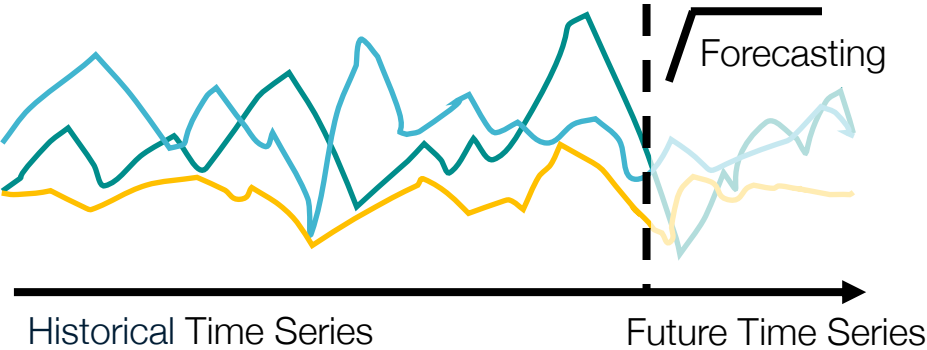
July 16th, 2025

Time Series Forecasting

Sundial



Predict future values based on historical data



Weather Forecasting



Assets Trading

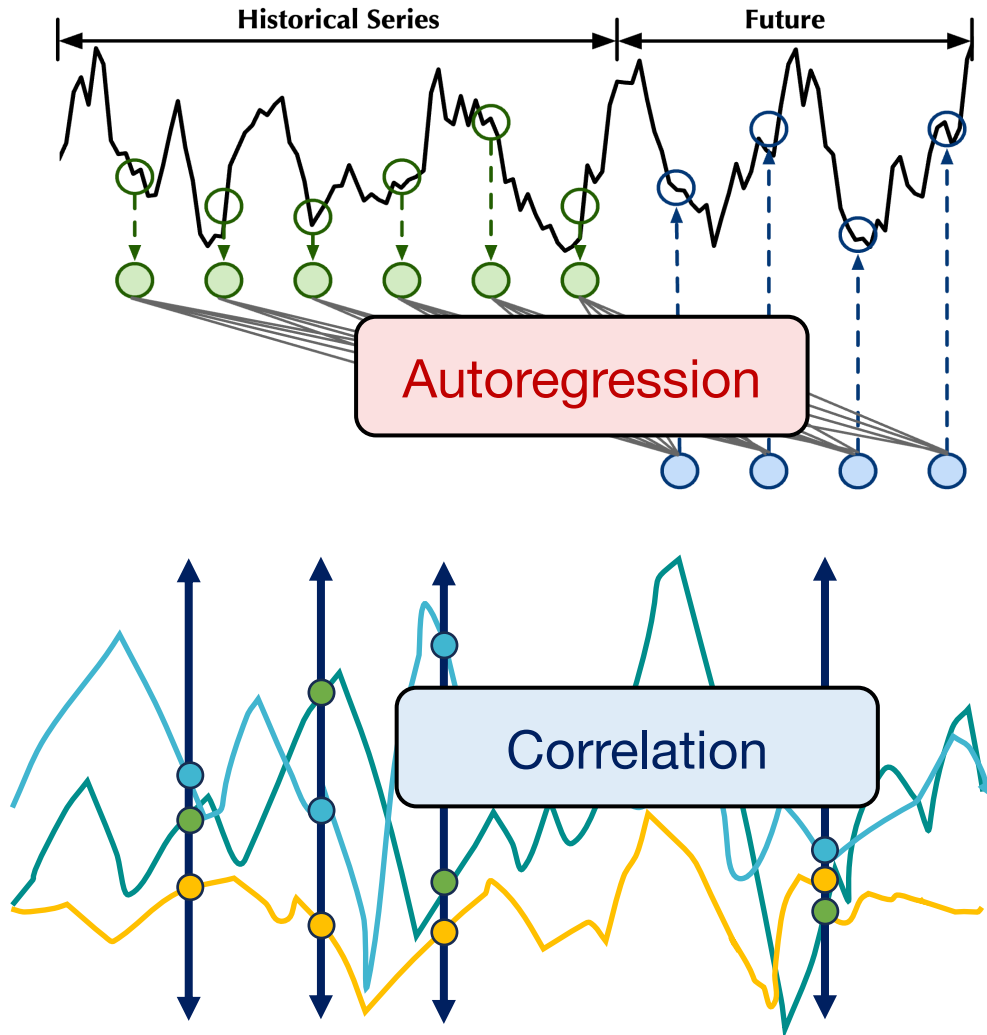


Supply Chain



Healthcare

The Golden Rule: Learn from History



Temporal Dependency (Univariate)

- Auto-Regression, Moving-Average

Variable Correlation (Multivariate)

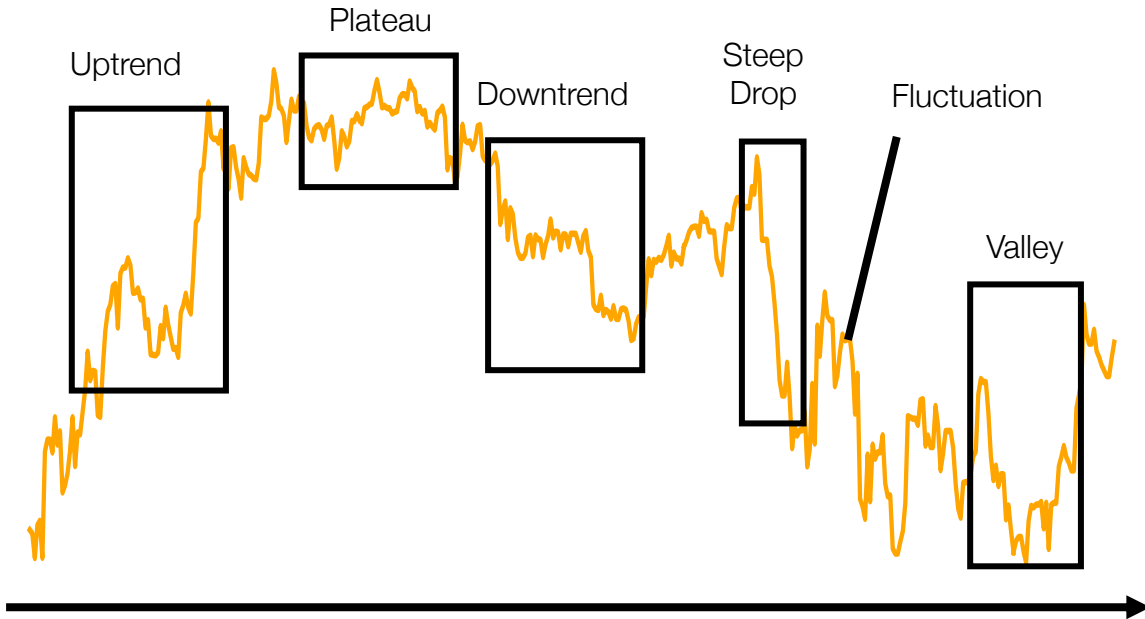
- VAR, Cointegration



Clive Granger

Nobel Prize in Economics

Challenges



Nonlinearity: The linear assumption may not hold for complicated variations

-> Require Good Factors (Features)



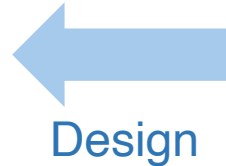
Adaptiveness: Statistical methods need to be fitted separately at different periods

-> Need Large Model Capacity

Deep Learning for Forecasting

Model Architecture

(Linear Layer; MLP; CNN; RNN; GNN
Transformer; State-Space Model...)



Statistics & Theory

(Autoregression; Moving-Average; Stationary:
Normalization, Differencing, Decomposition: STL;
Spectral Analysis: Auto-Correlation, FFT, Wavelets...)

Challenges: Data Hungry...

- Hard to train with **insufficient** data
- Fail to **generalize** to different datasets

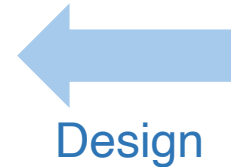
Challenges: Nonlinearity...

- Rely on good features
- Small model capacity

Time Series Foundation Models

Model Architecture

(Linear Layer; MLP; CNN; RNN; GNN
Transformer; State-Space Model...)



Design

Statistics & Theory

(Autoregression; Moving-Average; Stationary:
Normalization, Differencing, Decomposition: STL;
Spectral Analysis: Auto-Correlation, FFT, Wavelets...)

Challenges: Data Hungry...

Challenges: Nonlinearity...



We are still in the early stages...

Time Series

Foundation Models



Pre-training

Generality & Scalability

(Architecture: Transformer, Tokenization: Point/Patch,
Discrete/Continuous; Pre-training: Next-Token Prediction,
Masked Modeling, Generative Modeling; Data Curation...)

Challenges: Heterogeneity, ...

- ✓ Generalizable: zero-shot forecasting
- ✓ Good Factors: feature extractor of time series

Era of
Large
Models

Timeline

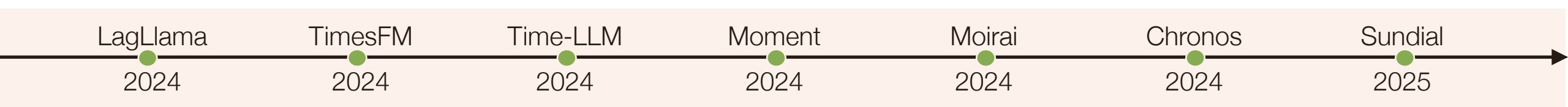
Stage 1: From Statistics/Theories to Deep Models



Stage 2: Unlocking the Capability of Deep Models



Stage 3: Towards Time Series Foundation Models



Not exhaustive. Additional important works, while not enumerated, remain integral to the process.

Heterogeneity

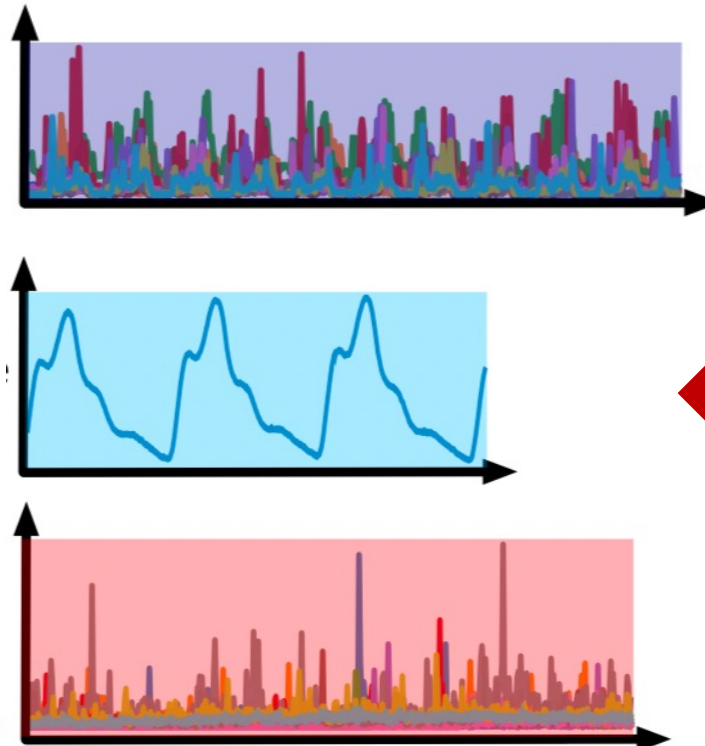
Time Series Are Highly Unstructured Compared To Natural Language

Diverse Shape/Freq./Scale

Dataset	Dim	Frequency
ETTh1, ETTh2	7	Hourly
ETTm1, ETTm2	7	15min
Exchange	8	Daily
Weather	21	10min
ECL	321	Hourly
Traffic	862	Hourly
Solar-Energy	137	10min
PEMS03	358	5min
PEMS04	307	5min
PEMS07	883	5min

Hard for unified pre-training

2D (or more) continuous values



1D discrete token sequences

Natural language processing (NLP) is a subfield of [computer science](#) and especially [artificial intelligence](#). It is primarily concerned with providing computers with the ability to process data encoded in [natural language](#) and is thus closely related to [information retrieval](#), [knowledge representation](#) and [computational linguistics](#), a subfield of [linguistics](#).

Major tasks in natural language processing are [speech recognition](#), [text classification](#), [natural language understanding](#), and [natural language generation](#).

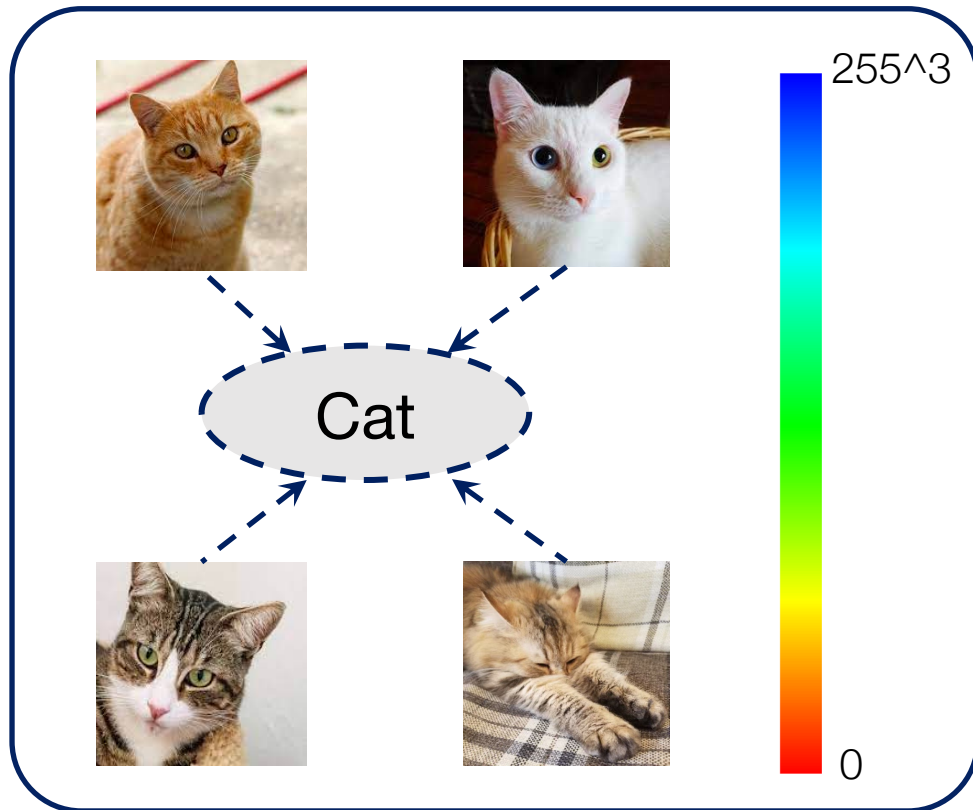
History [edit]

Further information: [History of natural language processing](#)

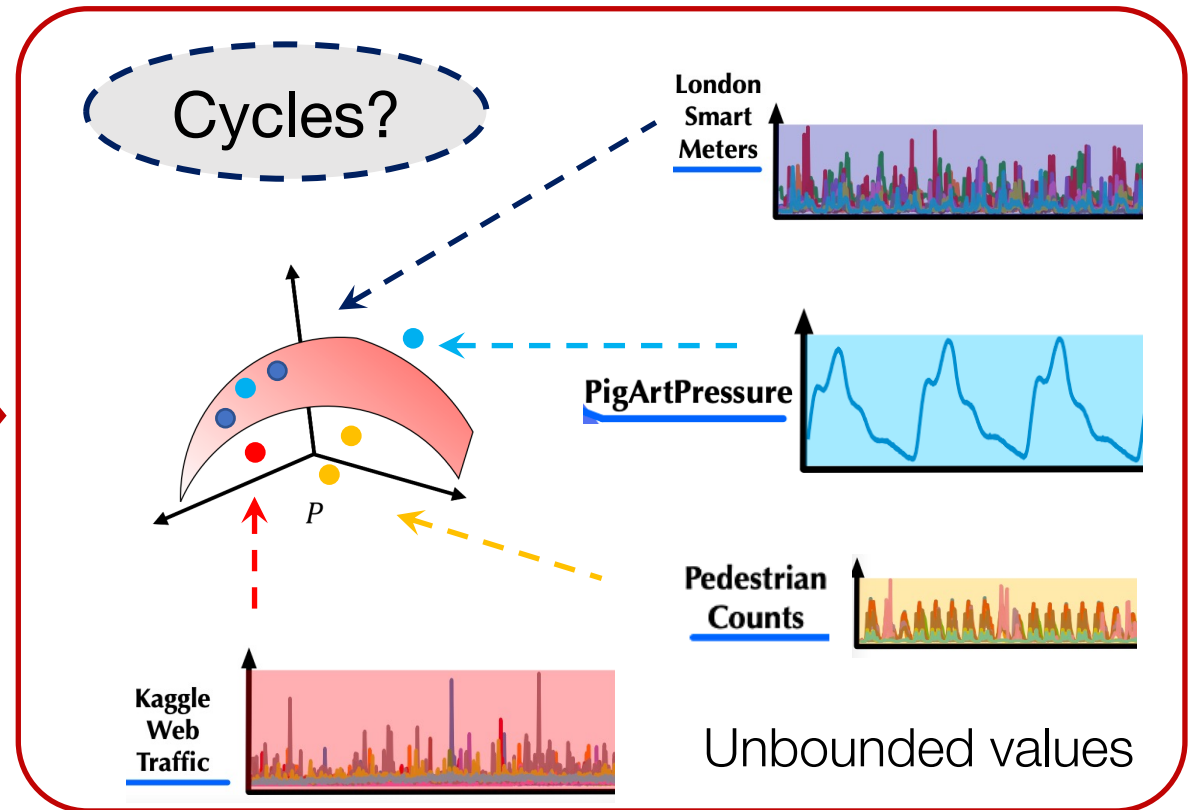
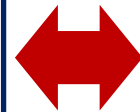
Natural language processing has its roots in the 1950s.^[1] Already in 1950, [Alan Turing](#) published an article titled "[Computing Machinery and Intelligence](#)" which proposed

Heterogeneity

Time Series Can Be More Ambiguous (Distributed) Compared To Images

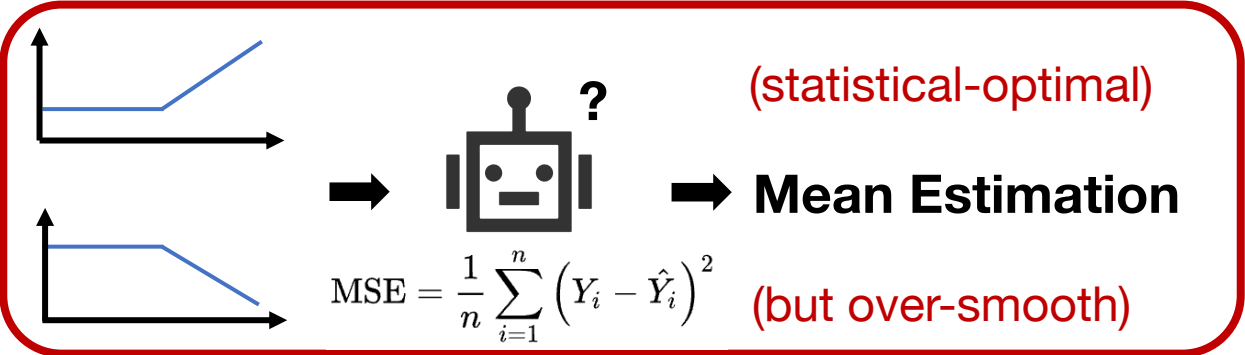
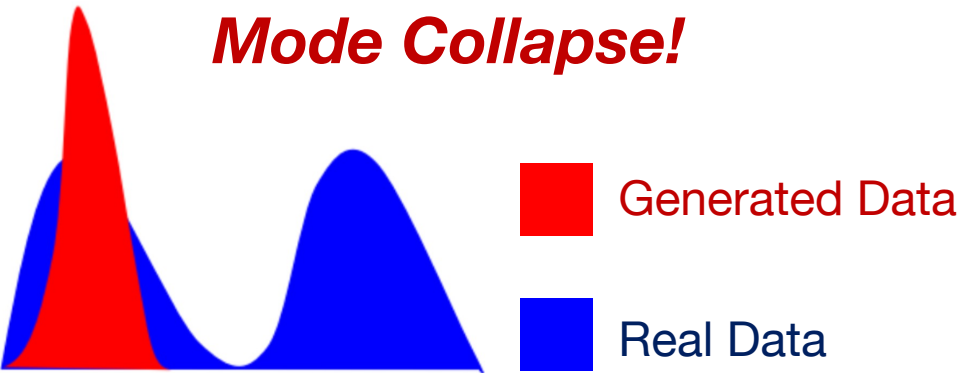


Align well with semantics

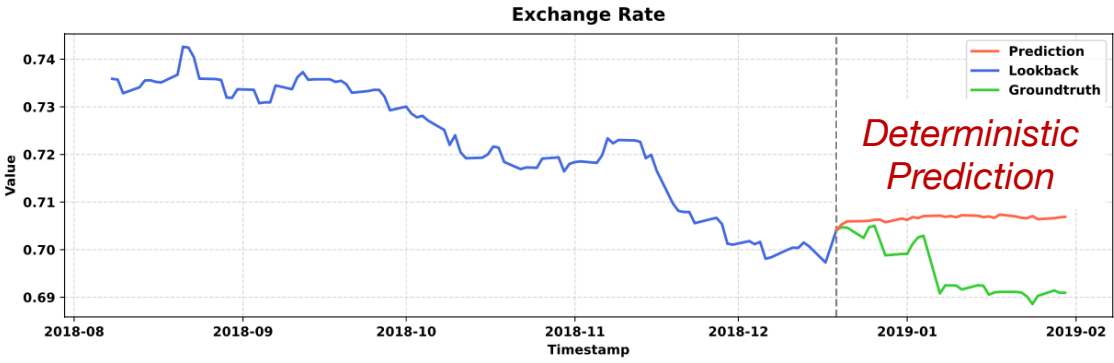
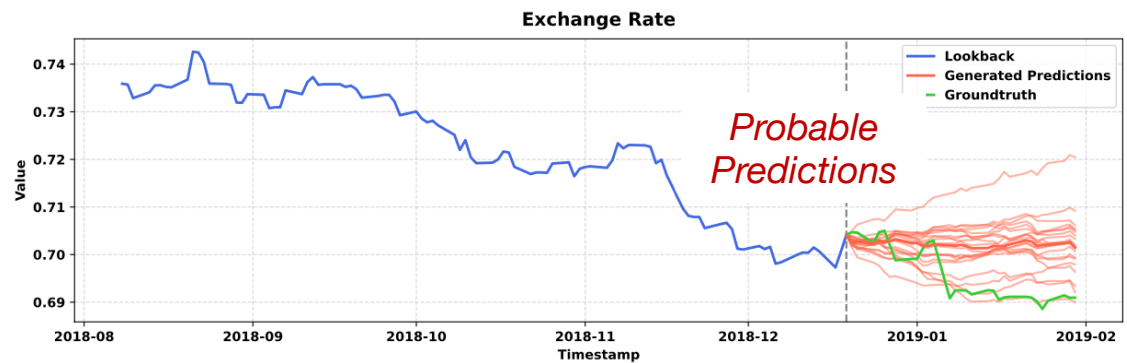


Hard to unify the semantics

Forecasting With Uncertainty



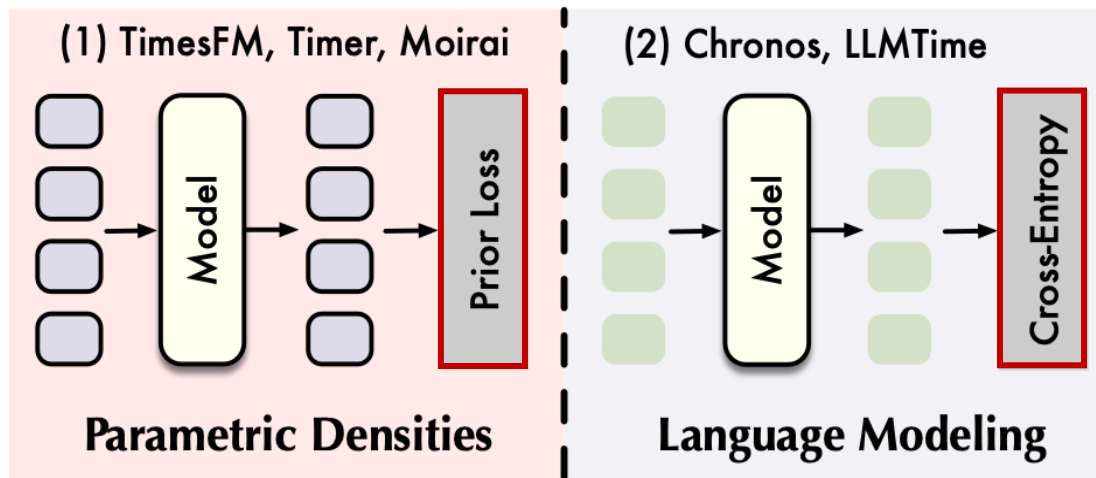
A Simple Prior Is The Scaling Bottleneck of Time Series Foundation Models



Real-World Observation is also a 'sampling' result

How: Can we use a more Flexible prior?

A Dilemma of Flexible Prior and Native Tokenization



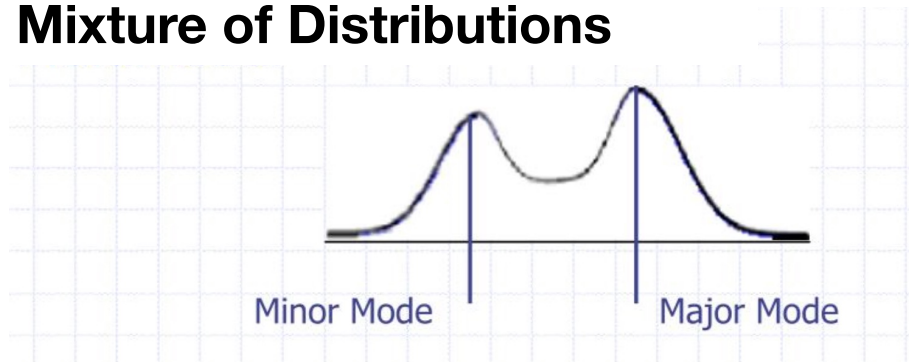
(1) Unimodal

- MSE - Mean Values
- MAE - Median Value
- Pinball Loss - Quantiles...



TSFM cannot presuppose its objective

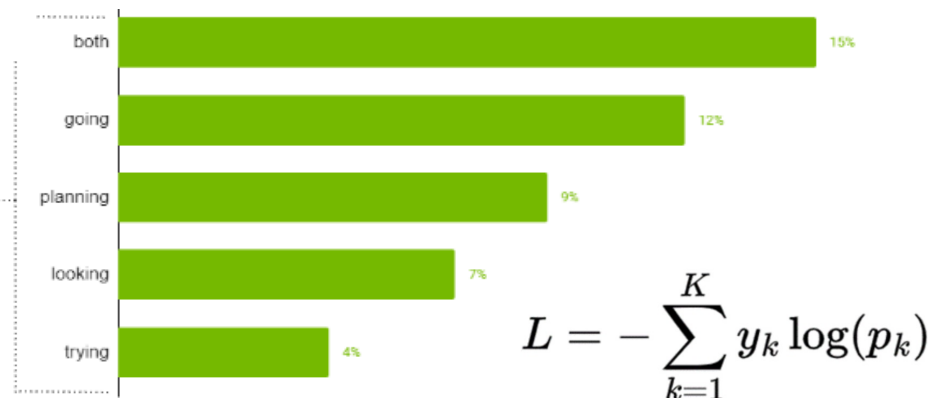
(2) Mixture of Distributions



Cannot assume #Mode in advance

(3) Categorical

My friend and I are



$$L = - \sum_{k=1}^K y_k \log(p_k)$$



Cross-Entropy needs discrete tokenization

A Dilemma of Flexible Prior and Native Tokenization

Time Series Is A Foreign Language To LLM



Embedding like LLM

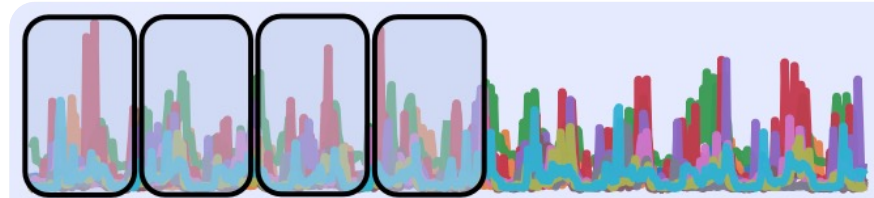
Discrete



- Two-stage round-off
- Out-of-vocabulary

Point-wise

- Long-context input
- Multi-step autoregression



Embedding like ViT

Continuous



- Lossless embedding
- Prior-free optimization?

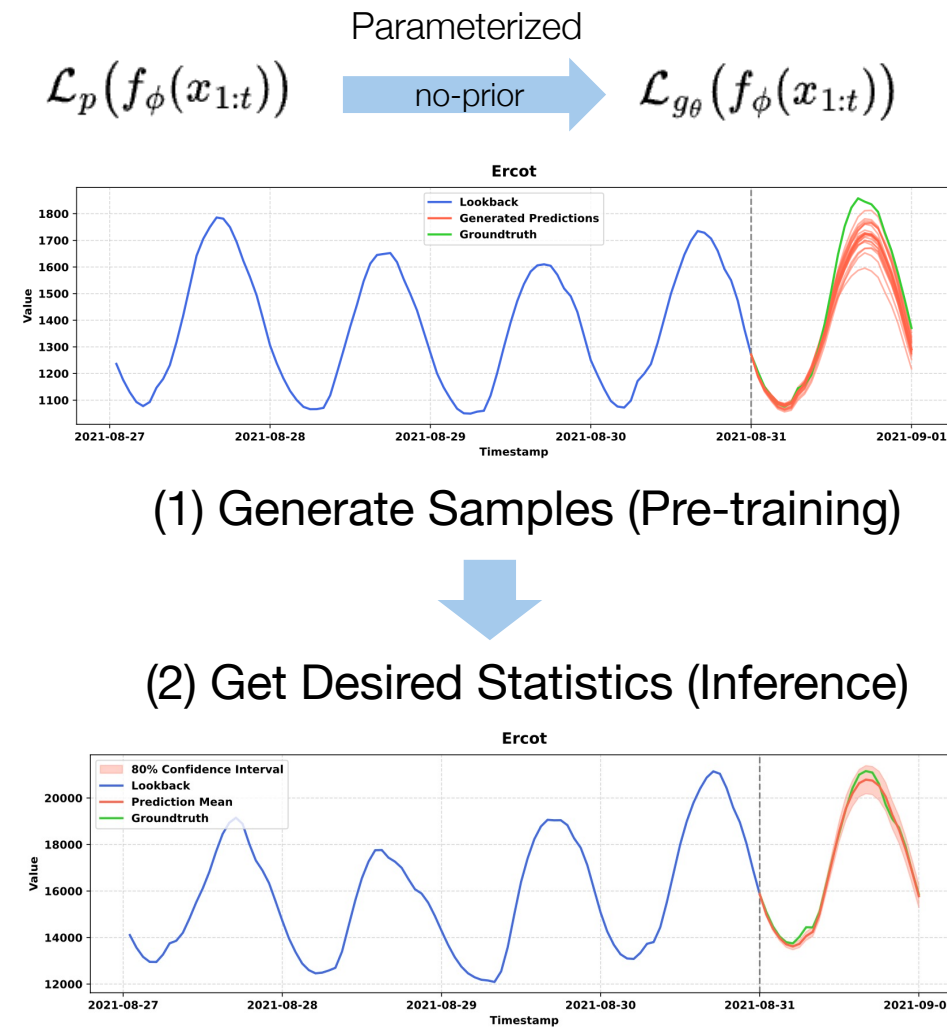
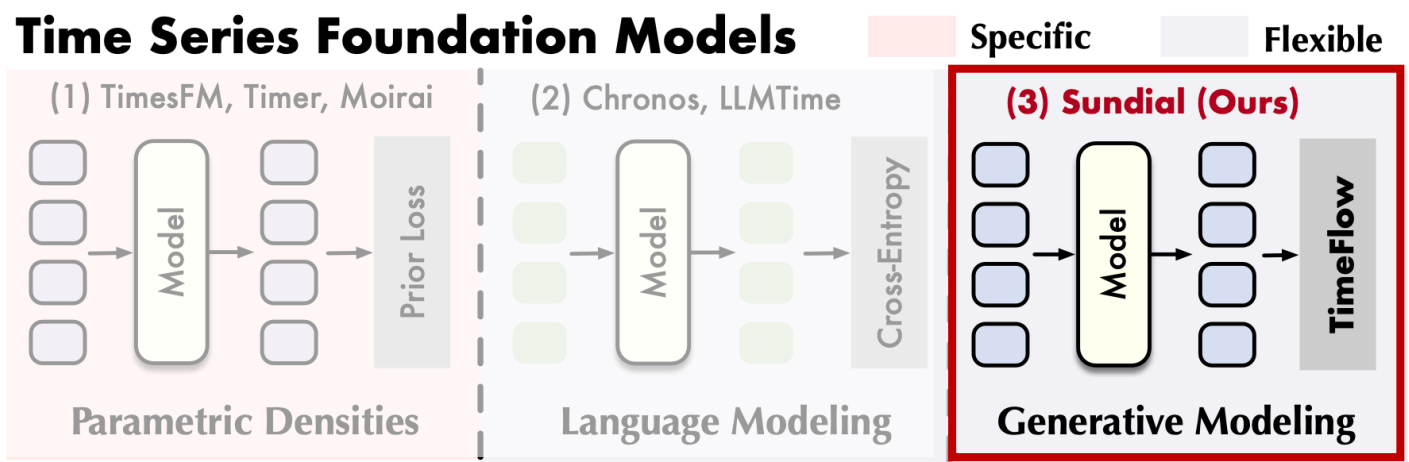
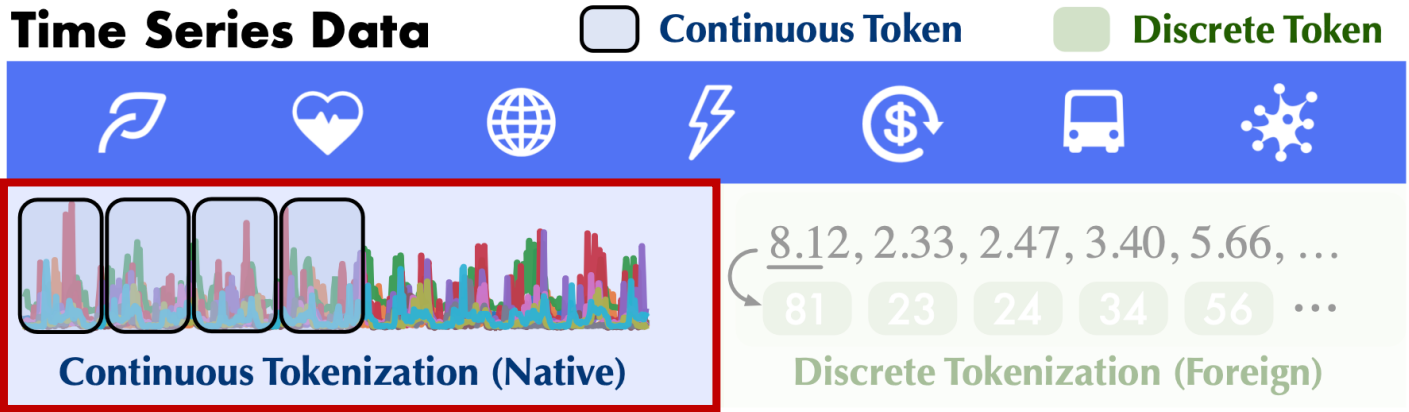
Patch-wise

- More semantic tokens
- Less computation costs

Learning Native and Flexible Language of Time Series

Native: Transformer without Discrete Tokenization

Flexible: Generative Forecasting



The First-Principle Design

ARMA (Theoretically Complete)

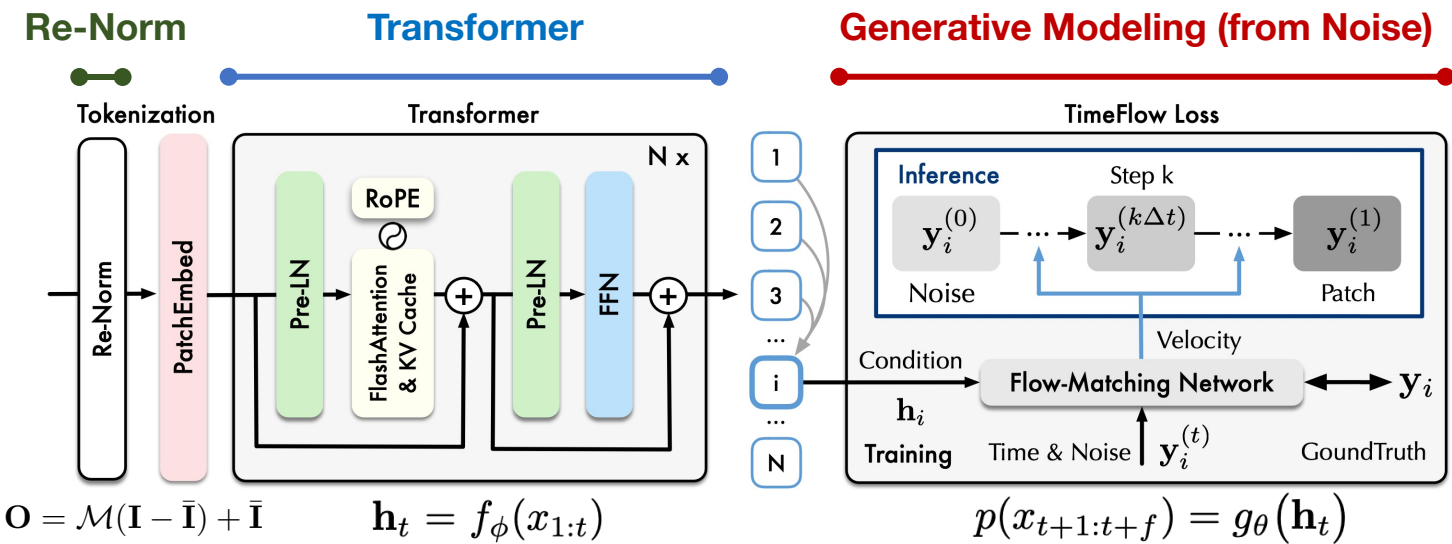
$$y'_t = c + \phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$



Statistics Model

- Mean Estimation
- Autoregressive (Point)
- Moving-Average
- Separately Fitting...

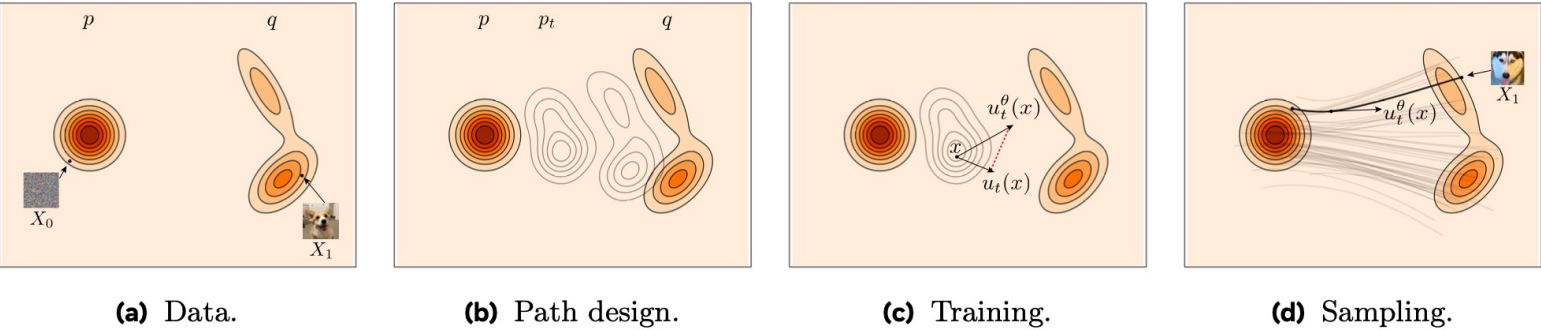
Sundial (Nonlinear and Adaptive)



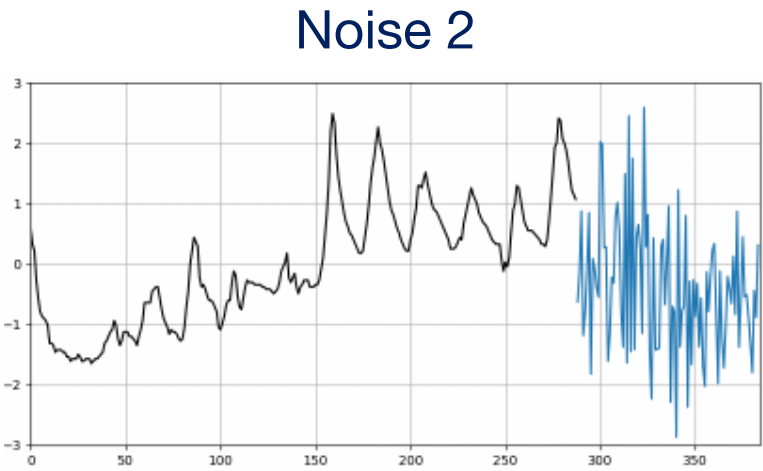
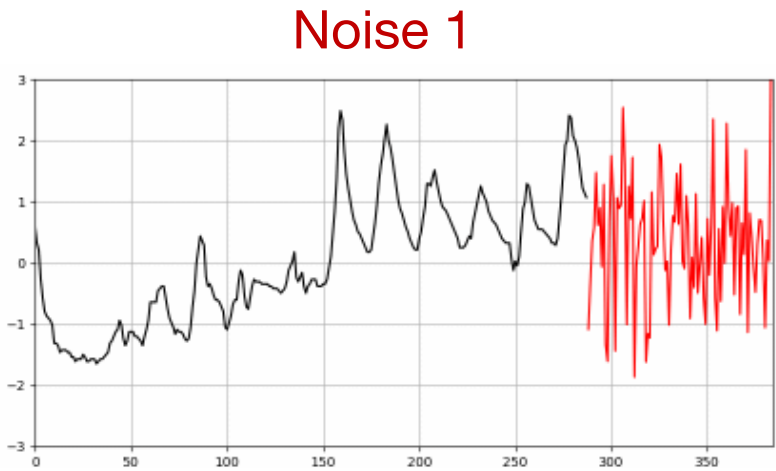
Deep Model

- Instance Re-Norm
- Decoder-Only (Patch)
- Conditional Generation
- In-Context Learning...

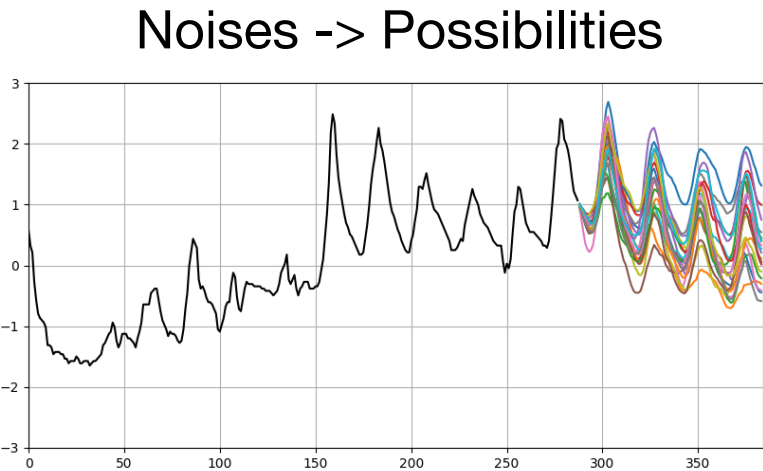
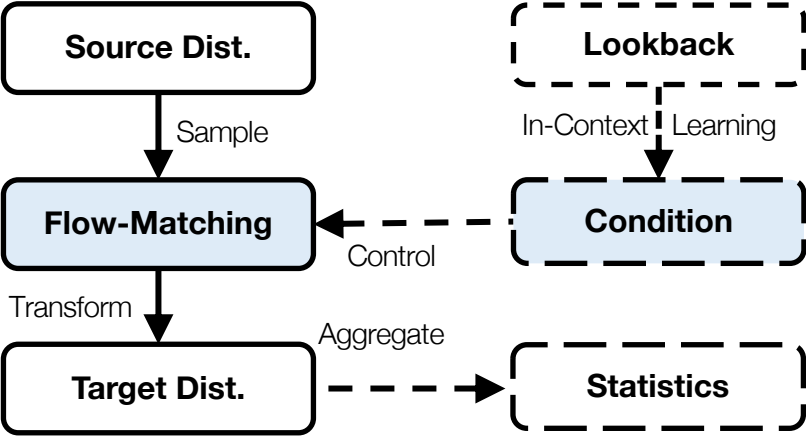
TimeFlow Loss



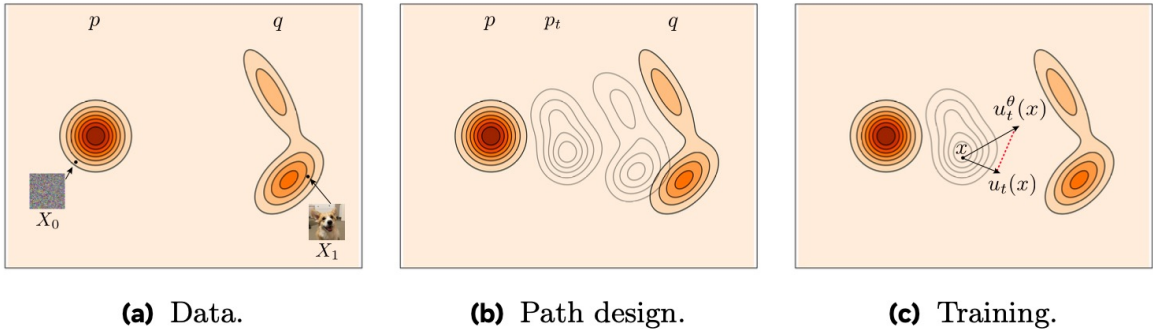
Conditional Flow-Matching



Pipeline of TimeFlow



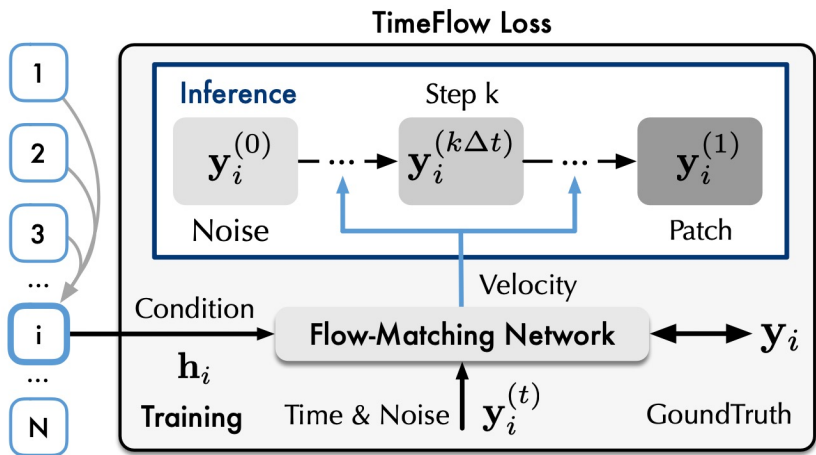
TimeFlow Loss



Why FM?

Objective	Error in Prediction						Avg.
	ETTM1	ETTM2	ETTh1	ETTh2	ECL	Weather	
TimeFlow	0.336	0.258	0.411	0.333	0.169	0.234	0.290
Diffusion	0.362	0.265	0.444	0.360	0.202	0.252	0.314
MSE	0.360	0.264	0.404	0.341	0.175	0.231	0.296

Conditional Flow-Matching



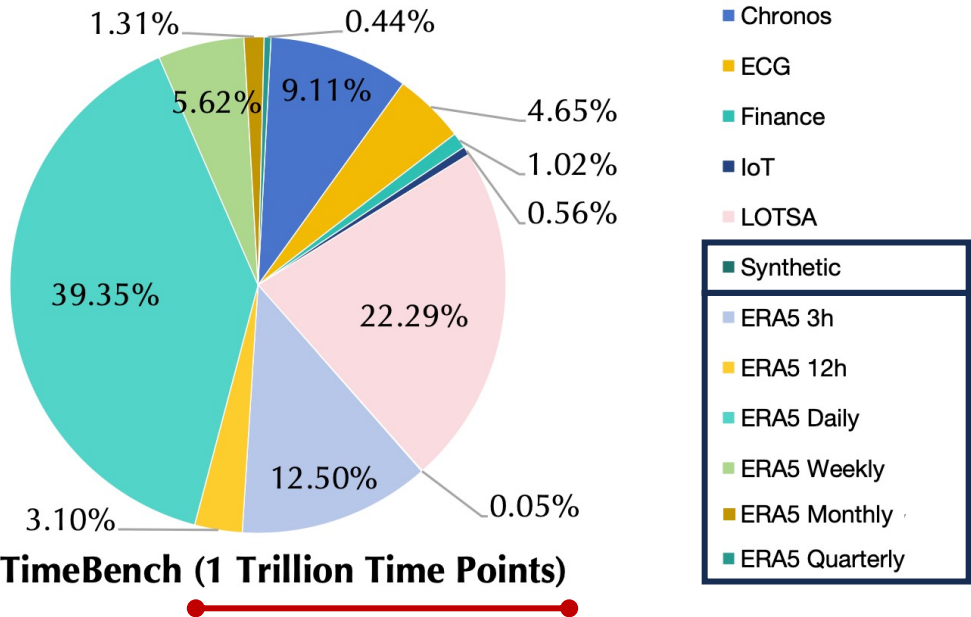
Velocity Prediction: Parameterized by a small MLP

$$\mathcal{L}_{\text{TimeFlow}} = \sum_{i=1}^N \left\| \underbrace{\text{FM-Net} \left(\mathbf{y}_i^{(t)}, t, \mathbf{h}_i \right)}_{\text{Predicted } v \text{ at } t} - \underbrace{\left(\mathbf{y}_i - \mathbf{y}_i^{(0)} \right)}_{\text{Real } v} \right\|^2$$

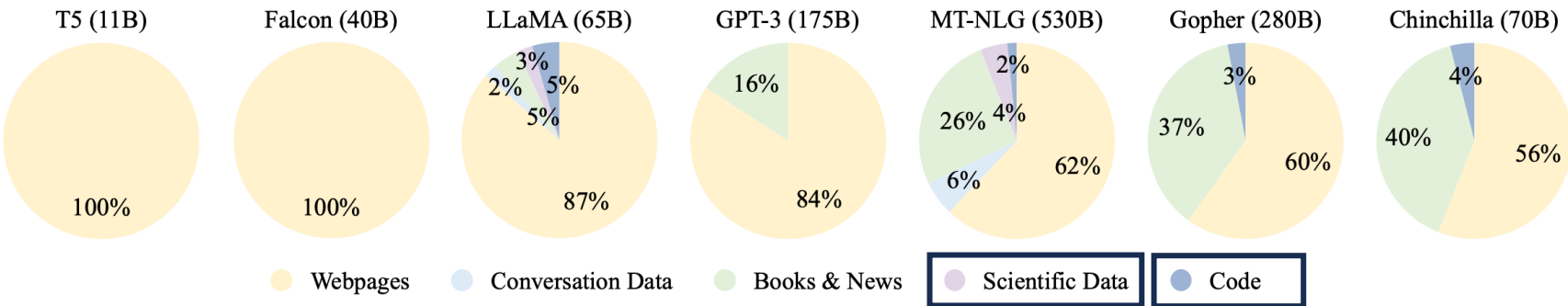
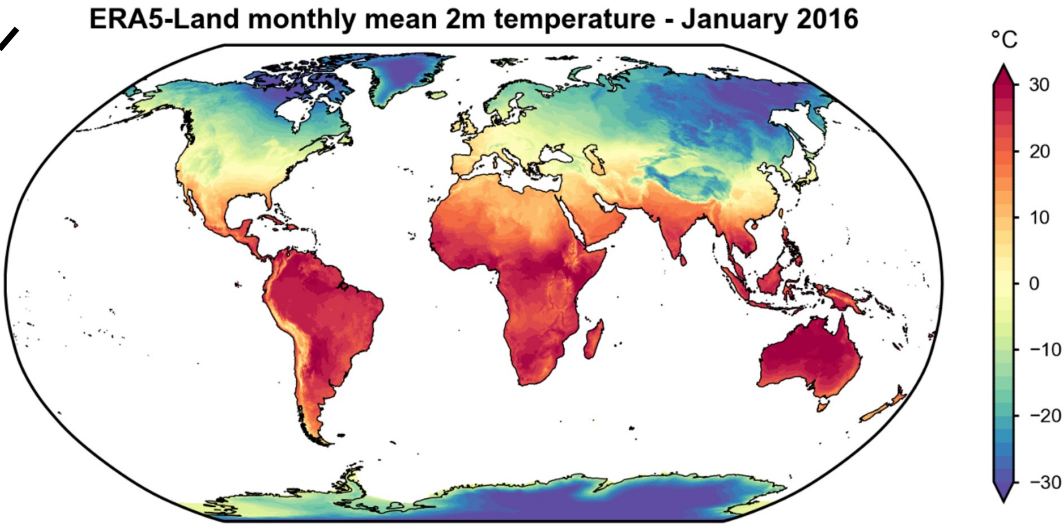
Push-Forward:

$$\underbrace{\mathbf{y}_i^{(t+\Delta t)}}_{\text{Predicted noise at } t} = \underbrace{\mathbf{y}_i^{(t)}}_{\text{Predicted noise at } t} + \underbrace{\text{FM-Net} \left(\mathbf{y}_i^{(t)}, t, \mathbf{h}_i \right) \Delta t}_{\text{Uniform-step trajectory}}$$

TimeBench: Data Collection



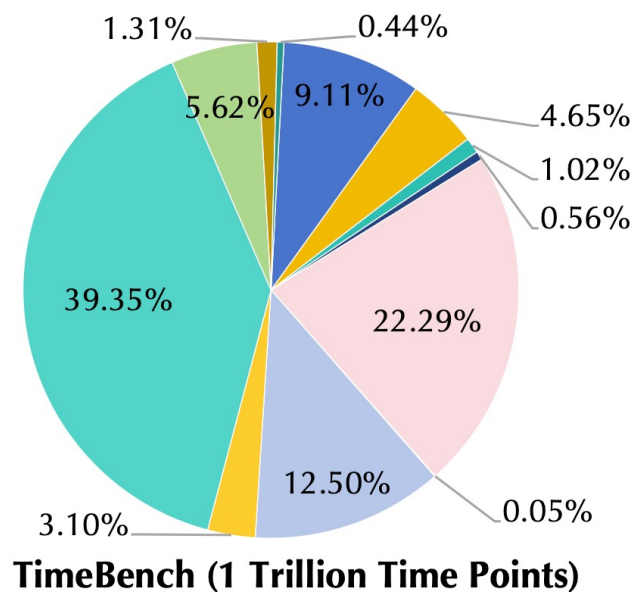
Why Synthetic and ERA5?



Principles

- Diverse
- Scientific
- Predictable

TimeBench: Data Curation



↓ (OOD to Evaluation)

Statistical Analysis

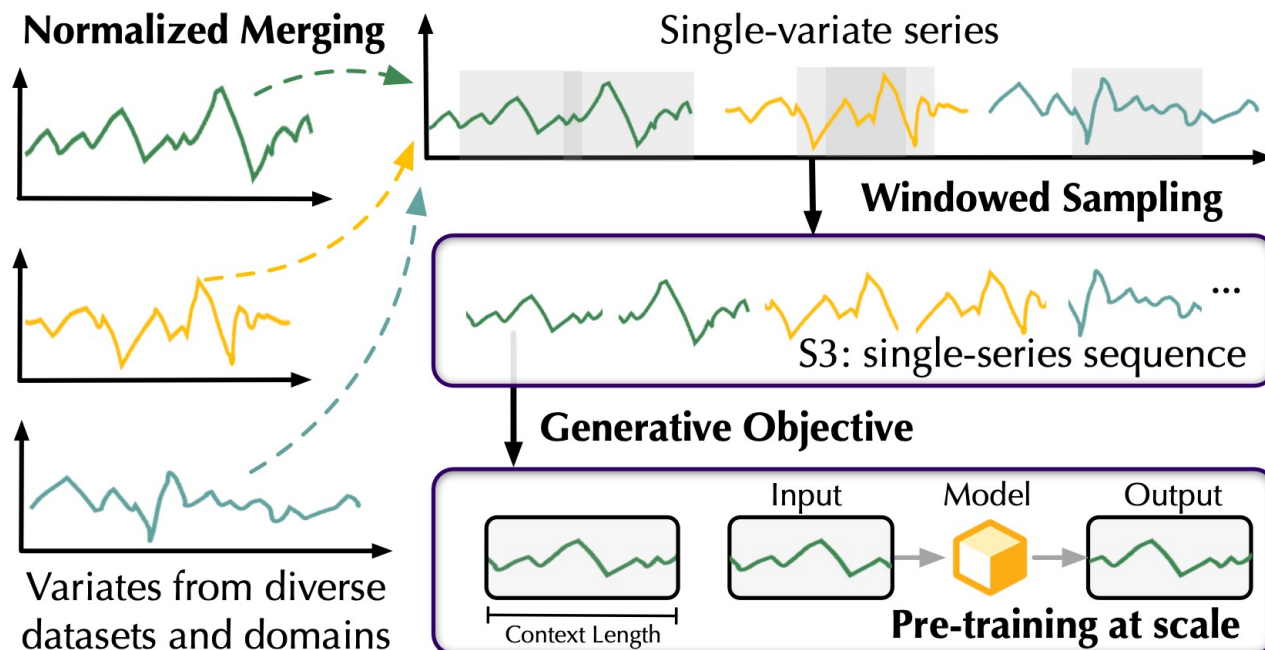
(Stationary: ADF Test; Predictability: FFT Mean, Entropy, Seasonality...)

Imputation & Anomaly

(Impute: Causal Mean Impute; Anomaly: $k\text{-}\sigma$, IQR...)

Preprocessing & Sampling

(Z-Score Norm; Single-Series Sequence; Domain-Balanced Weighting ...)

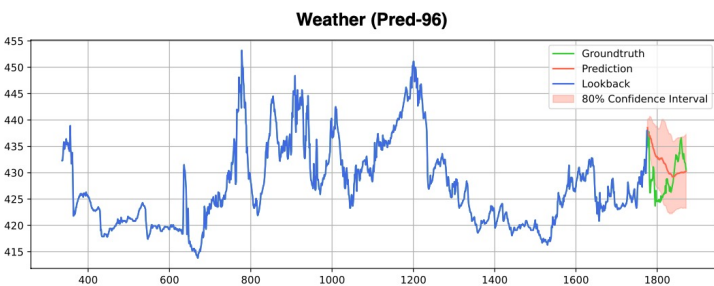
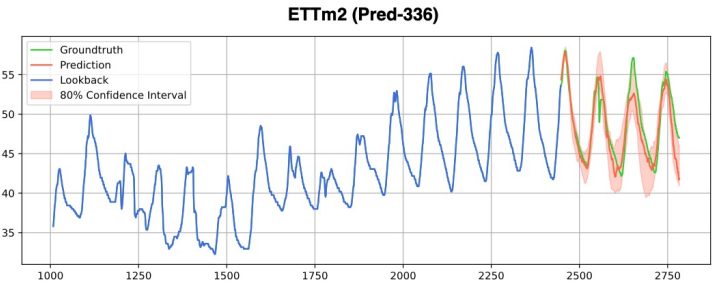
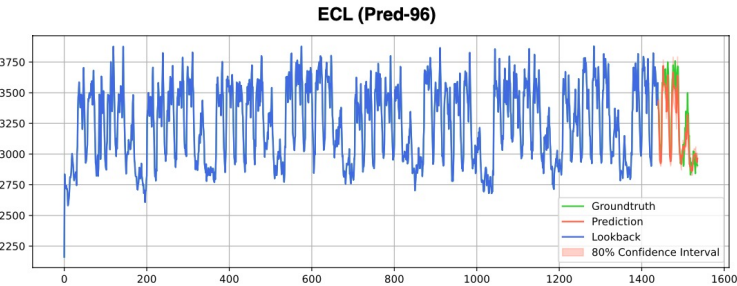


Zero-Shot Forecasting

Long-Term Forecasting ([Time-Series-Library](#)) 7.57% MSE ↓ 4.71% MAE ↓ (to Prev. SoTA)

Models	Sundial _{Small} (Ours)		Sundial _{Base} (Ours)		Sundial _{Large} (Ours)		Time-MoE _{Base} (2024b)		Time-MoE _{Large} (2024b)		Time-MoE _{Ultra} (2024b)		Timer (2024a)		Moirai _{Base} (2024)		Moirai _{Large} (2024)		Chronos _{Base} (2024)		Chronos _{Large} (2024)		TimesFM (2023b)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.354	0.388	<u>0.336</u>	<u>0.377</u>	0.331	0.369	0.394	0.415	0.376	0.405	0.356	0.391	0.373	0.392	0.406	0.385	0.422	0.391	0.645	0.500	0.555	0.465	0.433	0.418
ETTh2	0.265	0.324	<u>0.258</u>	<u>0.320</u>	0.254	0.315	0.317	0.365	0.316	0.361	0.288	0.344	0.273	0.336	0.311	0.337	0.329	0.343	0.310	0.350	0.295	0.338	0.328	0.346
ETTm1	0.390	<u>0.418</u>	0.411	0.434	0.395	0.420	0.400	0.424	<u>0.394</u>	0.419	0.412	0.426	0.404	0.417	0.417	0.419	0.480	0.439	0.591	0.468	0.588	0.466	0.473	0.443
ETTm2	0.340	0.387	0.333	0.387	<u>0.334</u>	0.387	0.366	0.404	0.405	0.415	0.371	0.399	0.347	0.388	0.362	<u>0.382</u>	0.367	0.377	0.405	0.410	0.455	0.427	0.392	0.406
ECL	<u>0.169</u>	<u>0.265</u>	<u>0.169</u>	<u>0.265</u>	0.166	0.262	-	-	-	-	-	-	0.174	0.278	0.187	0.274	0.186	0.270	0.214	0.278	0.204	0.273	-	-
Weather	0.233	<u>0.271</u>	<u>0.234</u>	0.270	0.238	0.275	0.265	0.297	0.270	0.300	0.256	0.288	0.256	0.294	0.287	0.281	0.264	0.273	0.292	0.315	0.279	0.306	-	-
1 st Count	7	2	<u>8</u>	5	16	16	0	1	0	0	2	1	1	3	0	2	0	<u>6</u>	0	0	0	0	0	0

Showcases



Zero-Shot Forecasting

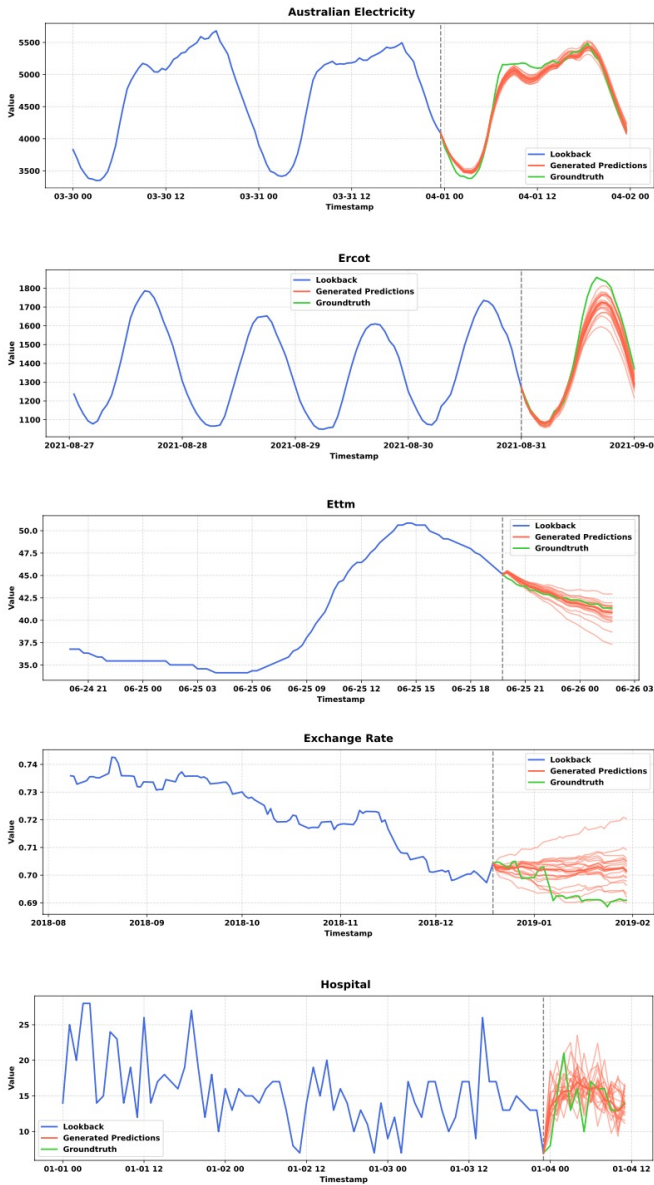
GIFT-Eval (Salesforce) *w/o Prior in Evaluation Metrics*

We introduce the General Time Series Forecasting Model Evaluation, GIFT-Eval, a pioneering benchmark aimed at promoting evaluation across diverse datasets. GIFT-Eval encompasses 24 datasets over 144,000 time series and 177 million data points, spanning seven domains, 10 frequencies, multivariate inputs, and prediction lengths ranging from short to long-term forecasts.

🏆 Overall 🏆 By Domain 🏆 By Frequency 🏆 By Term Length 🏆 By Variate Type 📖 About

T	model	MASE	CRPS
🟢	TiRex	0.650	0.421
🟢	Toto_Open_Base_1.0	0.673	0.437
🟢	sundial_base_128m	0.673	0.472
🟡	TTM-R2-Finetuned (code)	0.679	0.492
🟢	timesfm_2_0_500m (code)	0.680	0.465
🟢	TabPFN-TS	0.692	0.46
🟢	YingLong_300m	0.716	0.463
🟢	chronos_bolt_base (code)	0.725	0.485
🟢	YingLong_110m	0.726	0.471
🟢	YingLong_50m	0.738	0.479
🟢	chronos_bolt_small (code)	0.738	0.487

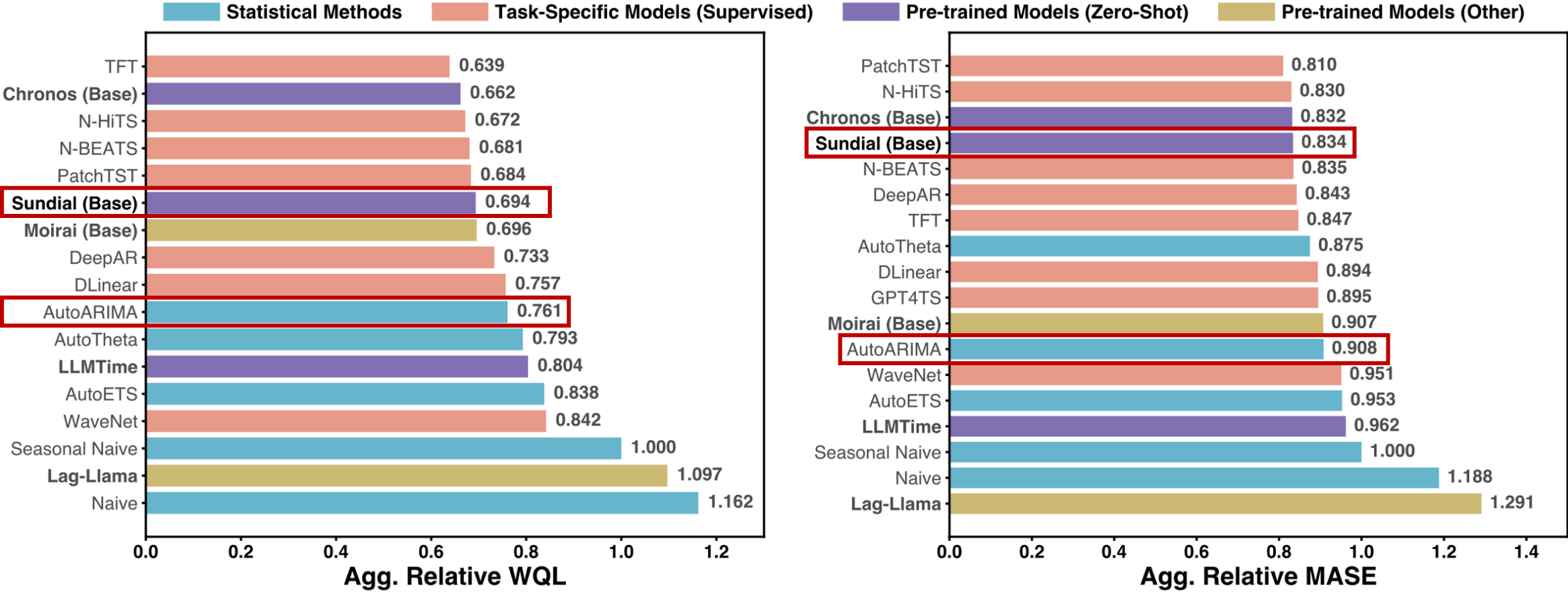
Generated Results



Zero-Shot Forecasting

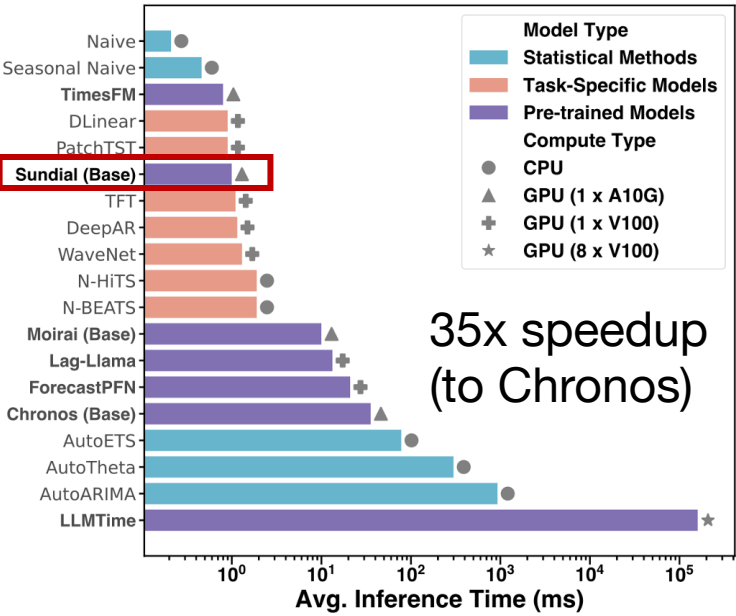
[FEV](#) (AutoGluon)

Surpass Auto-ARIMA(p,d,q) Fitting and Tuning on In-Distribution Data



Inference Speed

On GPU (A100, ms-level)



On CPU (Apple M1, sec-level)

Lookback Length	Prediction Length	# Generated Samples	Inference Time	Accelerate By
672	16	1	249ms	-
2880	16	1	510ms	FlashAttention
2880	720	1	510ms	Multi-Patch Prediction
2880	1440	1	789ms	KV Cache
2880	720	20	949ms	Shared Condition

Ref: Auto-ARIMA

Dataset Size

Manual Tuning

Automated Search

Small (<1000 pts)

10 min – 2 hours

30 sec – 5 min

Medium (100~100K pts)

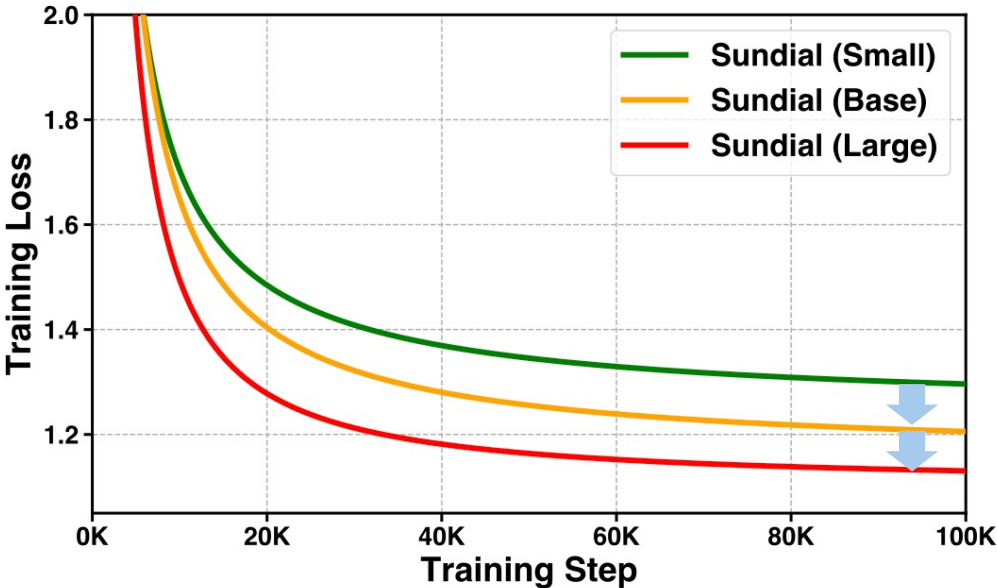
1 hour – 10 hours

5 min – 1 hour

Scaling

Scalable Training

- ✓ Large Model ↑
- ✓ More Data ↑

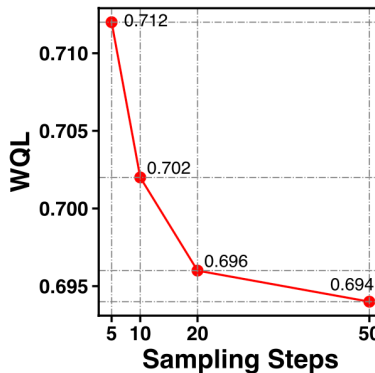
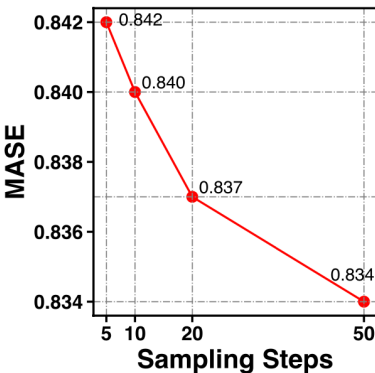
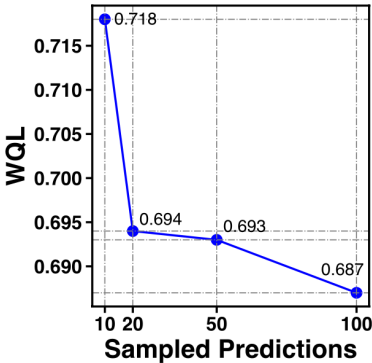
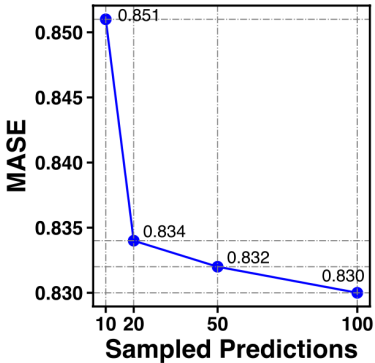


Models	Sundial _{Small} (Ours)		Sundial _{Base} (Ours)		Sundial _{Large} (Ours)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	0.354	0.388	<u>0.336</u>	<u>0.377</u>	0.331	0.369
ETTm2	0.265	0.324	<u>0.258</u>	<u>0.320</u>	0.254	0.315
ETTh1	0.390	<u>0.418</u>	0.411	0.434	0.395	0.420
ETTh2	0.340	0.387	0.333	0.387	<u>0.334</u>	0.387
ECL	<u>0.169</u>	<u>0.265</u>	<u>0.169</u>	<u>0.265</u>	0.166	0.262
Weather	0.233	<u>0.271</u>	<u>0.234</u>	0.270	0.238	0.275
1 st Count	7	2	<u>8</u>	5	16	16

Test-Time Calibration

- ✓ Use More Samples ↑
- ✓ Fine-Grained FM Steps ↑




Error in Prediction









Resource: Out-of-the-Box Models

Time Series Foundation Models Hugging Face

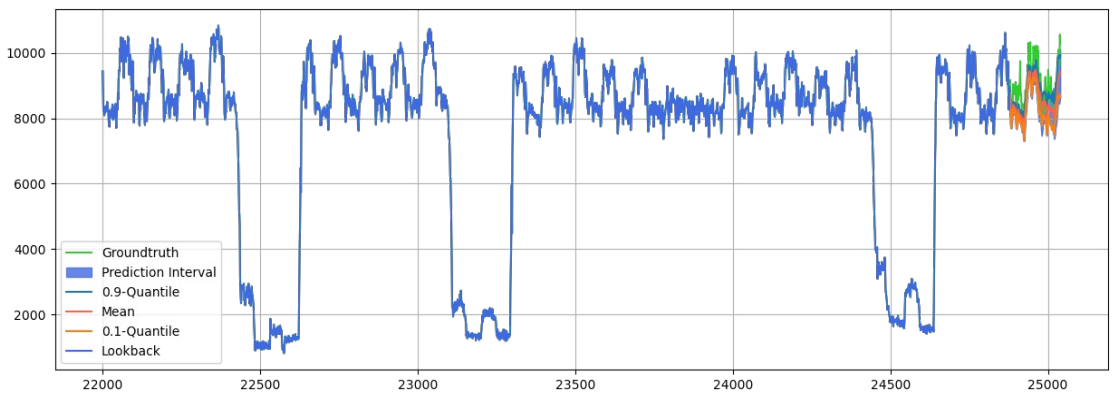
Pretrained models and datasets for out-of-the-box time series forecasting 

 **thuml/sundial-base-128m**
Time Series Forecasting • Updated 6 days ago •  10.5k •  15

 **thuml/timer-base-84m**
Time Series Forecasting • Updated 6 days ago •  44k •  43

 **thuml/UTSD**
Viewer • Updated 6 days ago •  867k •  8.22k •  27

Point forecasting, quantile, and interval estimation



More details: <https://github.com/thuml/Sundial>

Example of Sundial (Generative Forecasting) **No Training**

```
import torch
from transformers import AutoModelForCausalLM

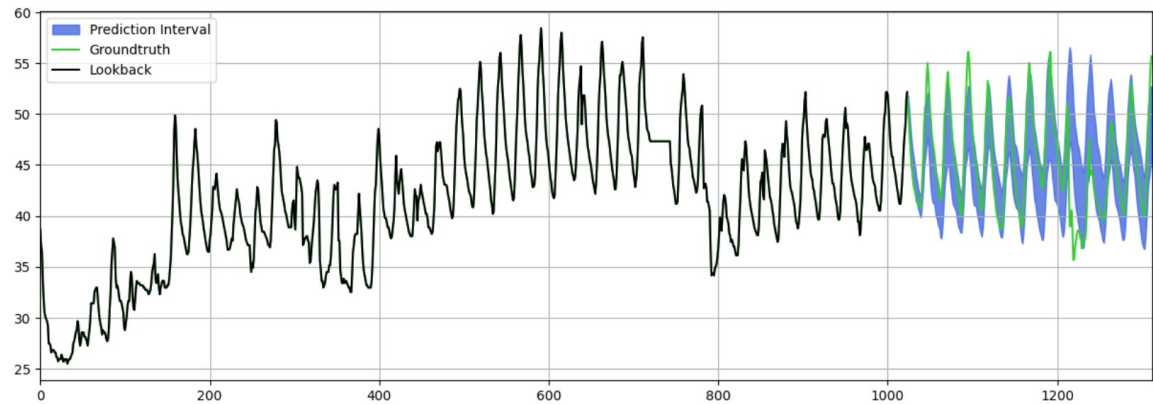
# load pretrain model
# supports different lookback/forecast lengths
model = AutoModelForCausalLM.from_pretrained('thuml/sundial-base-128m', trust_remote_code=True)

# prepare input
batch_size, lookback_length = 1, 2880
seqs = torch.randn(batch_size, lookback_length)

# Note that Sundial can generate multiple probable predictions
forecast_length = 96
num_samples = 20
```

Just-in-Time (CPU)

All you need is a HuggingFace account

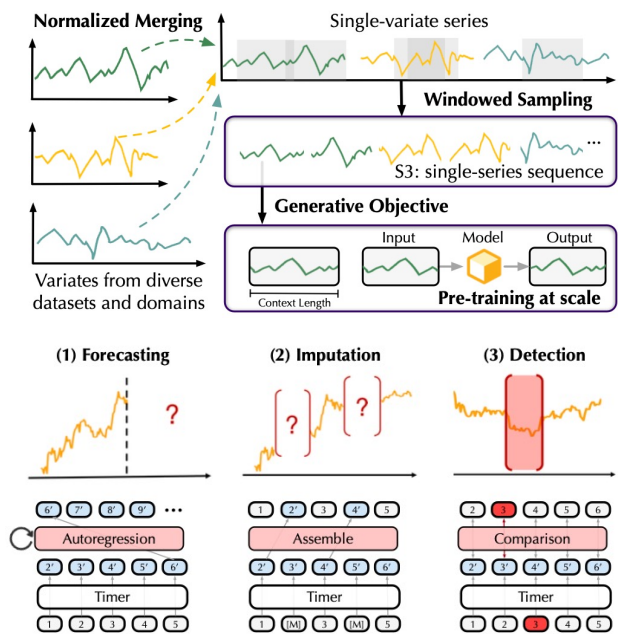


Code for fine-tuning will be available soon!

A Path Towards Time Series Foundation Model

Timer-v1 (ICML 2024)

- Unified Pre-training
- Multiple Tasks

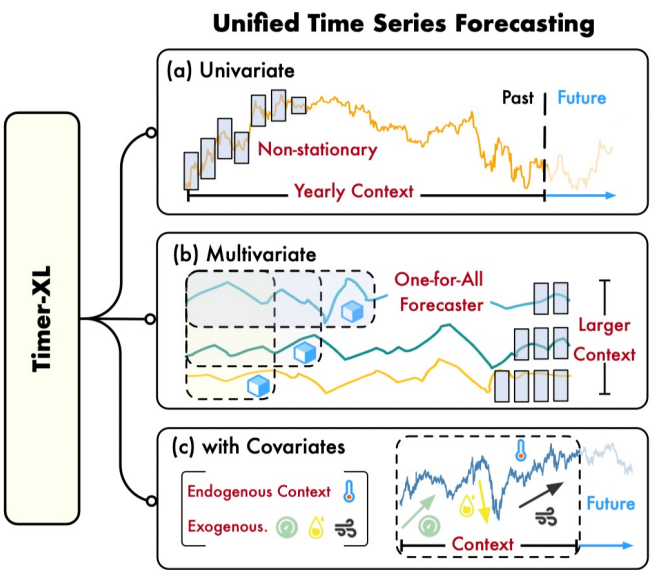


Multitask, Few-Shot

1440/67M/28B

Timer-XL (ICLR 2025)

- Long-Context Extension
- Unified Time Series Attention
- Enhanced Position Embedding

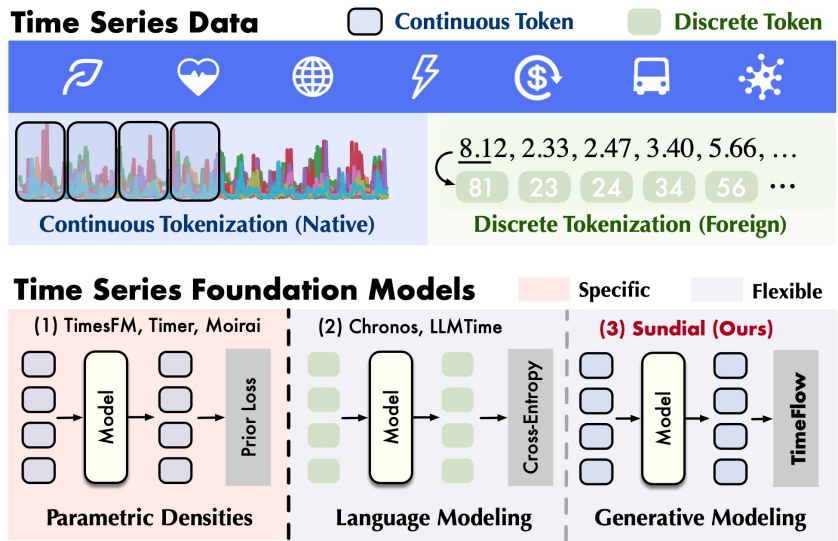


2D Context, Zero-Shot

2880/84M/260B

Sundial (ICML 2025)

- Trillion-Scale Pre-training
- Native Time Series Transformer
- Flow-Matching for Generation



Probabilistic, Flexible

2880/444M/1032B

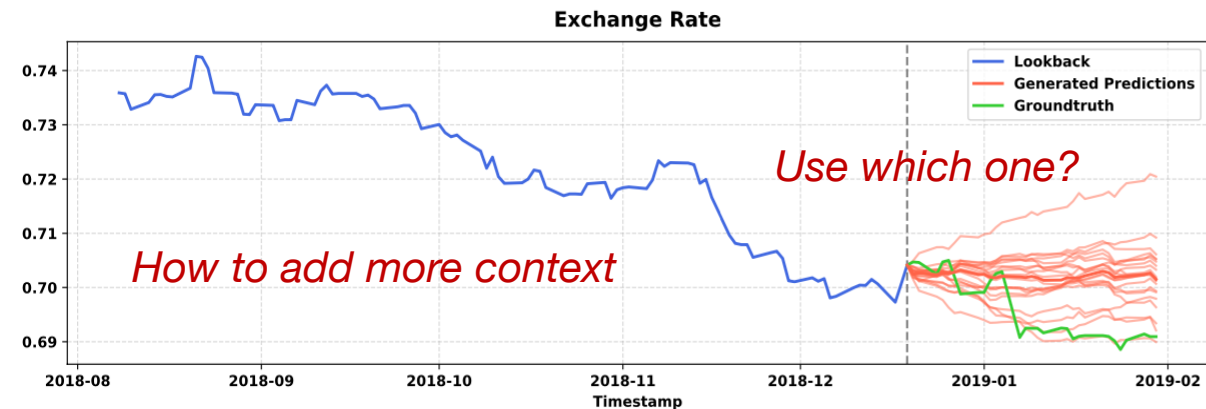
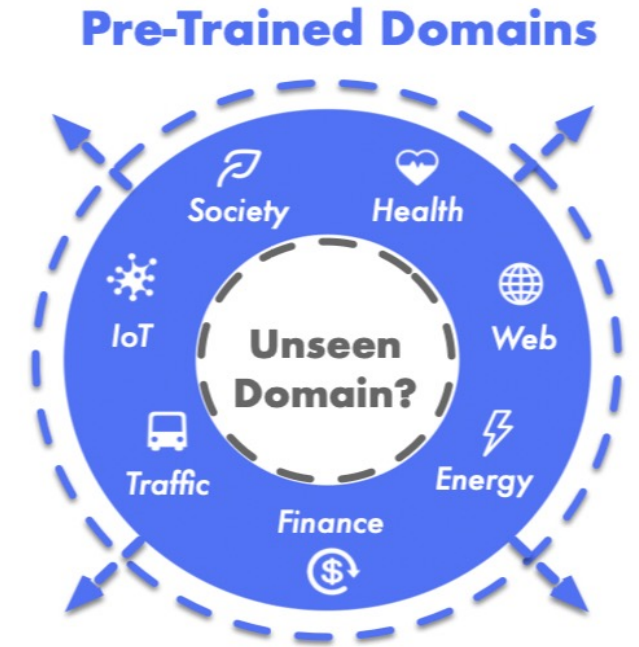
Limitations and Future Directions

From A Model Developer's Perspective

- Data: Effective Data Expansion/Synthetics
- Architecture: Unified Context Modeling
- Generalization: OOD and Ambiguity

From A Practitioner's Perspective

- Finetuning: Univariate -> Multivariate
- Domain-Specific: Physics-driven
- Generative + RL: Rewarded Predictions



Thank you!



Code



Paper



Yong Liu

liuyong21@mails.tsinghua.edu.cn



Guo Qin

qinguo24@mails.tsinghua.edu.cn



Zhiyuan Shi

shizy22@mails.tsinghua.edu.cn



Zhi Chen

chenzhi21@mails.tsinghua.edu.cn



Caiyin Yang

ycy23@mails.tsinghua.edu.cn



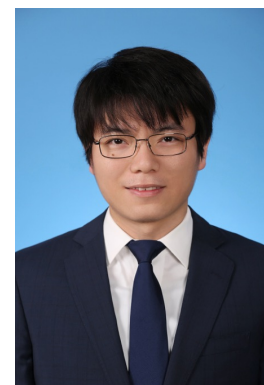
Xiangdong Huang

huangxdong@tsinghua.edu.cn



Jianmin Wang

jimwang@tsinghua.edu.cn



Mingsheng Long

mingsheng@tsinghua.edu.cn



清华大学
Tsinghua University

