World Simulators

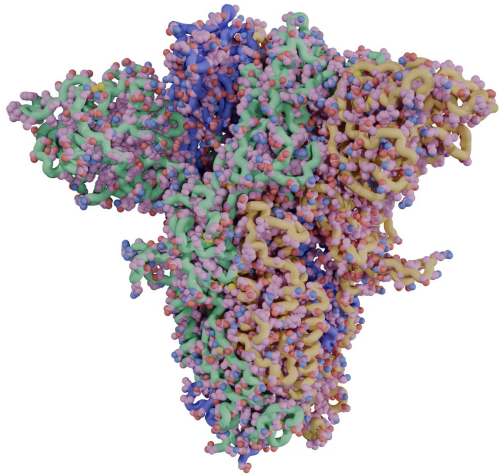
Toward Intelligent Dynamical System Simulation

Haixu Wu

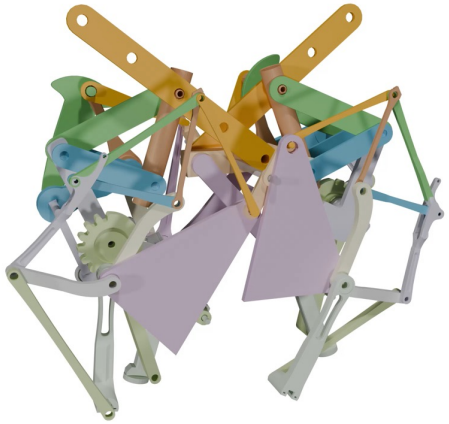
MIT CSAIL

May 11, 2026

Dynamical world



Protein



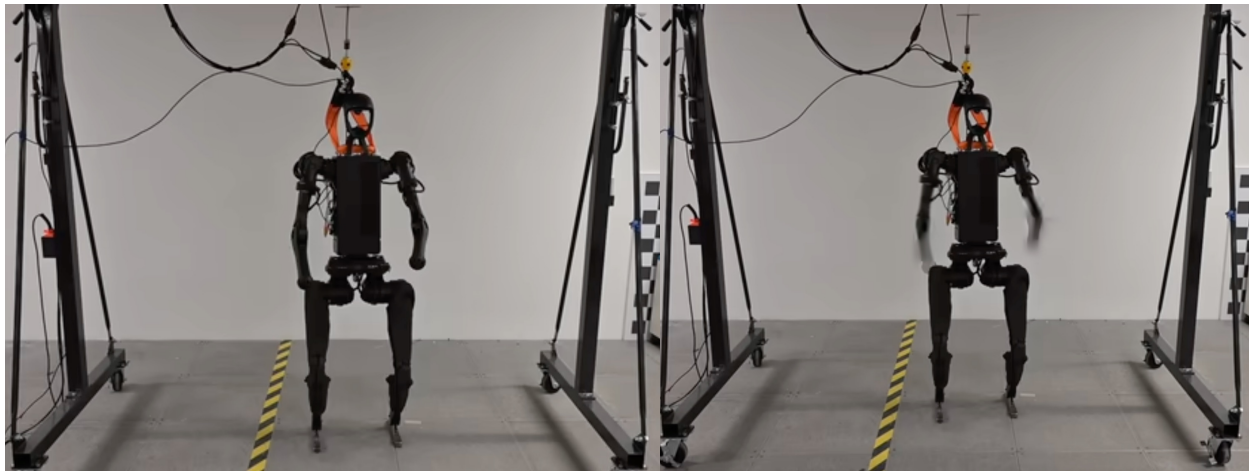
Rigid body

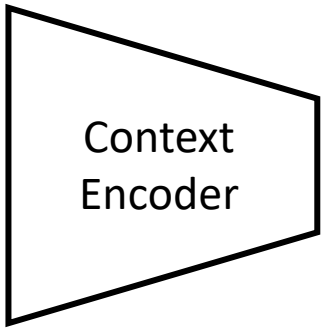
Weather

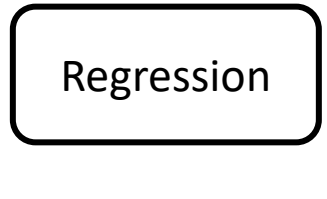
Finance

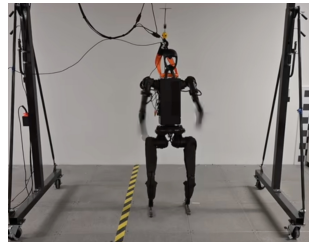
Scientific engineering, industrial design, risk prevention, investment …

Dynamical world modeling

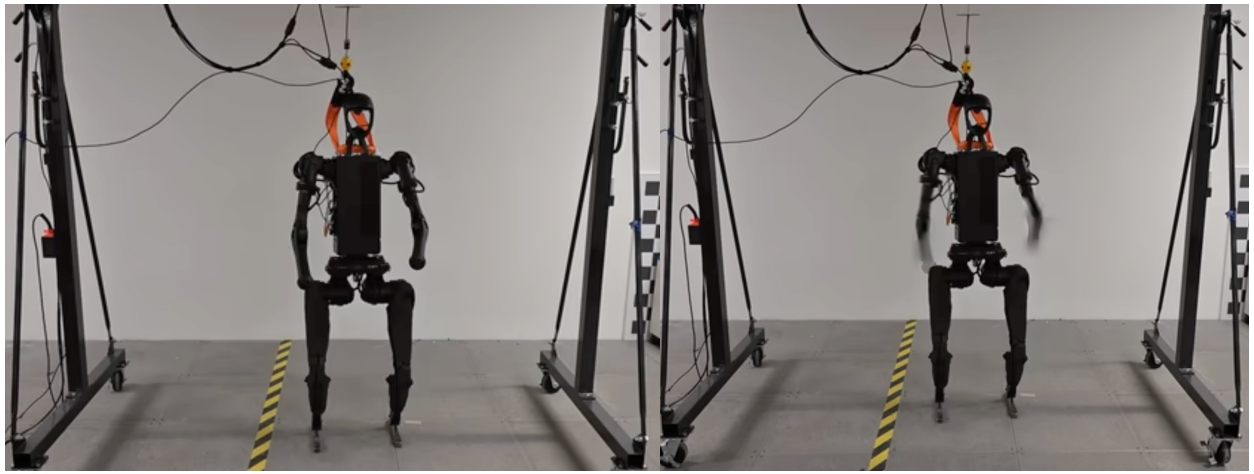


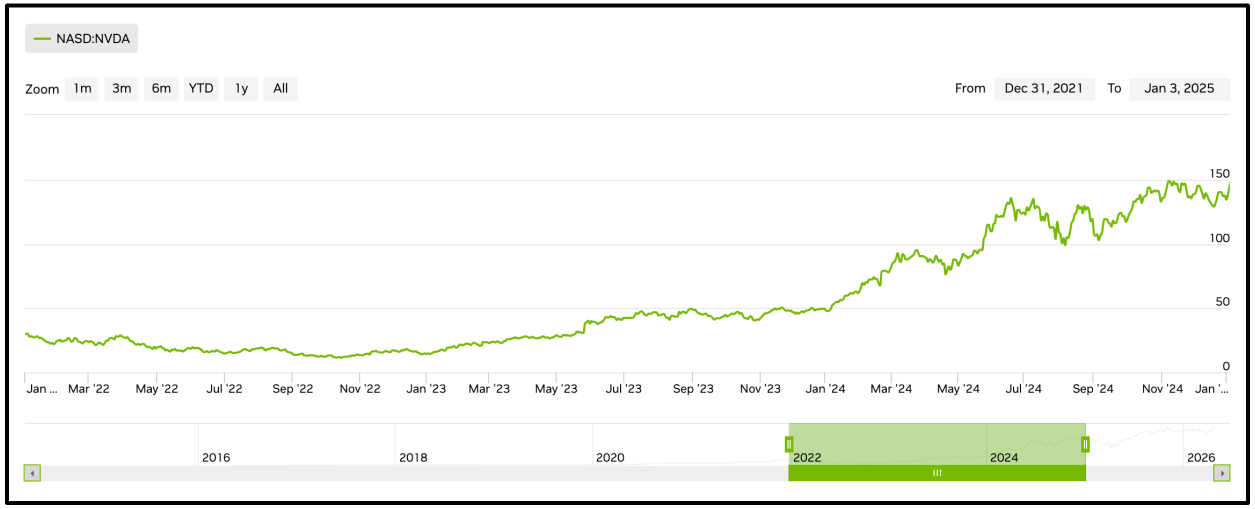






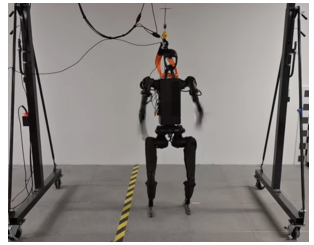
Dynamical world modeling





Context Encoder

Regression

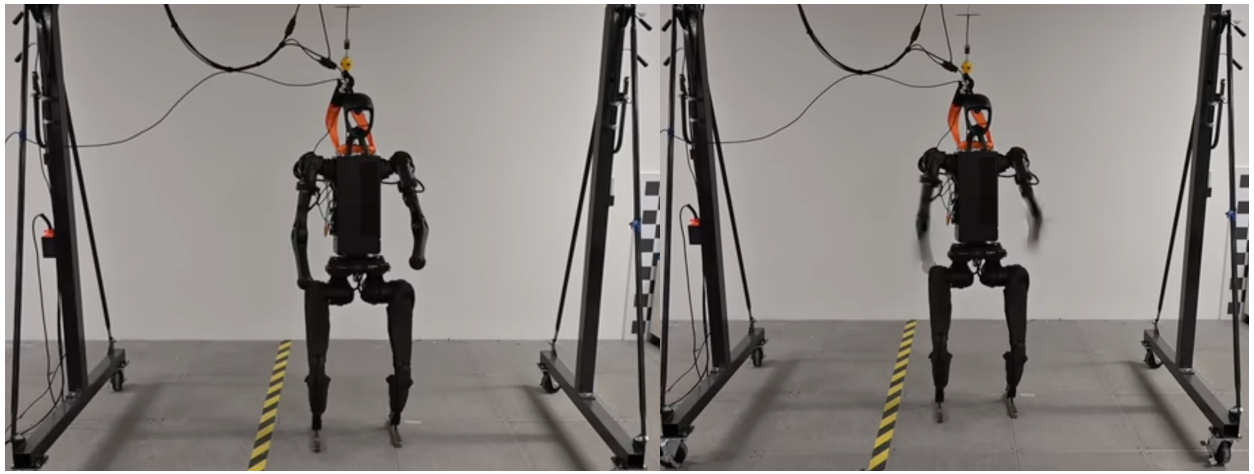


Context Encoder

Regression

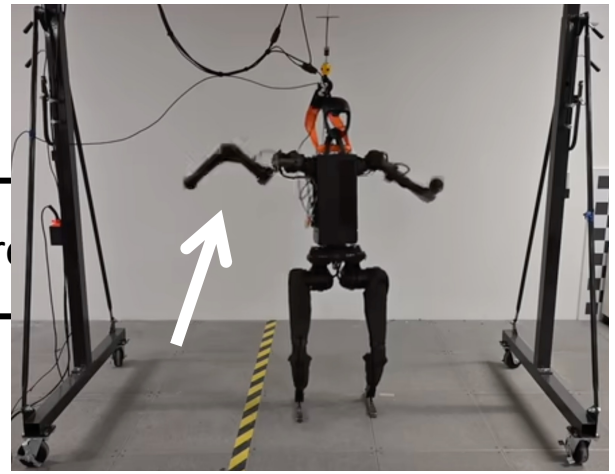
https://investor.nvidia.com/stock-info/stock-quote-and-chart/default.aspx

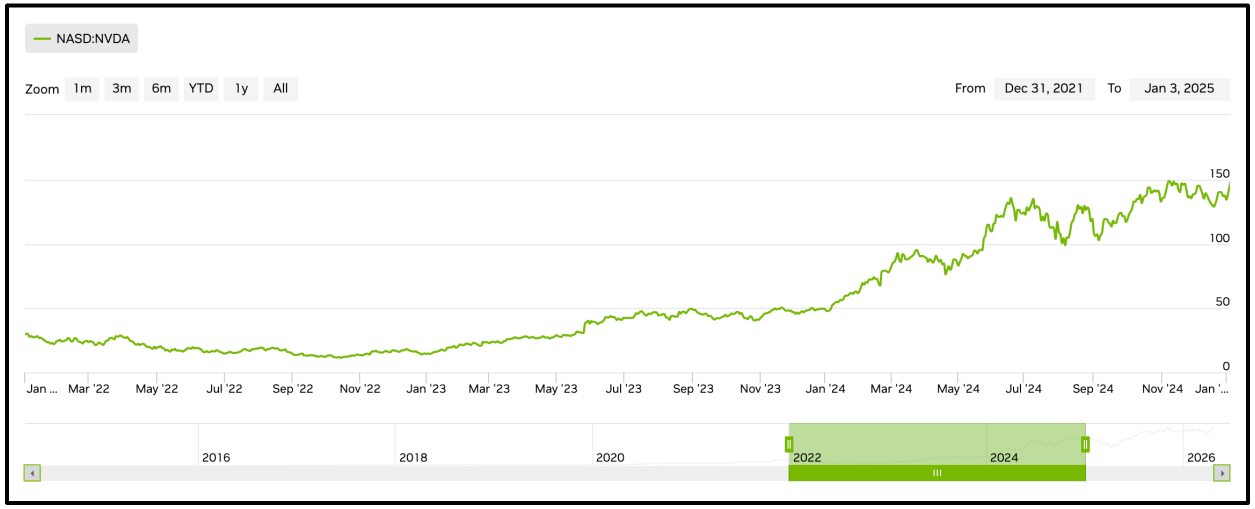
What actually happened



Context Encoder

Regr



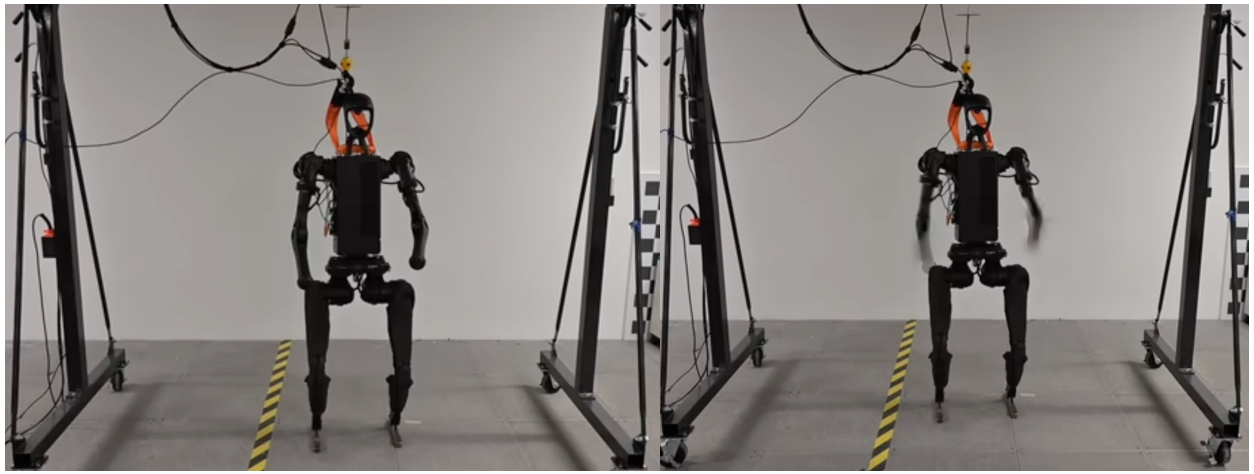


Context Encoder

Regression

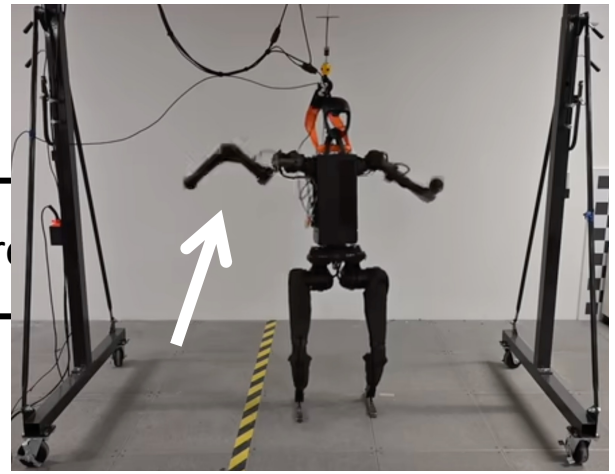
https://investor.nvidia.com/stock-info/stock-quote-and-chart/default.aspx

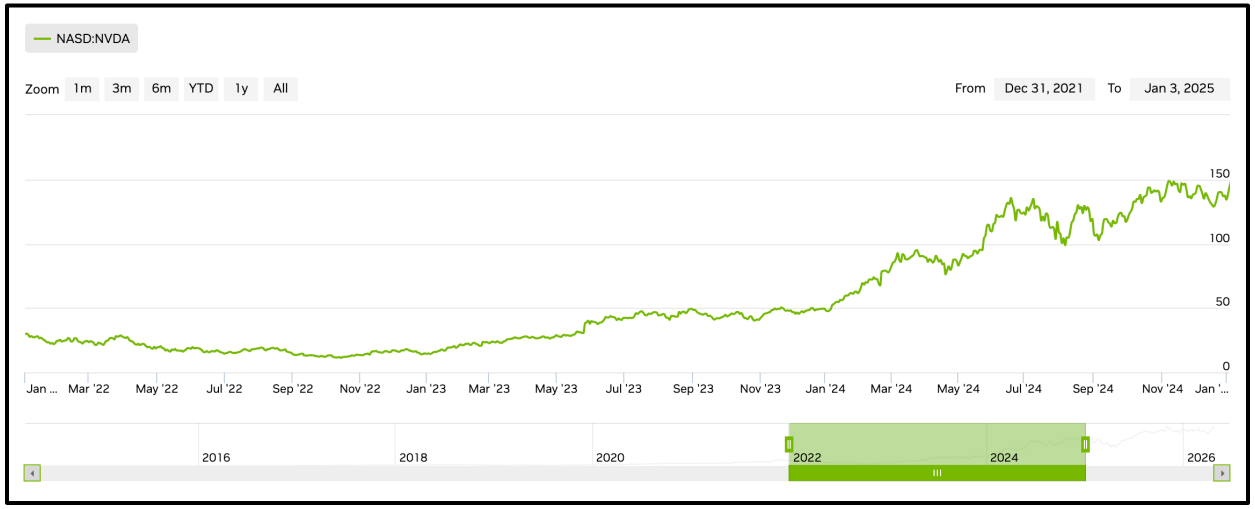
What actually happened



Context Encoder

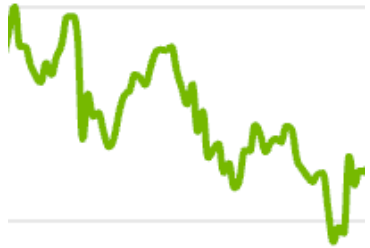
Regr





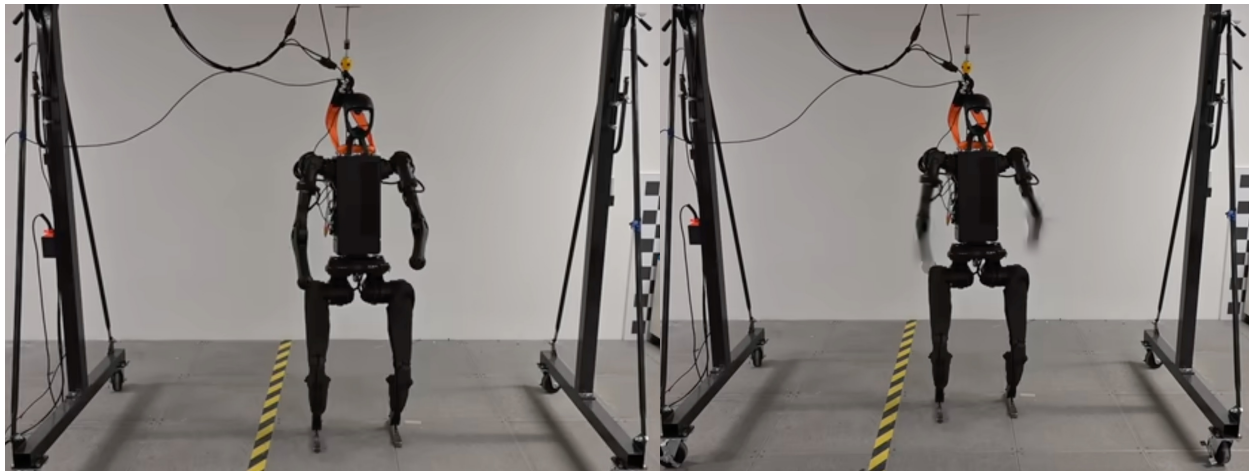
Context Encoder

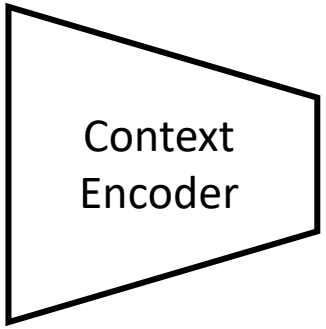
Regression

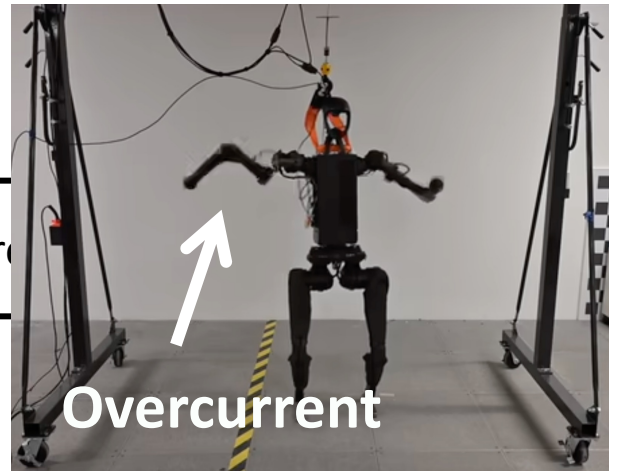


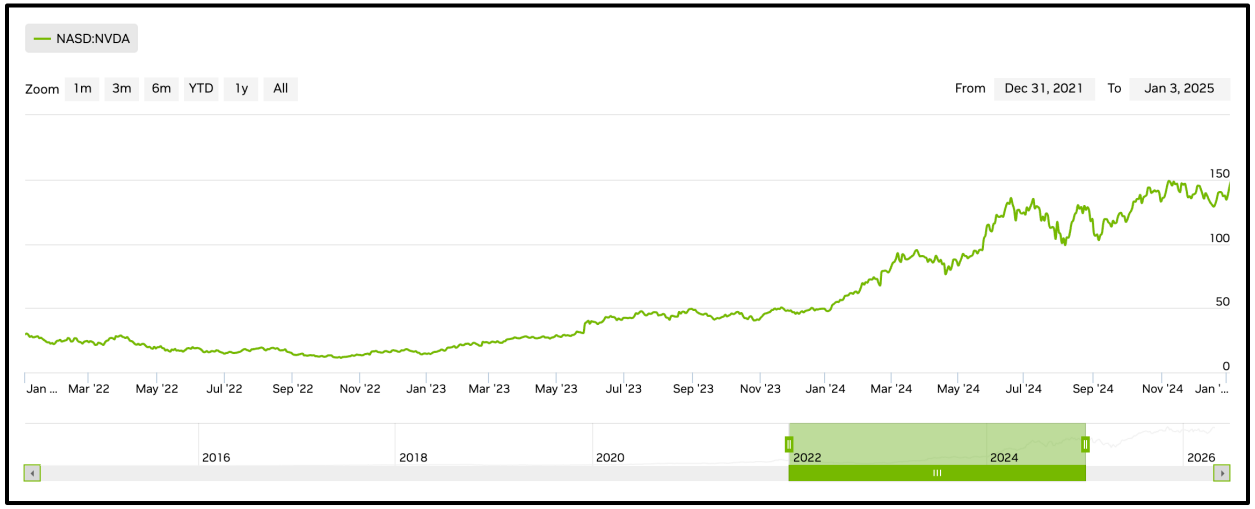
https://investor.nvidia.com/stock-info/stock-quote-and-chart/default.aspx

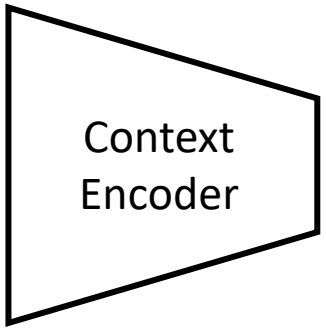
Why these happened

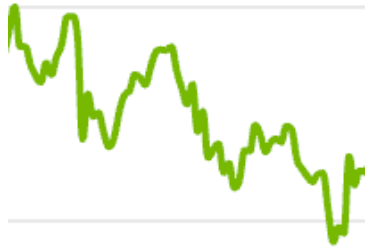












DeepSeek-R1 Release

https://investor.nvidia.com/stock-info/stock-quote-and-chart/default.aspx

Missing component: Imperfect information

IDEAS AI

Spatial Intelligence Is AI's Next Frontier

ADD TIME ON GOOGLE

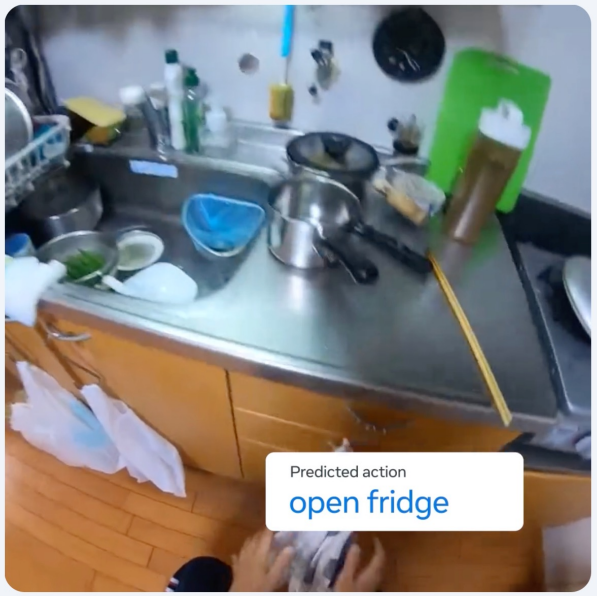
by Fei-Fei Li

Li is co-director of Stanford's Human-Centered AI Institute and co-founder CEO of World Labs

DEC 11, 2025 7:52 AM ET

Language + 3D scene

By Prof. FeiFei Li



Unlock world understanding

V-JEPA 2 delivers exceptional motion understanding as well as leading visual reasoning capabilities when combined with language modeling.

Anticipate what's next

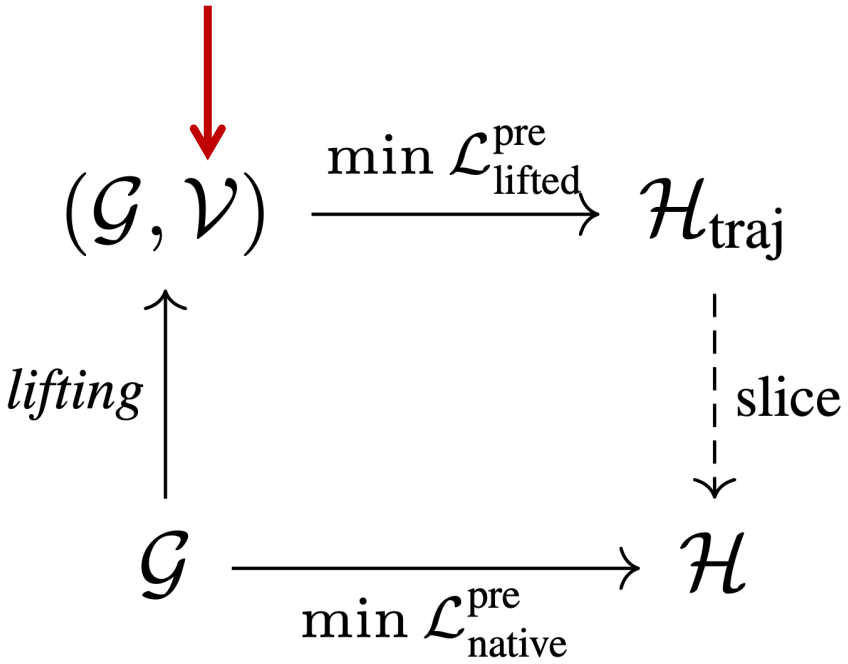
V-JEPA 2 can make predictions about how the world will evolve, setting a new state-of-the-art in anticipating actions from contextual cues.

Language + Interaction

By Prof. Yann LeCun

Lifted framework: The initial idea

Augmented Space



GeoPT: Scaling Physics Simulation via Lifted Geometric Pre-Training

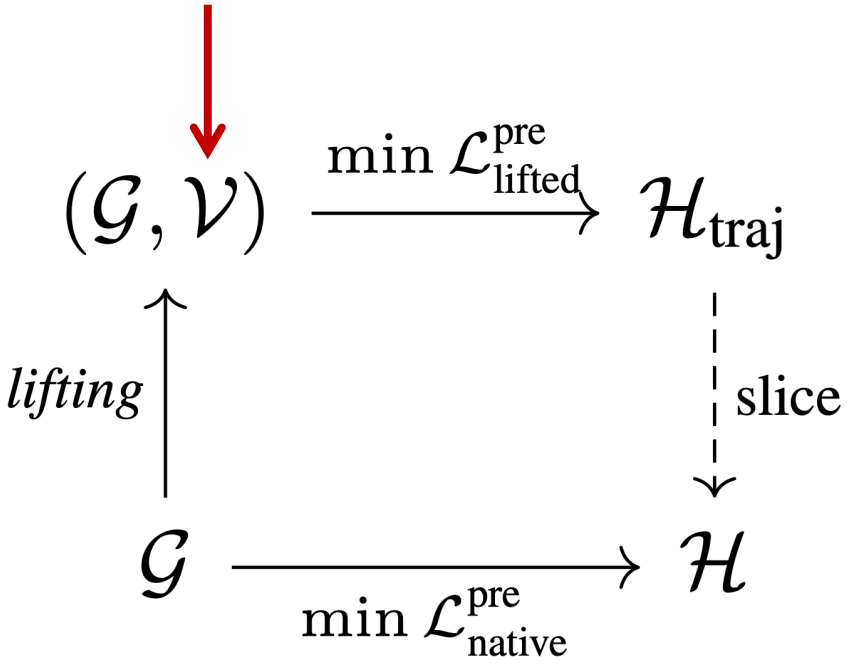
Haixu Wu* Minghao Guo* Zongyi Li Zhiyang Dou Mingsheng Long Kaiming He Wojciech Matusik

(* = equal contribution)

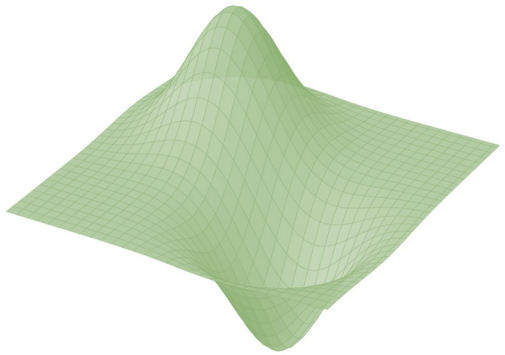
Wu et al., GeoPT: Scaling Physics Simulation via Lifted Geometric Pre-Training, ICML 2026 (ICLR FM4Science Best Paper)

Lifted framework

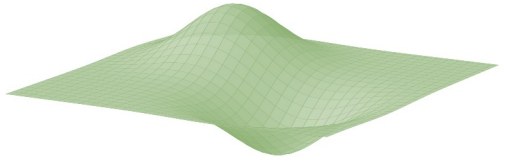
Augmented Space: Go beyond simple pattern recognition and learn intelligence.



True mechanisms lie in:



Real data manifold

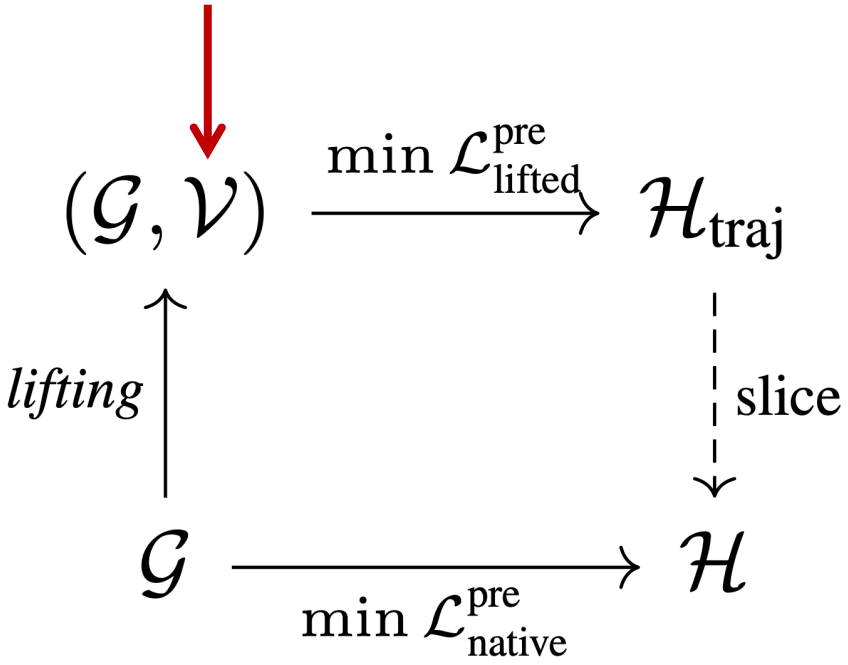


Observation manifold
(Smoothed, simplified)

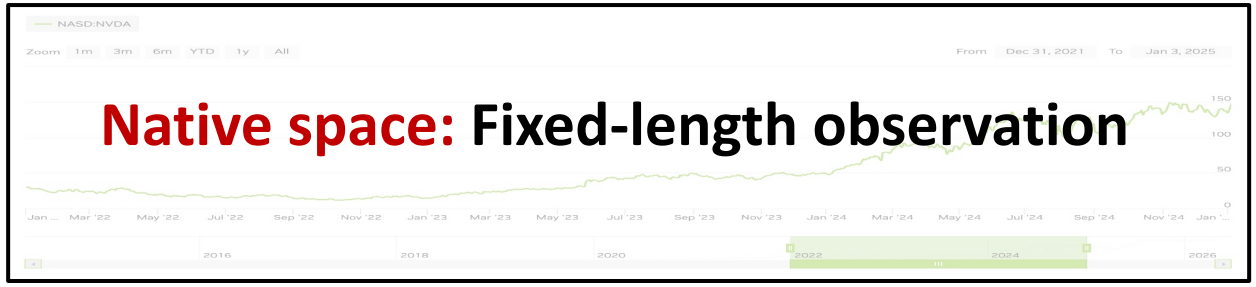
Wu et al., GeoPT: Scaling Physics Simulation via Lifted Geometric Pre-Training, ICML 2026 (ICLR FM4Science Best Paper)

Lifted framework for time series

Augmented Space: Go beyond simple pattern recognition and learn intelligence.



Lifted space: longer context, covariates…

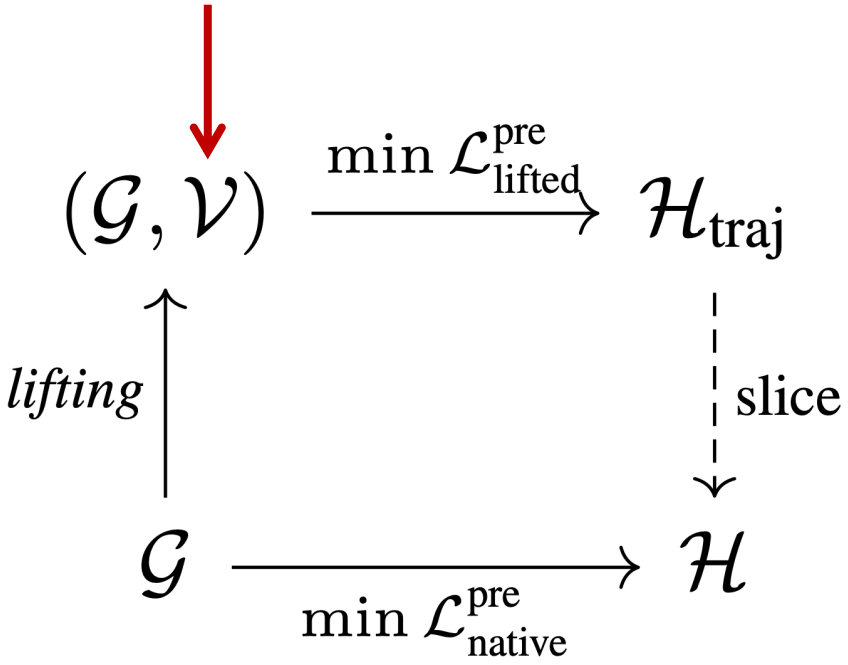


Native space: Fixed-length observation

Wu et al., GeoPT: Scaling Physics Simulation via Lifted Geometric Pre-Training, ICML 2026 (ICLR FM4Science Best Paper)

Lifted framework for time series

Augmented Space: Go beyond simple pattern recognition and learn intelligence.



Q1: How to effectively augment the observation?

Q2: How to learn from high-dimensional space?

Wu et al., GeoPT: Scaling Physics Simulation via Lifted Geometric Pre-Training, ICML 2026 (ICLR FM4Science Best Paper)

Lifted Framework in Dynamical Systems (Time Series)

$p(\mathbf{x})$ ---------- **Vanilla Time Series Forecasting**

$p(\mathbf{x}, \mathbf{x}_{\text{long-term}})$ ------ **Long-term Forecasting** — Autoformer

$p(\mathbf{x}, \text{ex})$ ---------- **Forecasting with Exogenous Variables** — TimeXer

$p(\mathbf{x}|\mathbf{z})$ ---------- **Large-scale Pre-training** — Some Discussion

* Omit the shared conditional variables, such as past observations.

Lifted Framework in Dynamical Systems (Time Series)

$p(\mathbf{x})$ - - - - - - - - - - - - Vanilla Time Series Forecasting

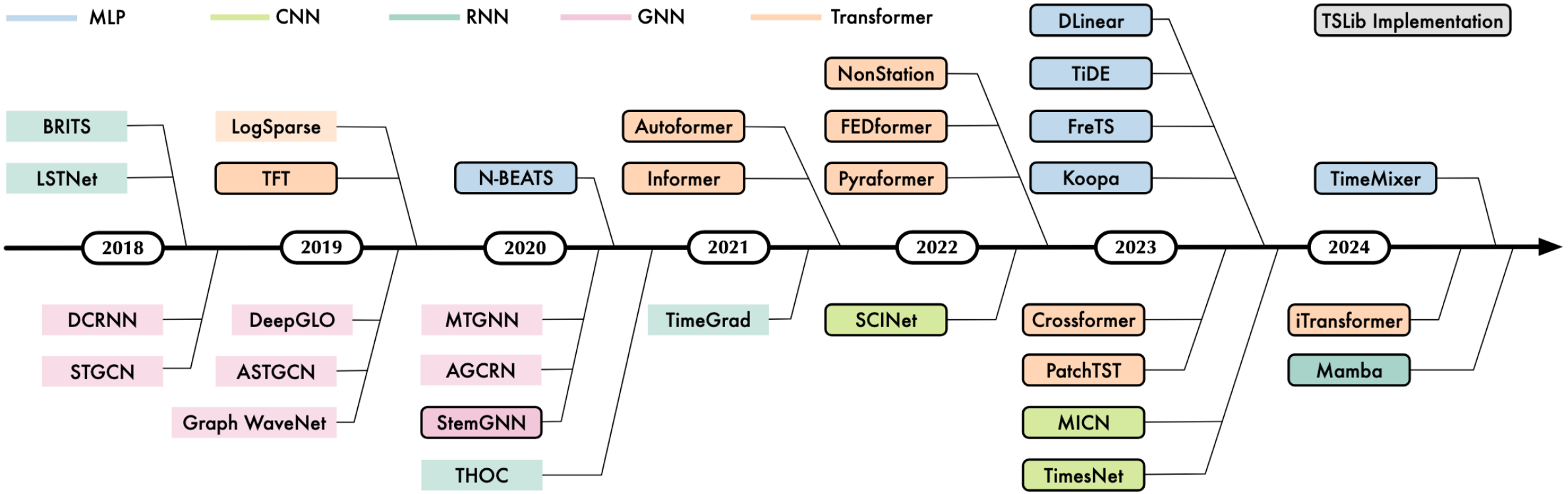
$p(\mathbf{x}, \mathbf{x}_{\text{long-term}})$ - - - - - - **Long-term Forecasting** — Autoformer

$p(\mathbf{x}, \text{ex})$ - - - - - - - - - - - - - - Forecasting with Exogenous Variables — TimeXer

$p(\mathbf{x}|\mathbf{z})$ - - - - - - - - - - - - - - Large-scale Pre-training — Some Discussion

* Omit the shared conditional variables, such as past observations.

Part 1. Long-term Time Series Forecasting



https://github.com/thuml/Time-Series-Library

Part 1. Long-term Time Series Forecasting

Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting

Haixu Wu, Jiehui Xu, Jianmin Wang, Mingsheng Long (✉)
School of Software, BNRist, Tsinghua University, China
{whx20,xjh20}@mails.tsinghua.edu.cn, {jimwang,mingsheng}@tsinghua.edu.cn

Haixu Wu | Jiehui Xu | Jianmin Wang | Mingsheng Long

[Autoformer, NeurIPS 2021, 6000+ citations]

Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting

Yong Liu*, Haixu Wu*, Jianmin Wang, Mingsheng Long✉
School of Software, BNRist, Tsinghua University, China
{liuyong21,whx20}@mails.tsinghua.edu.cn, {jimwang,mingsheng}@tsinghua.edu.cn

Yong Liu | Haixu Wu | Jianmin Wang | Mingsheng Long

[Non-Stationary Transformers, NeurIPS 2022, 1000+ citations]

Published as a conference paper at ICLR 2023

TIMESNET: TEMPORAL 2D-VARIATION MODELING FOR GENERAL TIME SERIES ANALYSIS

Haixu Wu*, Tengge Hu*, Yong Liu*, Hang Zhou, Jianmin Wang, Mingsheng Long✉
School of Software, BNRist, Tsinghua University, Beijing 100084, China
{whx20,liuyong21,htg21,h-zhou18}@mails.tsinghua.edu.cn
{jimwang,mingsheng}@tsinghua.edu.cn

Haixu Wu | Tengge Hu | Yong Liu | Hang Zhou | Jianmin Wang | Mingsheng Long

[TimesNet, ICLR 2023, 3000+ citations]

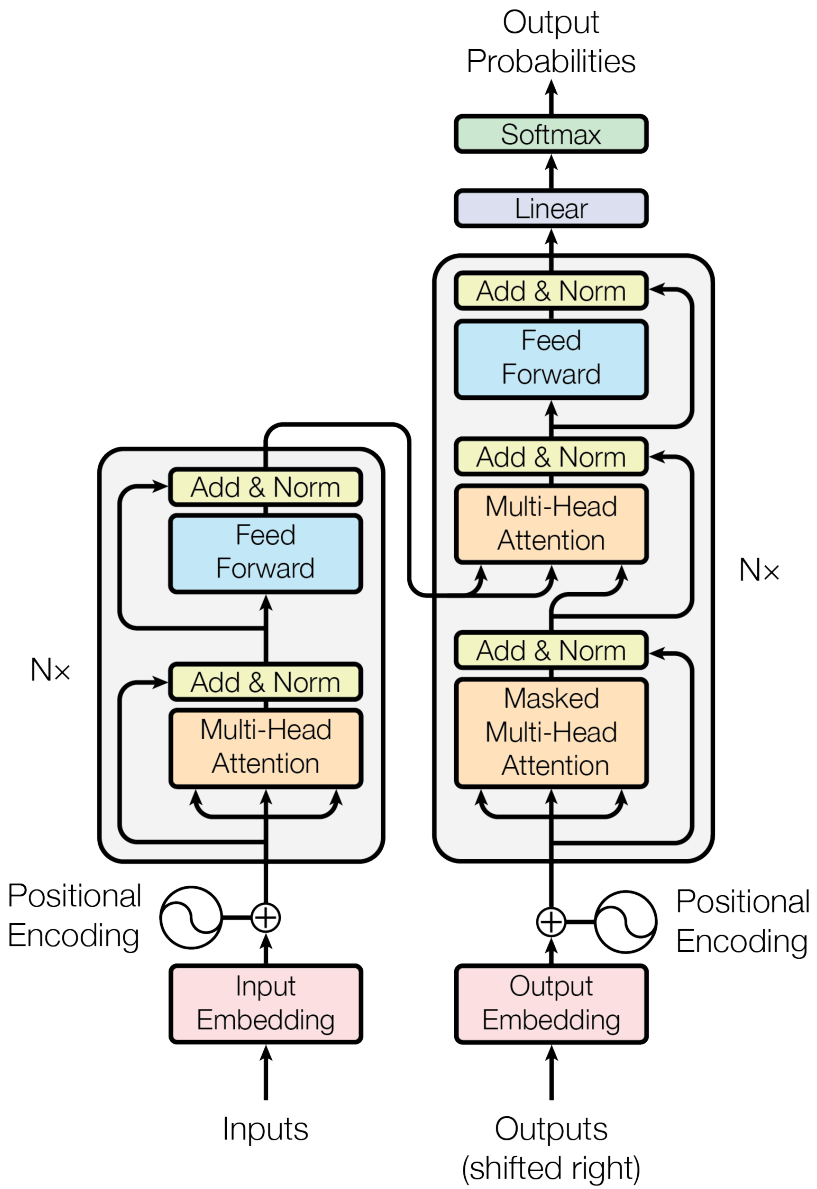
Published as a conference paper at ICLR 2024

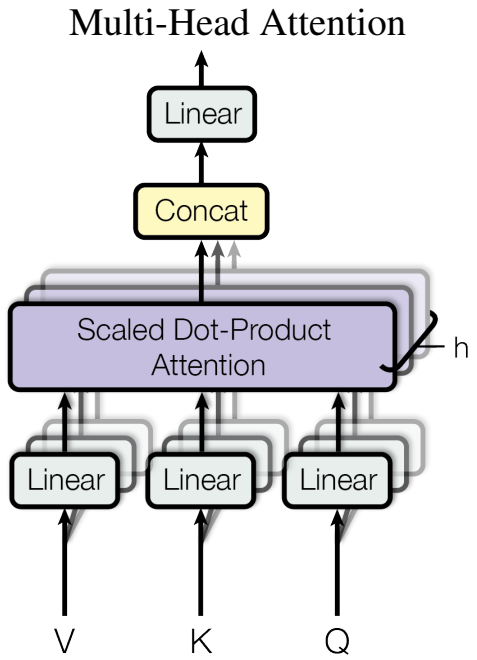
TIMEMIXER: DECOMPOSABLE MULTISCALE MIXING FOR TIME SERIES FORECASTING

Shiyu Wang[1]*, Haixu Wu[2]*, Xiaoming Shi[1], Tengge Hu[2], Huakun Luo[2], Lintao Ma[1]✉, James Y. Zhang[1], Jun Zhou[1]✉
[1]Ant Group, Hangzhou, China [2]Tsinghua University, Beijing, China
{weiming.wsy,lintao.mlt,peter.sxm,james.z,jun.zhoujun}@antgroup.com,
{wuhx23,htg21,luohk19}@mails.tsinghua.edu.cn

[TimeMixer, ICLR 2024, 1000+ citations]

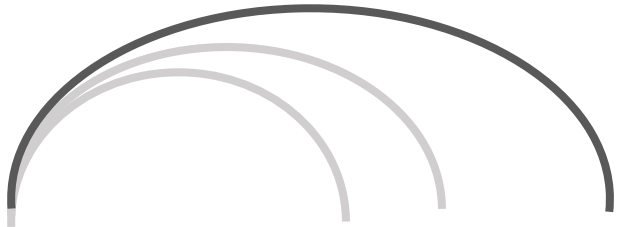
Problem definition





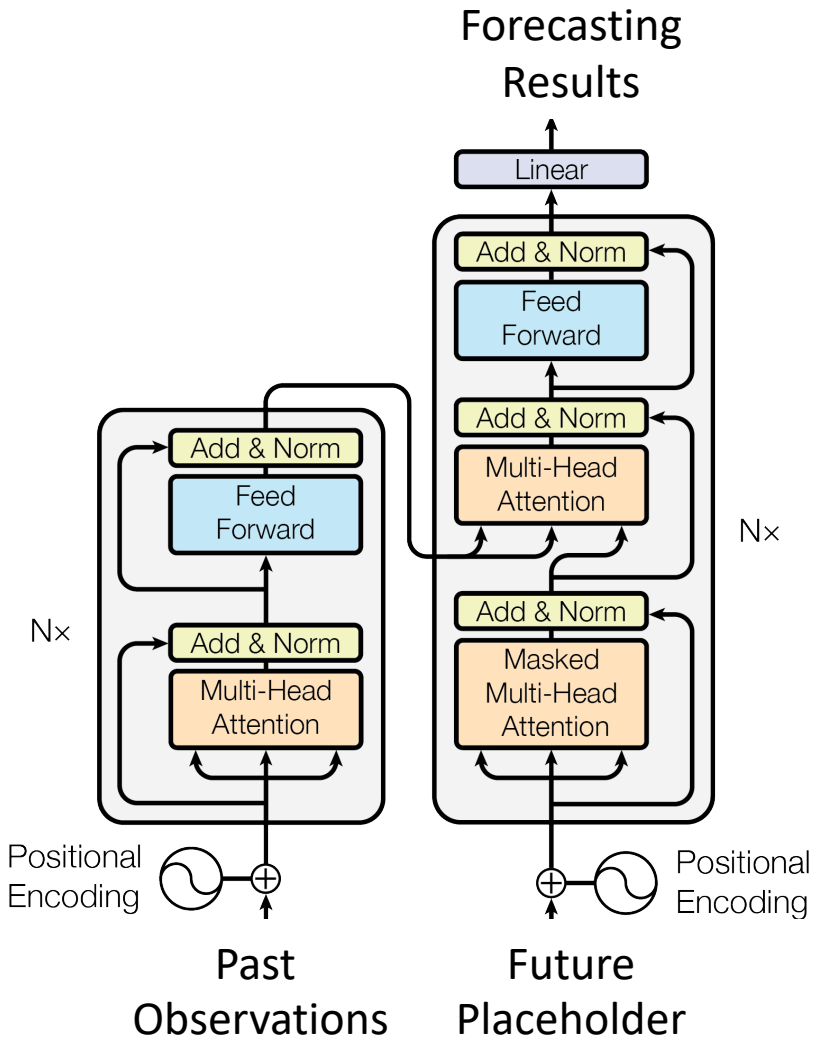
Modeling the relation of words

with **point-wise Self-Attention**

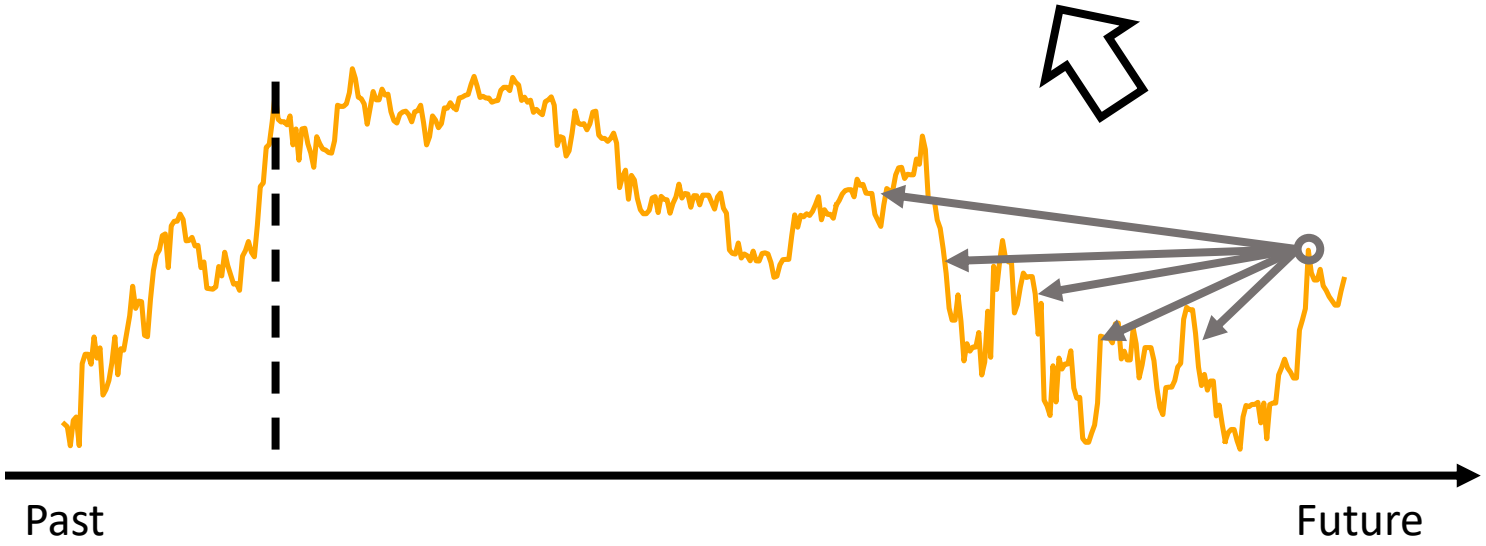


Autoformer is for long-term Forecasting.

Golden problem: Temporal correlations



- Modeling the **temporal dependencies** with **point-wise Self-Attention**
- Aggregate the representations for forecasting



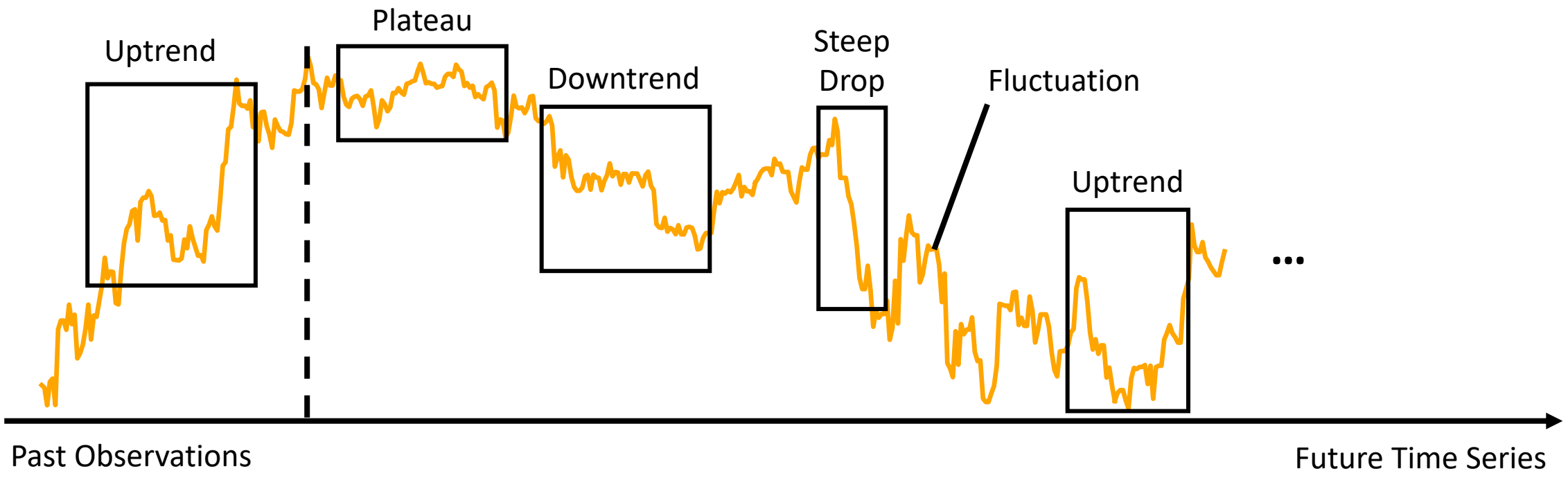
Informer [Zhou et al. AAAI2021], Reformer [Kitaev et al. ICLR2020], LogTrans [Li et al. NeurIPS19]

Autoformer (NeurIPS 2021, 5 years ago)

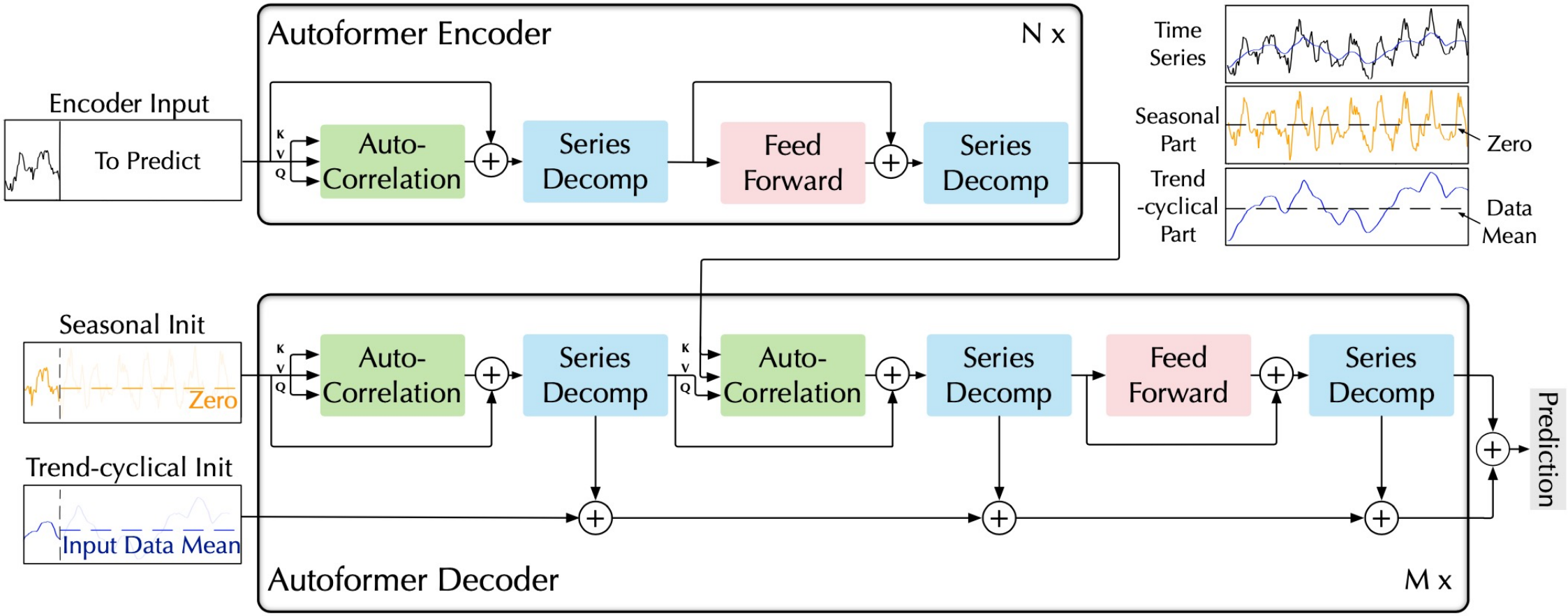
Longer Forecasting Horizon

Intricate Temporal Patterns

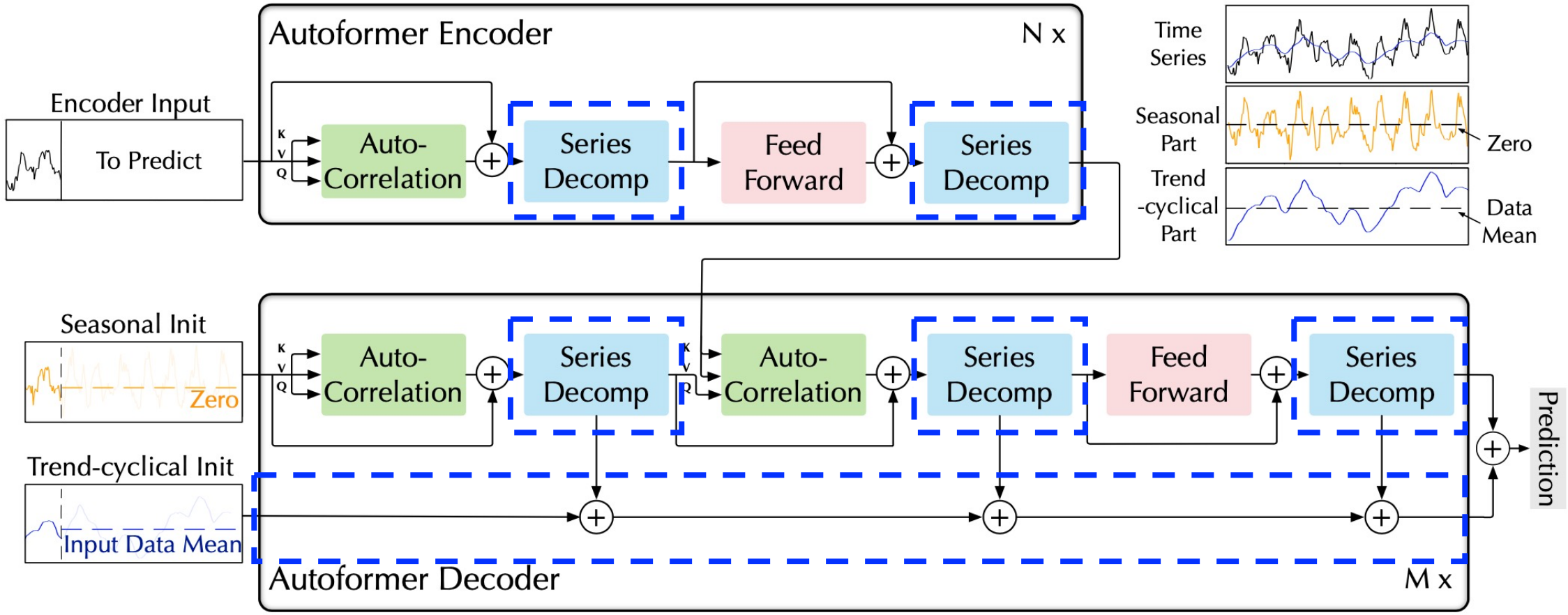
Deal with Long Series (complexity)



Overall Architecture



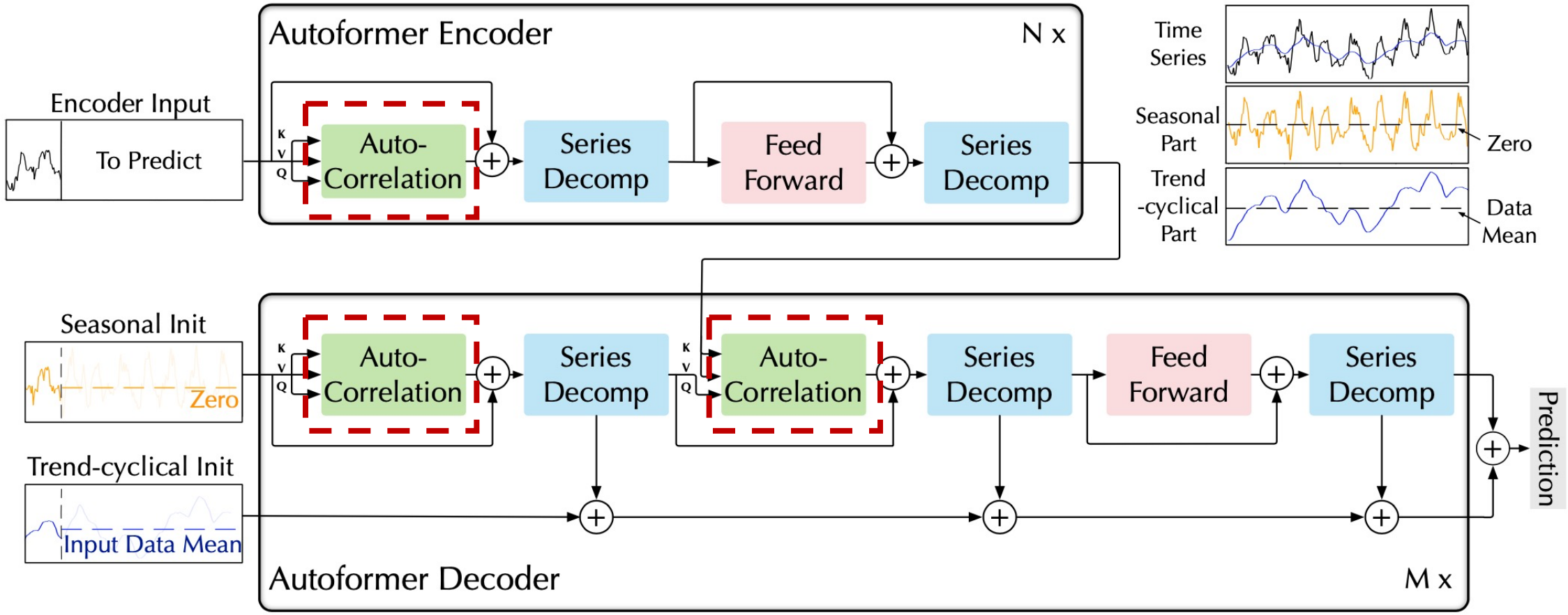
Deep decomposition architecture



Decomposition architecture for intricate temporal patterns.

Be expanded by Non-stationary Transformer and gradually become the standard usage in normalization.

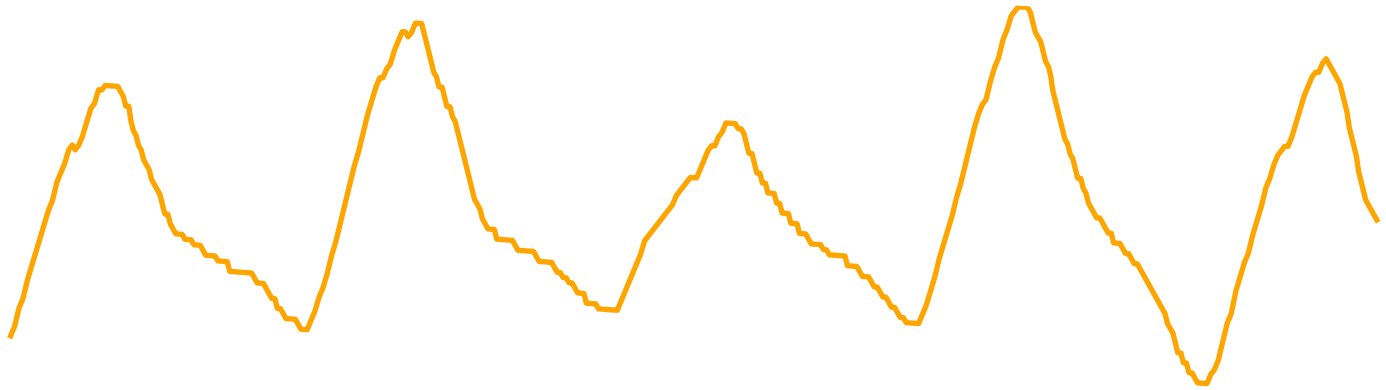
Series-wise Auto-Correlation mechanism



Series-wise Auto-Correlation for temporal correlation modeling.

The ideas of frequency domain analysis and series-wise modeling are widely used in subsequent models.

Series-wise Auto-Correlation mechanism



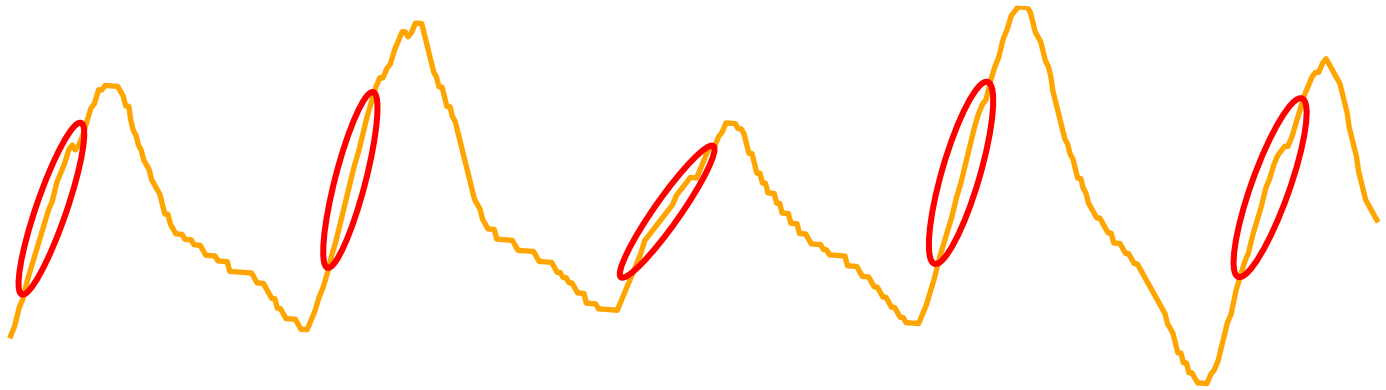
Benefited from the deep decomposition,
the **seasonal part** is highlighted with **periodicity.**

Conducts the **dependencies discovery** and **representation aggregation** at the series level.

Series-wise Auto-Correlation mechanism

Period-based dependencies

The same phase position of different periods



Benefited from the deep decomposition,
the **seasonal part** is highlighted with **periodicity.**

Conducts the **dependencies discovery** and **representation aggregation** at the series level.

Series-wise Auto-Correlation mechanism

Discover period-based dependencies with autocorrelation in stochastic process:

$$\mathcal{R}_{\mathcal{X}\mathcal{X}}(\tau) = \lim_{L\to\infty} \frac{1}{L} \sum_{t=0}^{L-1} \mathcal{X}_t \mathcal{X}_{t-\tau}.$$

Autocorrelation reflects the time delay similarity,

and corresponds to the confidence of period estimation.

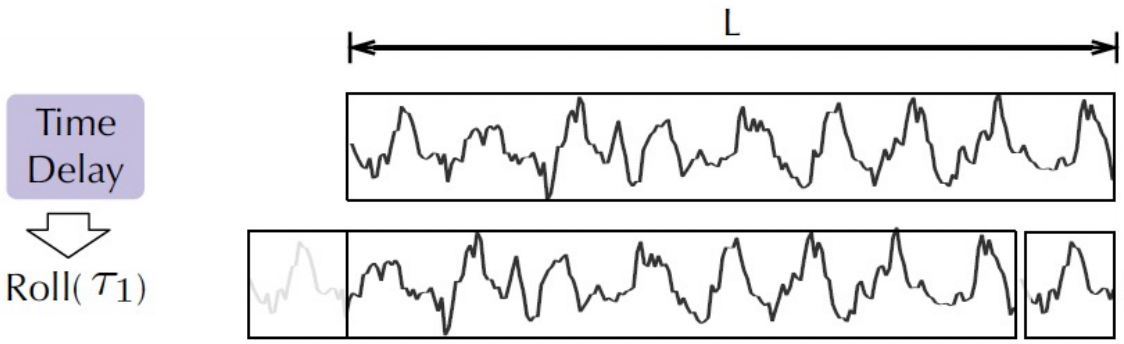
Series-wise Auto-Correlation mechanism

Discover period-based dependencies with autocorrelation in stochastic process:

$$\mathcal{R}_{\mathcal{X}\mathcal{X}}(\tau) = \lim_{L\to\infty} \frac{1}{L} \sum_{t=0}^{L-1} \mathcal{X}_t \mathcal{X}_{t-\tau}.$$

Autocorrelation reflects the time delay similarity,

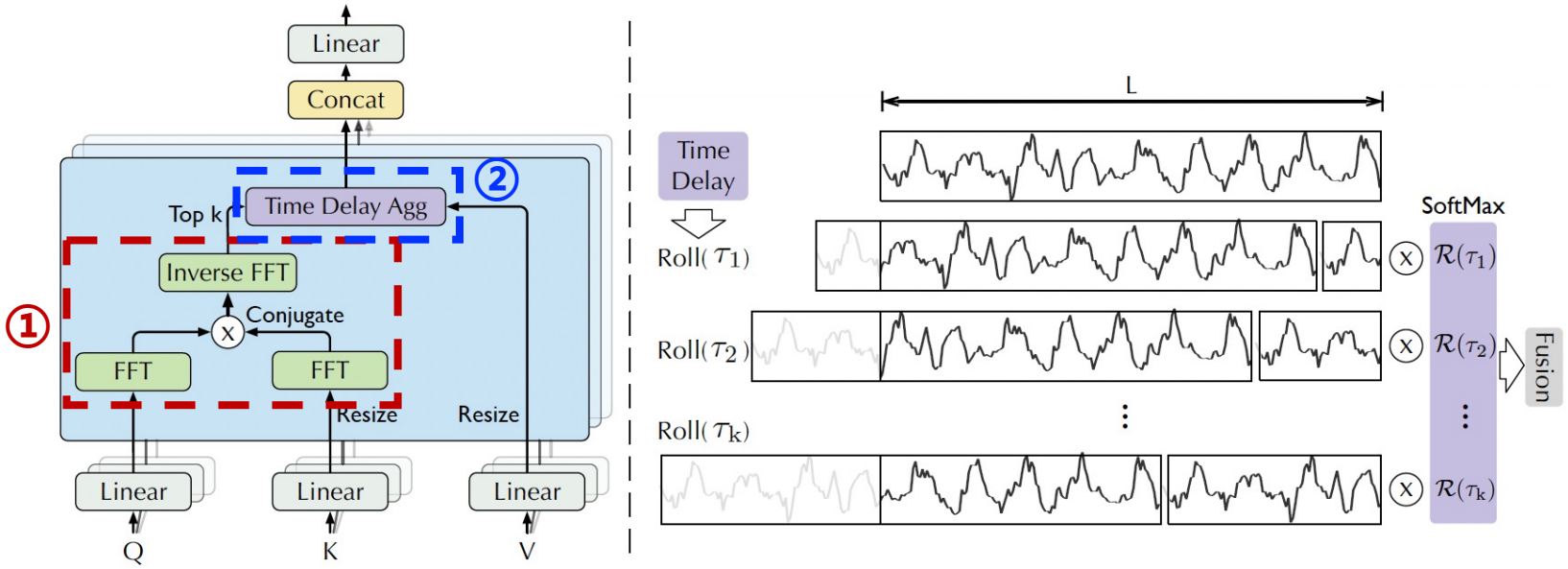
and corresponds to the confidence of period estimation.

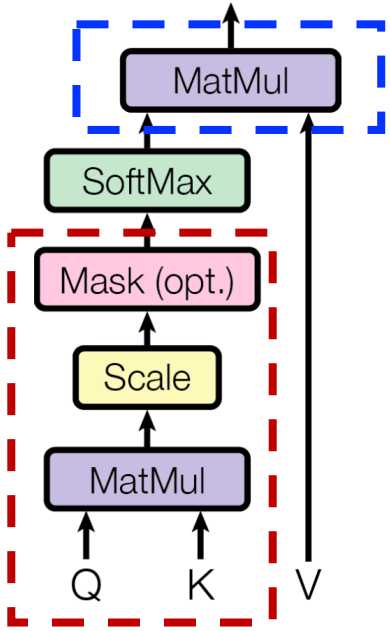


Larger autocorrelation $\mathcal{R}(\tau)$ means

- stronger time delay similarity w.r.t. τ
- more confidence of period length as τ

Auto-Correlation: Architecture design

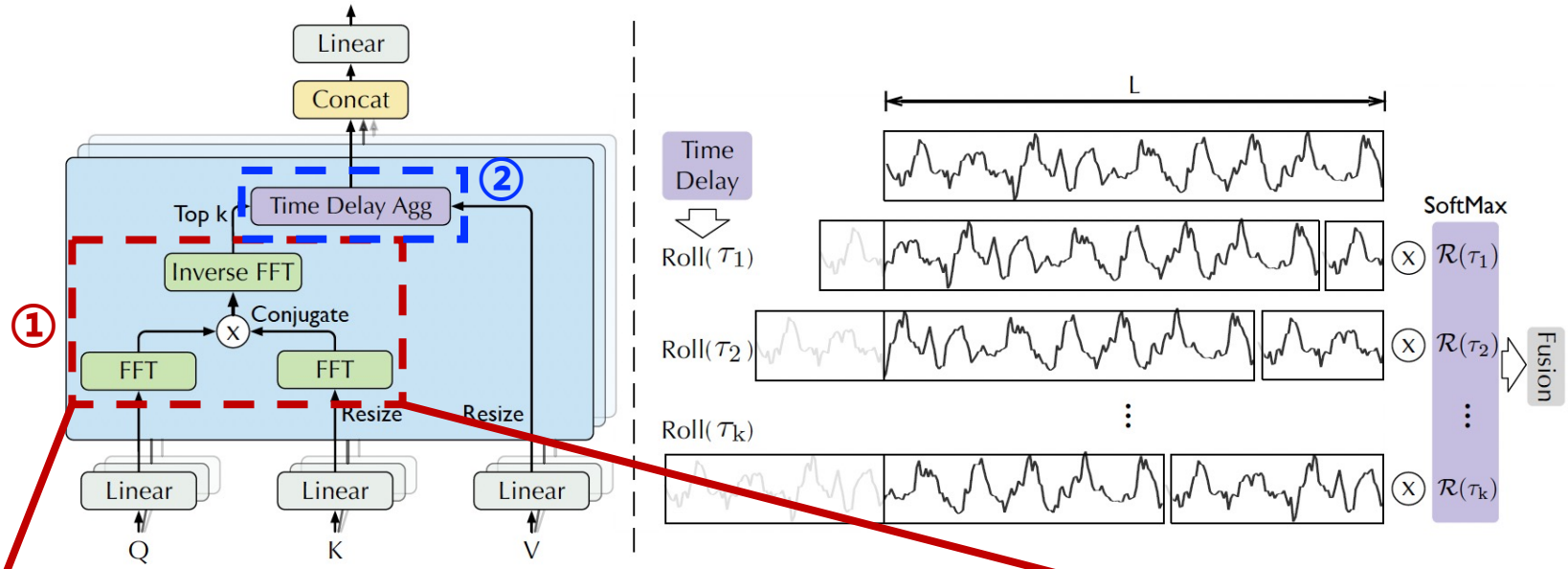




① Discover period-based dependencies

② Aggregate similar sub-processes from different periods

Auto-Correlation: Architecture design



Efficient computation of autocorrelation

with *Wiener–Khinchin theorem* by FFT

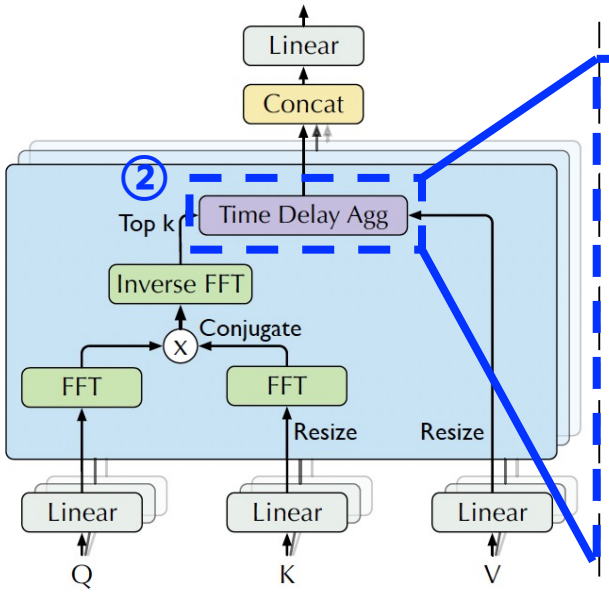
$$\mathcal{S}_{\mathcal{X}\mathcal{X}}(f) = \mathcal{F}(\mathcal{X}_t)\,\mathcal{F}^*(\mathcal{X}_t) = \int_{-\infty}^{\infty} \mathcal{X}_t e^{-i2\pi t f}\,\mathrm{d}t \overline{\int_{-\infty}^{\infty} \mathcal{X}_t e^{-i2\pi t f}\,\mathrm{d}t}$$

$$\mathcal{R}_{\mathcal{X}\mathcal{X}}(\tau) = \mathcal{F}^{-1}(\mathcal{S}_{\mathcal{X}\mathcal{X}}(f)) = \int_{-\infty}^{\infty} \mathcal{S}_{\mathcal{X}\mathcal{X}}(f) e^{i2\pi f \tau}\,\mathrm{d}f,$$

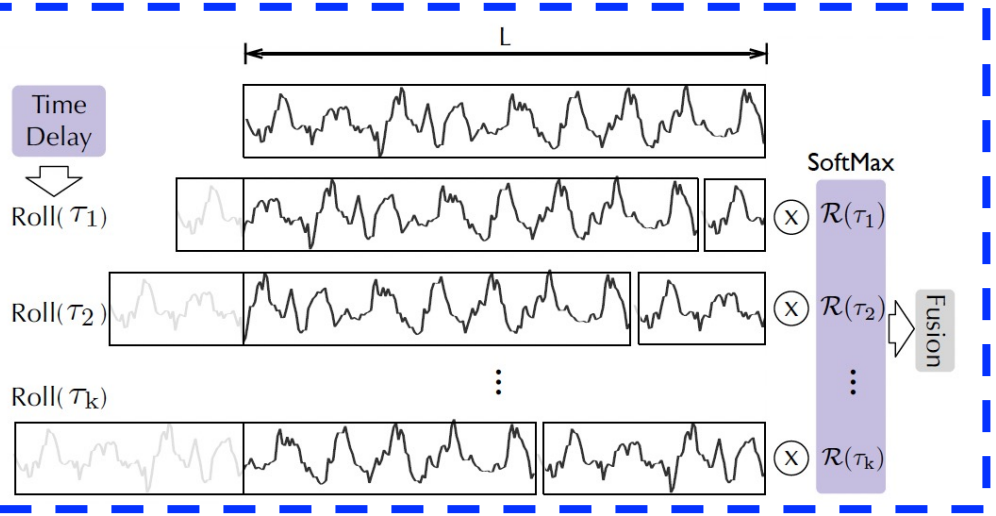
Discover period-based dependencies

with inherent $\mathrm{O}(L \log L)$ complexity

Auto-Correlation: Architecture design



Aggregate representations from similar sub-processes



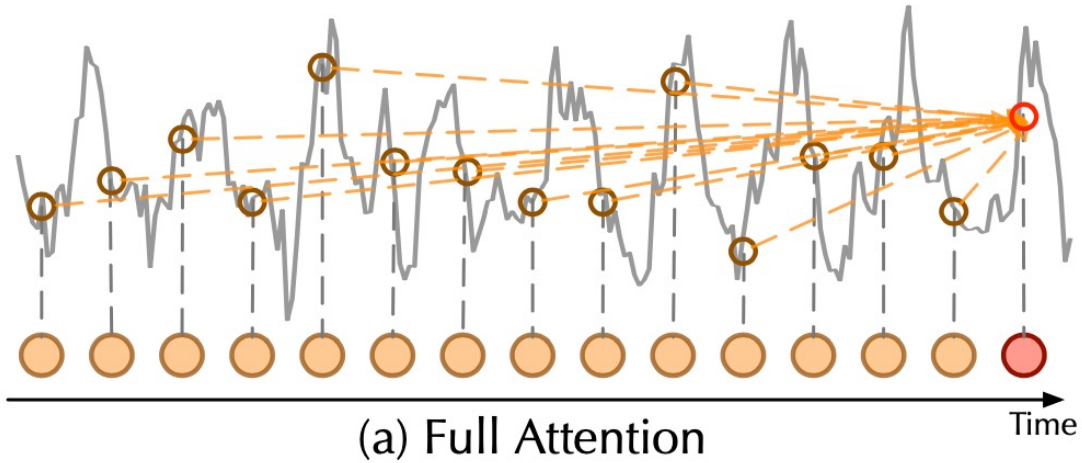
$$\tau_1, \cdots, \tau_k = \underset{\tau \in \{1, \cdots, L\}}{\arg \operatorname{Topk}} \left(\mathcal{R}_{\mathcal{Q},\mathcal{K}}(\tau) \right)$$ Select the Top k period lengths

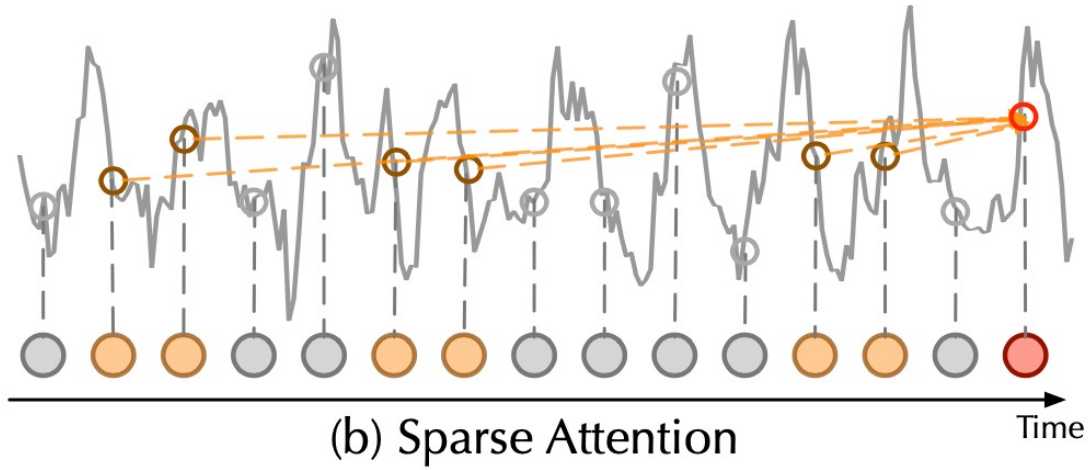
$$\widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}(\tau_1), \cdots, \widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}(\tau_k) = \operatorname{SoftMax}\left(\mathcal{R}_{\mathcal{Q},\mathcal{K}}(\tau_1), \cdots, \mathcal{R}_{\mathcal{Q},\mathcal{K}}(\tau_k) \right)$$ Normalization

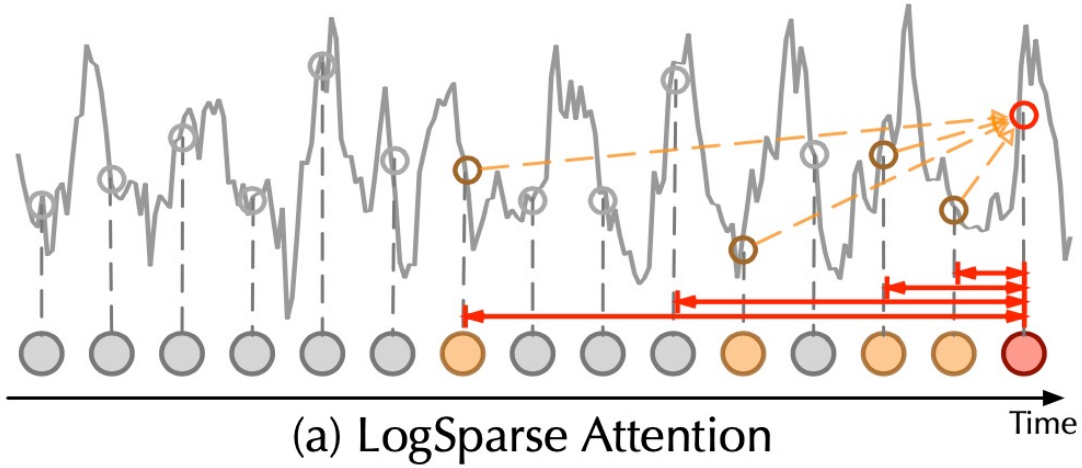
$$\operatorname{AutoCorrelation}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \sum_{i=1}^{k} \operatorname{Roll}(\mathcal{V}, \tau_k) \widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}(\tau_k),$$

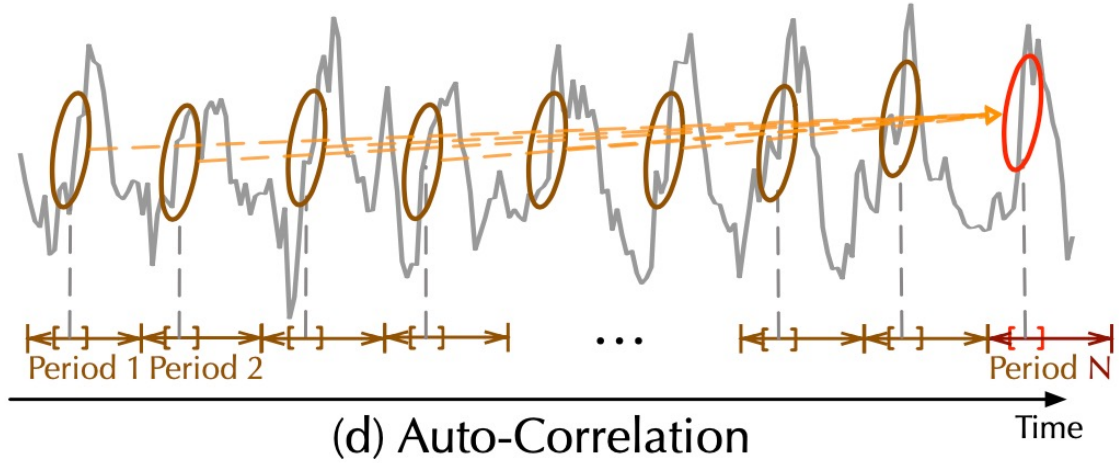
Align the delayed series,
Aggregate sub-series representations

Auto-Correlation v.s. Self-Attention Family



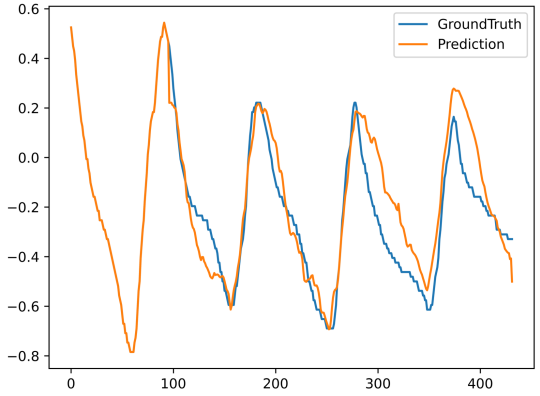




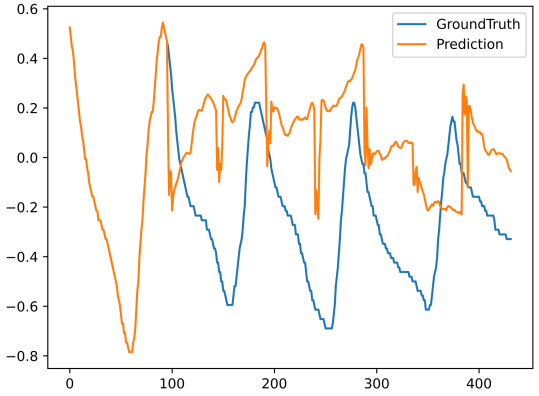


Auto-Correlation extends the point-wise aggregation to series-wise.

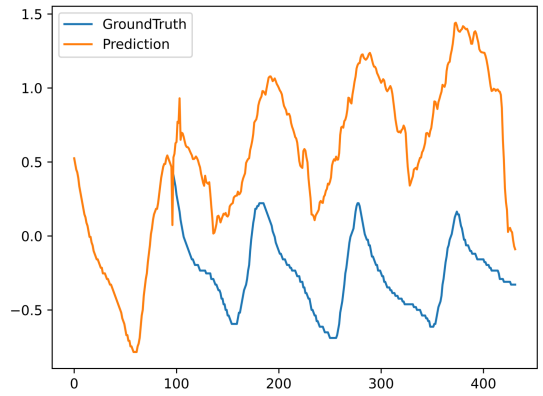
Some showcases



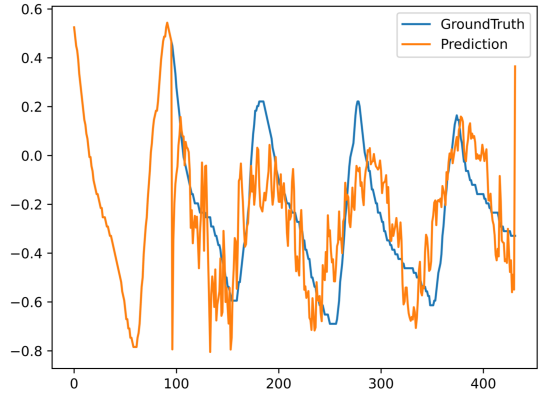
Autoformer



Informer

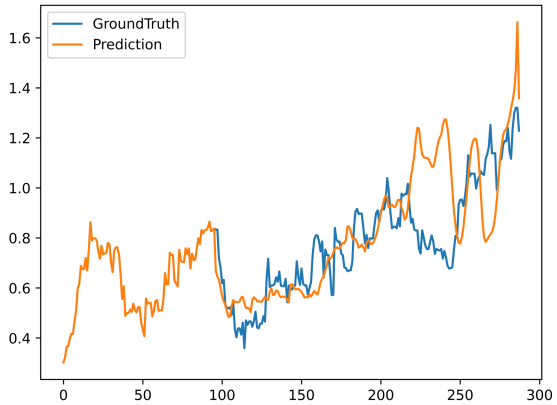


LogTrans

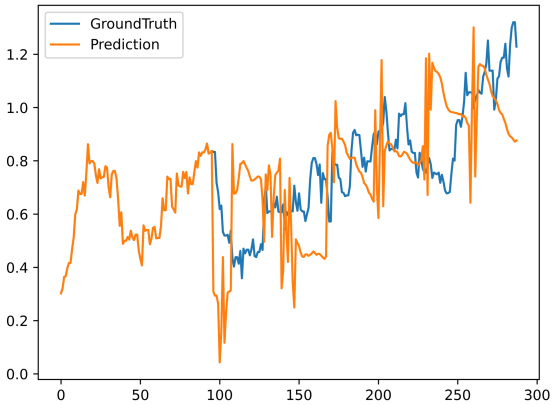


Reformer

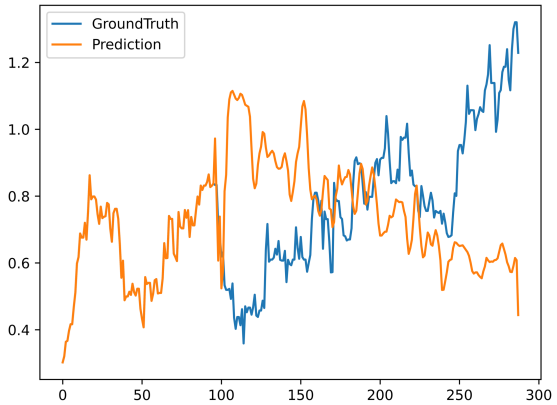
(1) ETT dataset with input-96-predict-336 (Energy, with obvious periodicity)



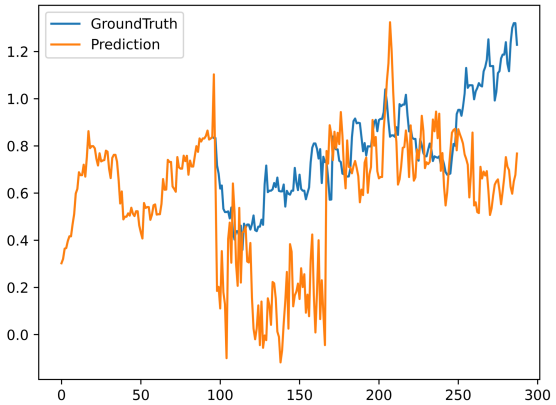
Autoformer



Informer



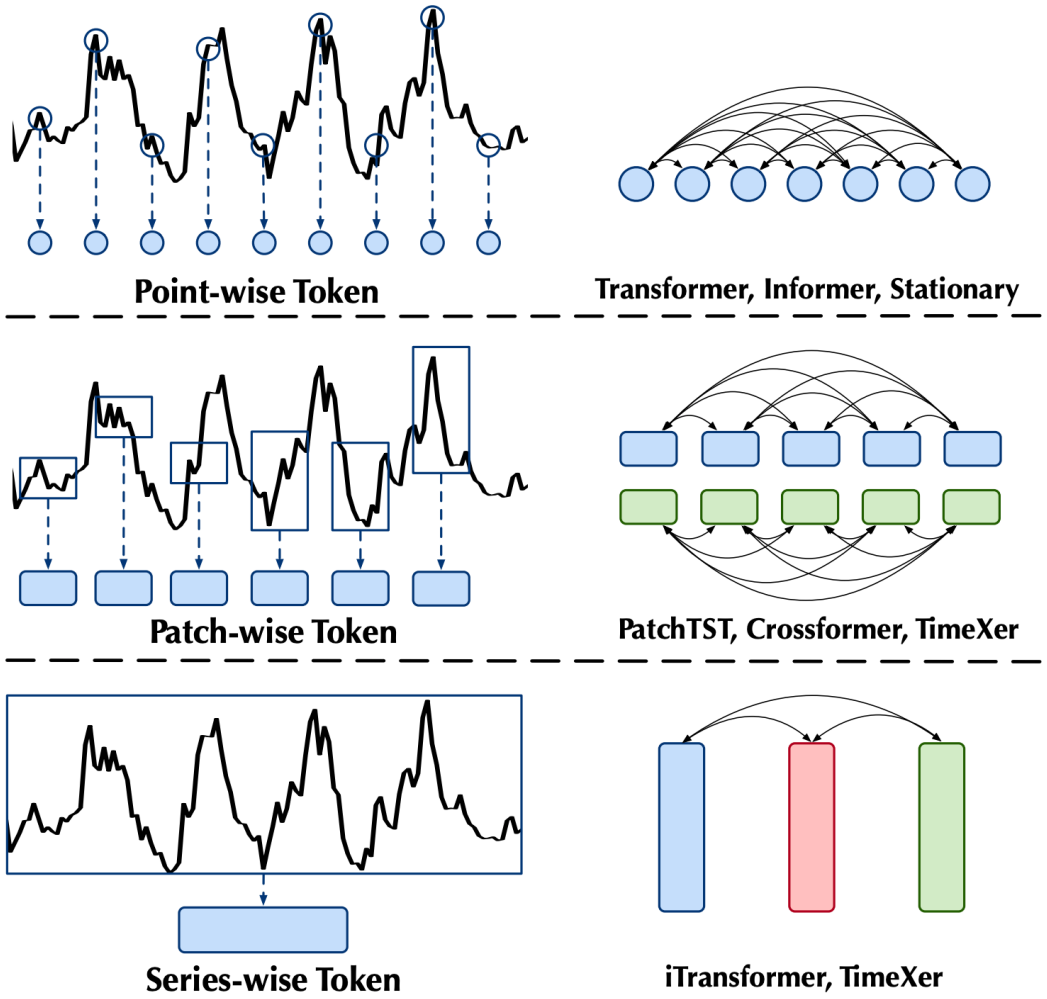
LogTrans



Reformer

(2) Exchange dataset with input-96-predict-192 (Economics, without obvious periodicity)

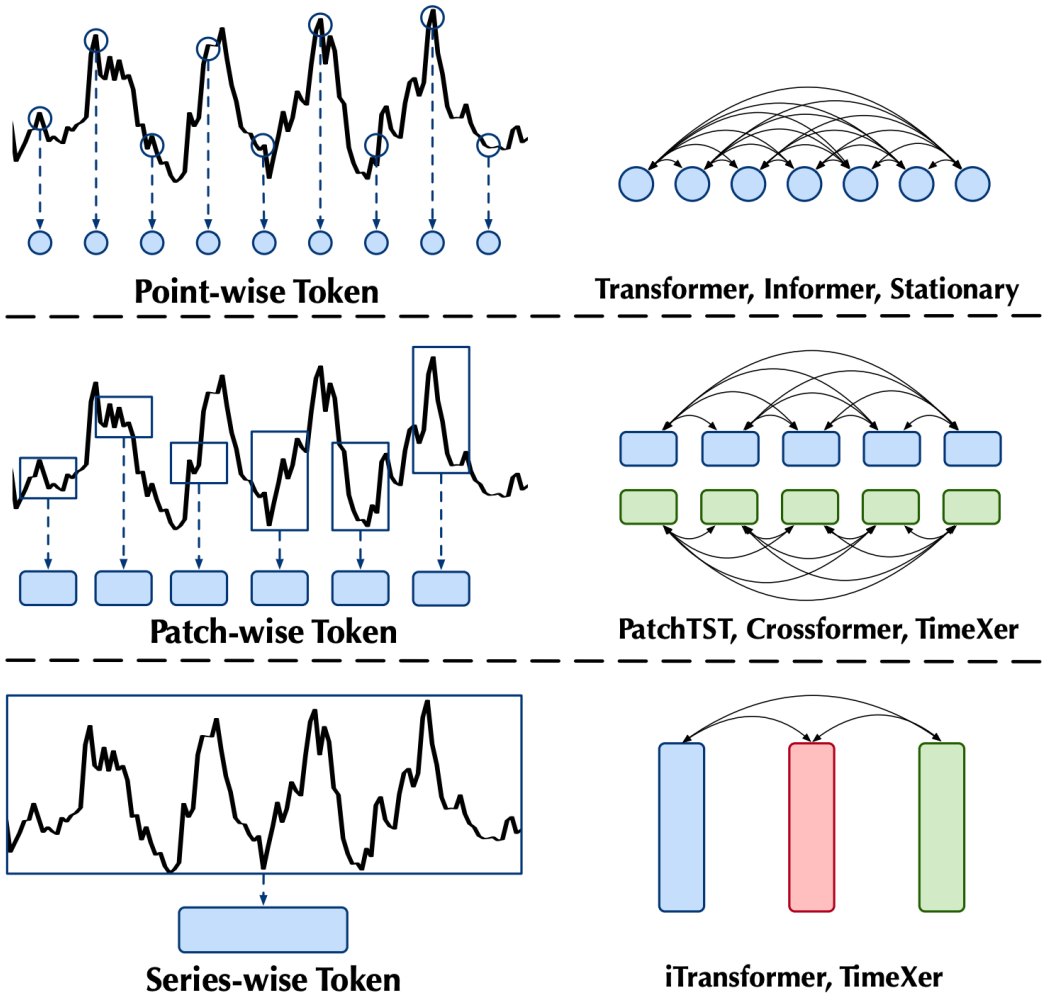
The modern way to do so…



- ✓ Change the input rather than the inside attention
- Include the series-wise dynamics in the model
- Better representation learning

PatchTST [Nie et al., ICLR 2023], iTransformer [Liu et al., ICLR 2024]

The modern way to do so…



✓ Change the input rather than the inside attention

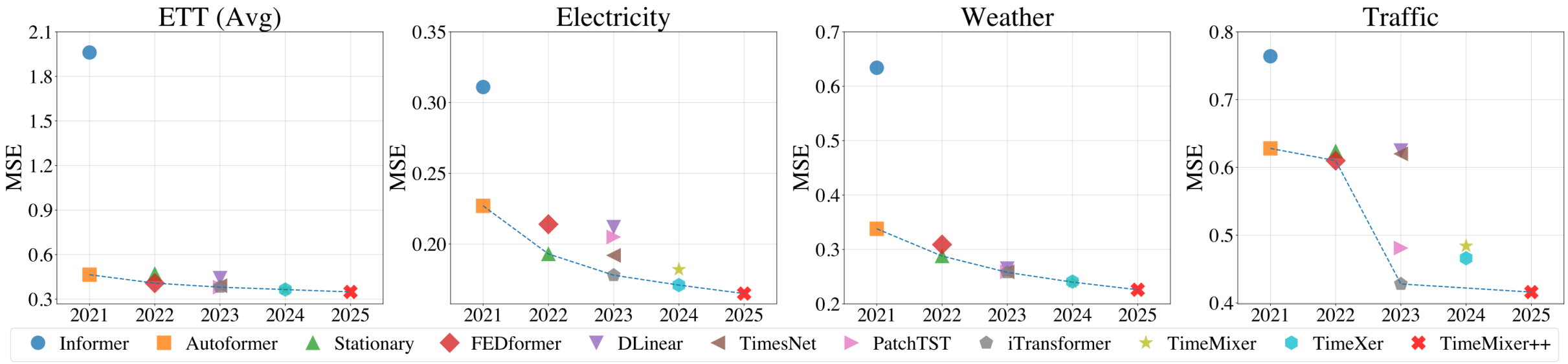
- Include the series-wise dynamics in the model
- Better representation learning

✓ Maintain the vanilla attention mechanism

- Potential scalability
- FlashAttention [NeurIPS 2022] can resolve efficiency

PatchTST [Nie et al., ICLR 2023], iTransformer [Liu et al., ICLR 2024]

The saturated accuracy



Time-series-only forecasting is nearly saturated.

The relative performance promotion is less than 3%.

Wang & Wu et al., Accuracy Law for the Future of Deep Time Series Forecasting. arXiv 2025

Accuracy law: The best accuracy that can achieve

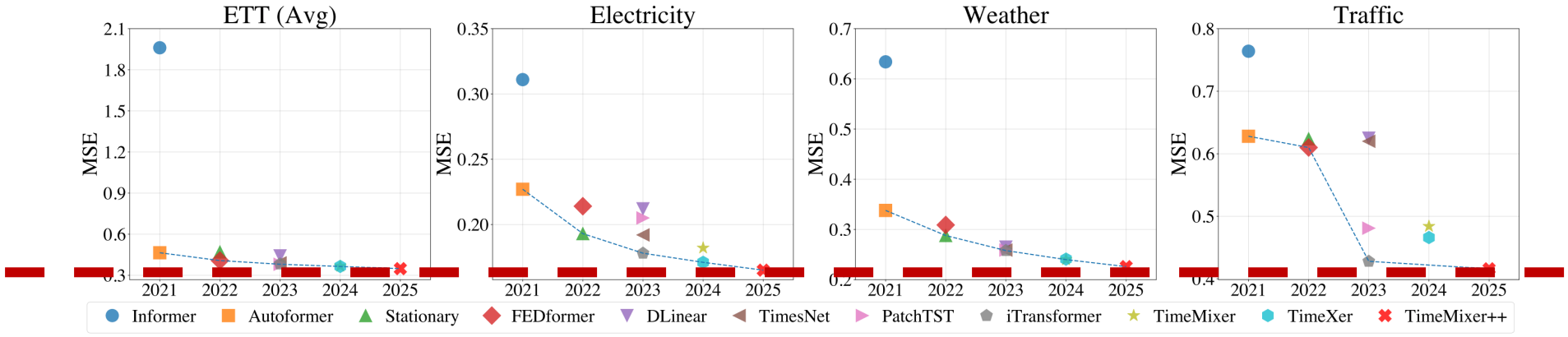
The relationship between **time series complexity** and **the best accuracy**

Accuracy law: The best accuracy that can achieve

The relationship between **time series complexity** and **the best accuracy**

- How to estimate the “best forecasting accuracy”?

The lowest accuracy achieved by state-of-the-art deep time series forecasting models.



Take the lowest value

Accuracy law: The best accuracy that can achieve

The relationship between **time series complexity** and **the best accuracy**

- How to estimate the “best forecasting accuracy”?

The lowest accuracy achieved by state-of-the-art deep time series forecasting models.

- How to measure the “time series complexity”?

Classical forecastability metrics: ADF, ACF, ForeCA

Accuracy law: The best accuracy that can achieve

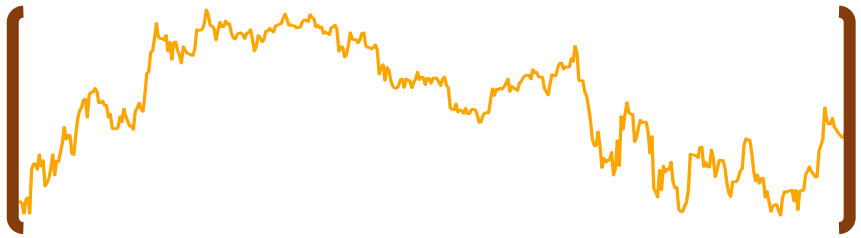
The relationship between **time series complexity** and **the best accuracy**

- How to estimate the “best forecasting accuracy”?

The lowest accuracy achieved by state-of-the-art deep time series forecasting models.

- How to measure the “time series complexity”?

Classical forecastability metrics: ADF, ACF, ForeCA



Series-wise metric

Sliding-window prediction paradigm

Wang & Wu et al., Accuracy Law for the Future of Deep Time Series Forecasting. arXiv 2025

Accuracy law: The best accuracy that can achieve

The relationship between **time series complexity** and **the best accuracy**

- How to estimate the “best forecasting accuracy”?

The lowest accuracy achieved by state-of-the-art deep time series forecasting models.

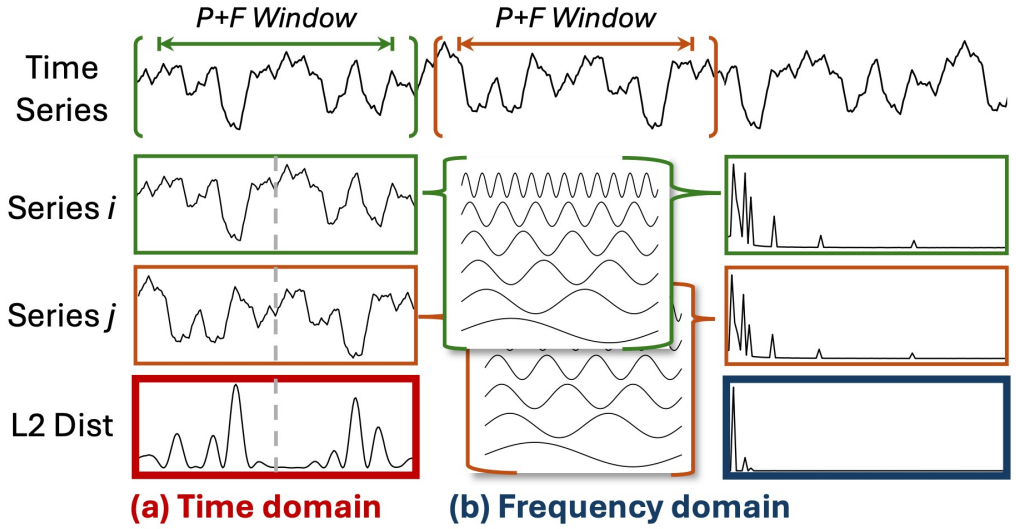
- How to measure the “time series complexity”?

Classical forecastability metrics: ADF, ACF, ForeCA

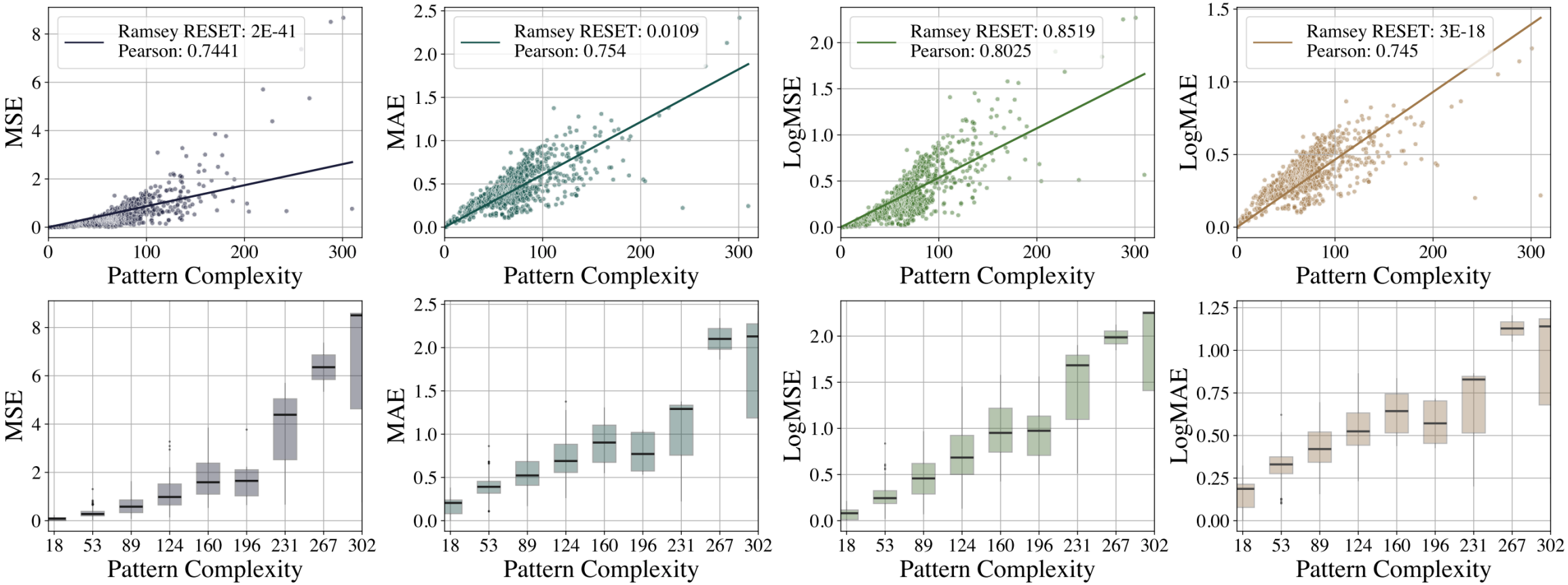
*Ours (**window-wise pattern complexity**)*

$$\{\mathbf{x}_{i:(i+P+F)}\}_i = \text{Split}(\mathbf{x}),\ \{\mathbf{A}_i\}_i = \left\{\text{Amp}\left(\text{FFT}\left(\mathbf{x}_{i:(i+P+F)}\right)\right)\right\}_i$$

$$\text{Complexity}(\mathbf{x}) = \text{tr}(\text{Cov}(\{\mathbf{A}_i\})) = \frac{1}{N}\sum\nolimits_{1\le i\le N} ||\mathbf{A}_i - \bar{\mathbf{A}}||_2^2.$$



Accuracy law (fitted from 3000+ experiments)

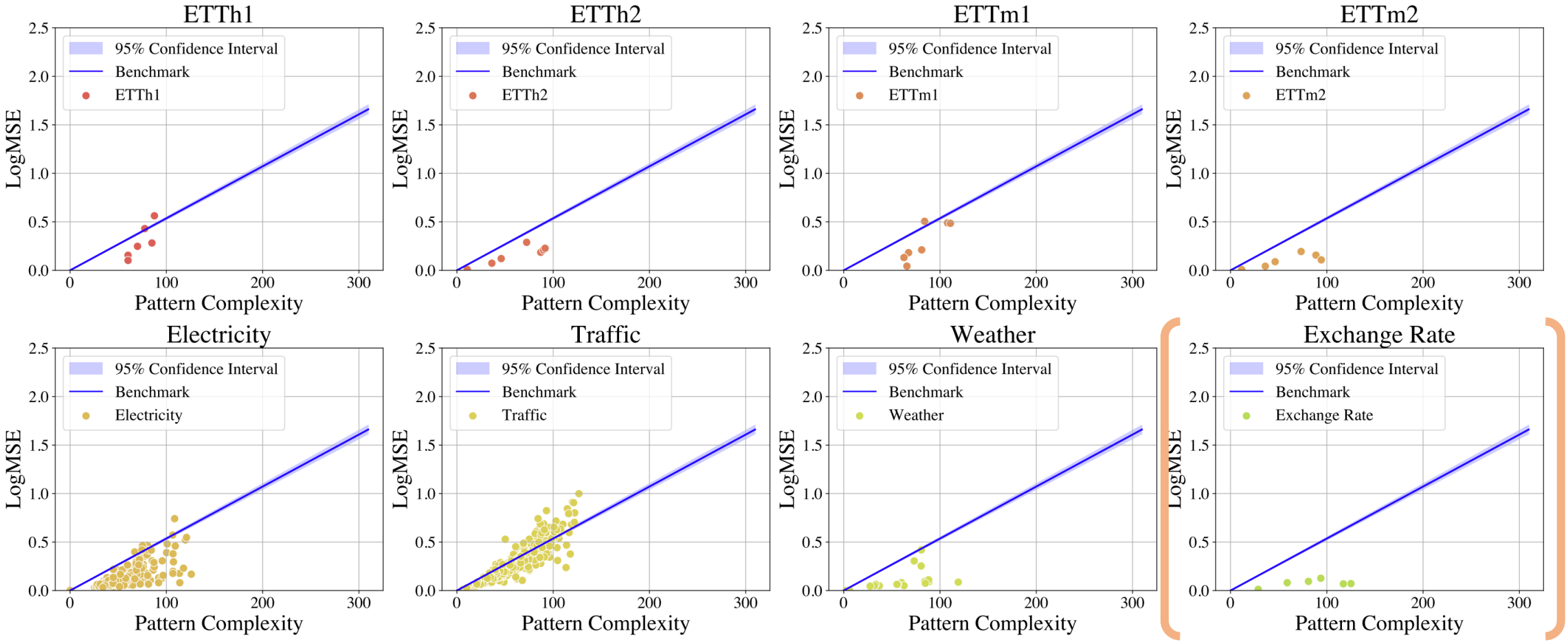


Univariate forecasting:

$$\mathrm{MSE} \approx \exp\left(\alpha \cdot \mathrm{Complexity}(\mathbf{x})\right) - 1$$

Wang & Wu et al., Accuracy Law for the Future of Deep Time Series Forecasting. arXiv 2025

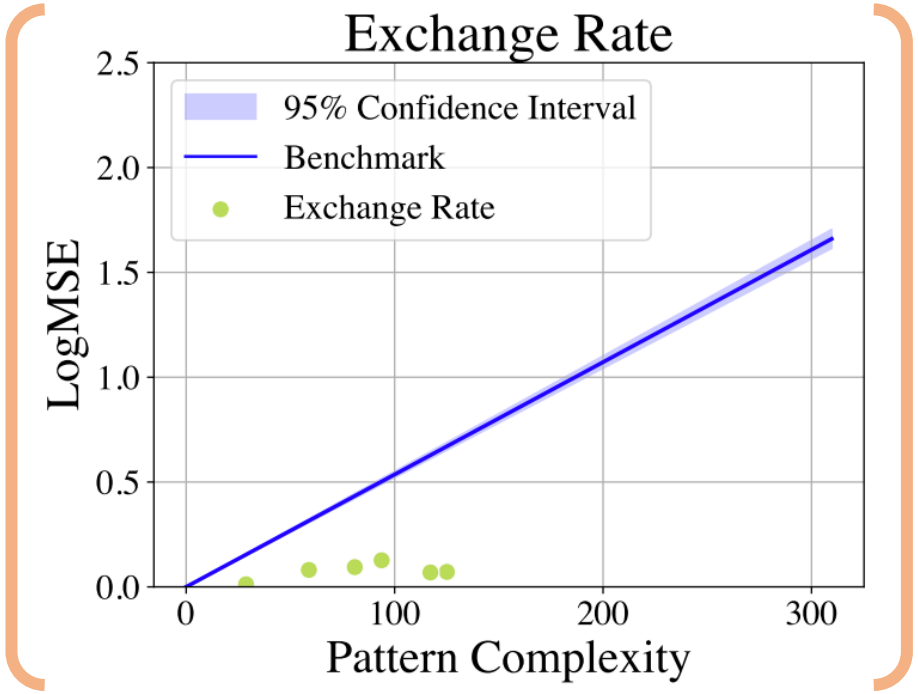
Saturated tasks (under the univariate forecasting setting)



Saturated task: Current MSE is smaller than the estimated value by accuracy law.

Saturated tasks (under the univariate forecasting setting)

Do not indicate the “easy task” but the “saturated task” (needs more than past observation)



Wang & Wu et al., Accuracy Law for the Future of Deep Time Series Forecasting. arXiv 2025

Lifted Framework in Dynamical Systems (Time Series)

$p(\mathbf{x})$ - - - - - - - - - - - - Vanilla Time Series Forecasting

$p(\mathbf{x}, \mathbf{x}_{\text{long-term}})$ - - - - - - Long-term Forecasting — Autoformer

$p(\mathbf{x}, \mathrm{ex})$ - - - - - - - - - - - - - - **Forecasting with Exogenous Variables** — TimeXer

$p(\mathbf{x}|\mathbf{z})$ - - - - - - - - - - - - - Large-scale Pre-training — Some Discussion

* Omit the shared conditional variables, such as past observations.

Part 2. Forecasting with Exogenous Variables

How to break the "accuracy law"? ***Go beyond univariate forecasting!***

[Supply]

Solar, Coal, Wind

[Electricity Price]

[Demand]

Resident, Commercial, Industry

Part 2. Forecasting with Exogenous Variables

NEURAL INFORMATION PROCESSING SYSTEMS

TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables

Yuxuan Wang[*], **Haixu Wu**[*], **Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, Mingsheng Long**[✉]

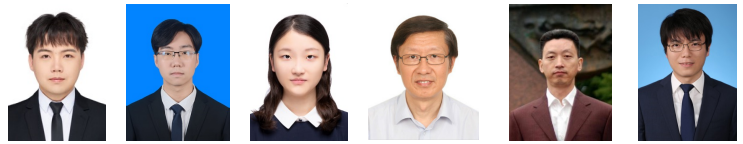
Yuxuan Wang Haixu Wu Jiaxiang Dong Guo Qin Haoran Zhang Yong Liu Yunzhong Qiu Jianmin Wang Mingsheng Long

[TimeXer, NeurIPS 2024, 500+ Citations]

SCIENCE CHINA Information Sciences

METADATA MATTERS FOR TIME SERIES: INFORMATIVE FORECASTING WITH TRANSFORMERS

Jiaxiang Dong[*], **Haixu Wu**[*], **Yuxuan Wang**[*], **Li Zhang, Jianmin Wang, Mingsheng Long**[✉]
School of Software, BNRist, Tsinghua University, Beijing 100084, China
{djx20,wuhx23,wangyuxu22}@mails.tsinghua.edu.cn
{lizhang,jimwang,mingsheng}@tsinghua.edu.cn

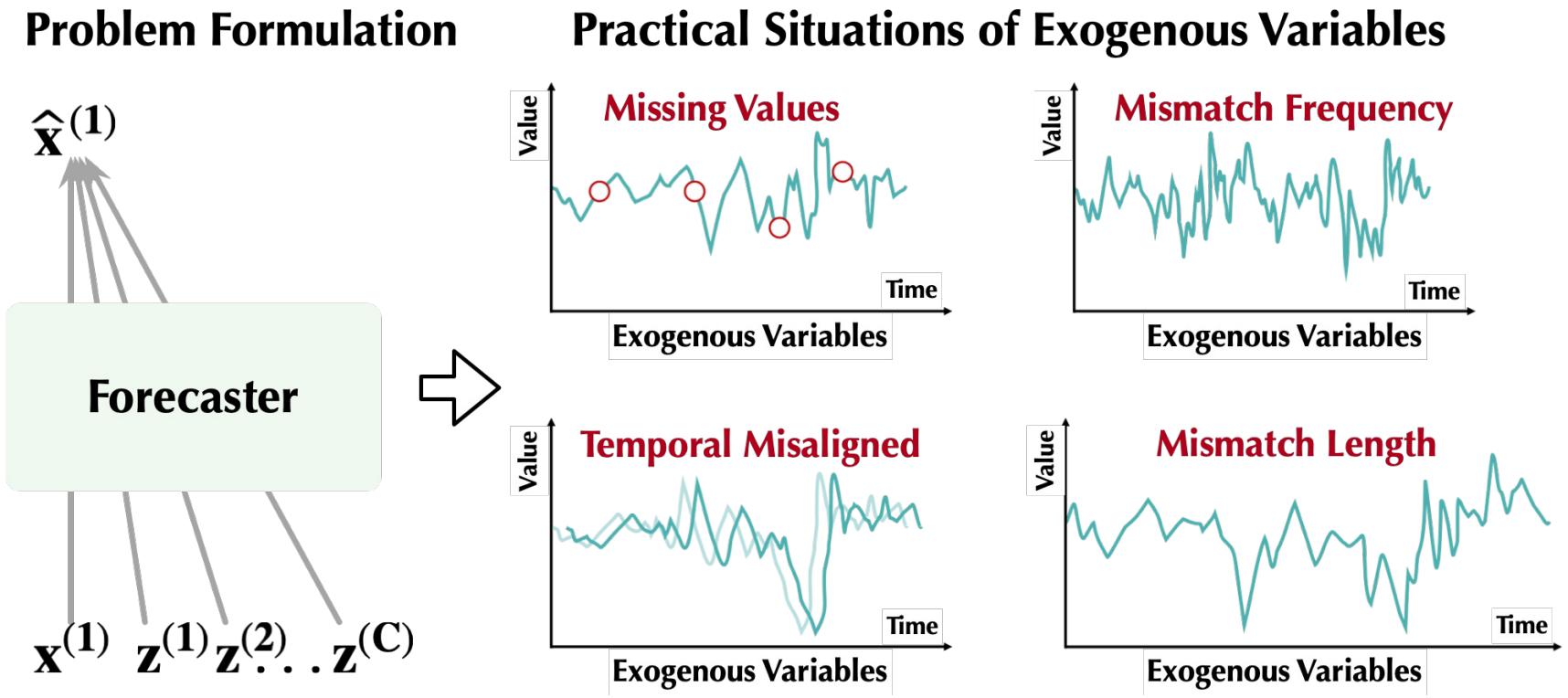


Jiaxiang Dong Haixu Wu Yuxuan Wang Li Zhang Jianmin Wang Mingsheng Long

[MetaTST, Science China Information Sciences 2025]

Problem definition

The exogenous variables are introduced to the forecaster for informative purposes without the need for forecasting.

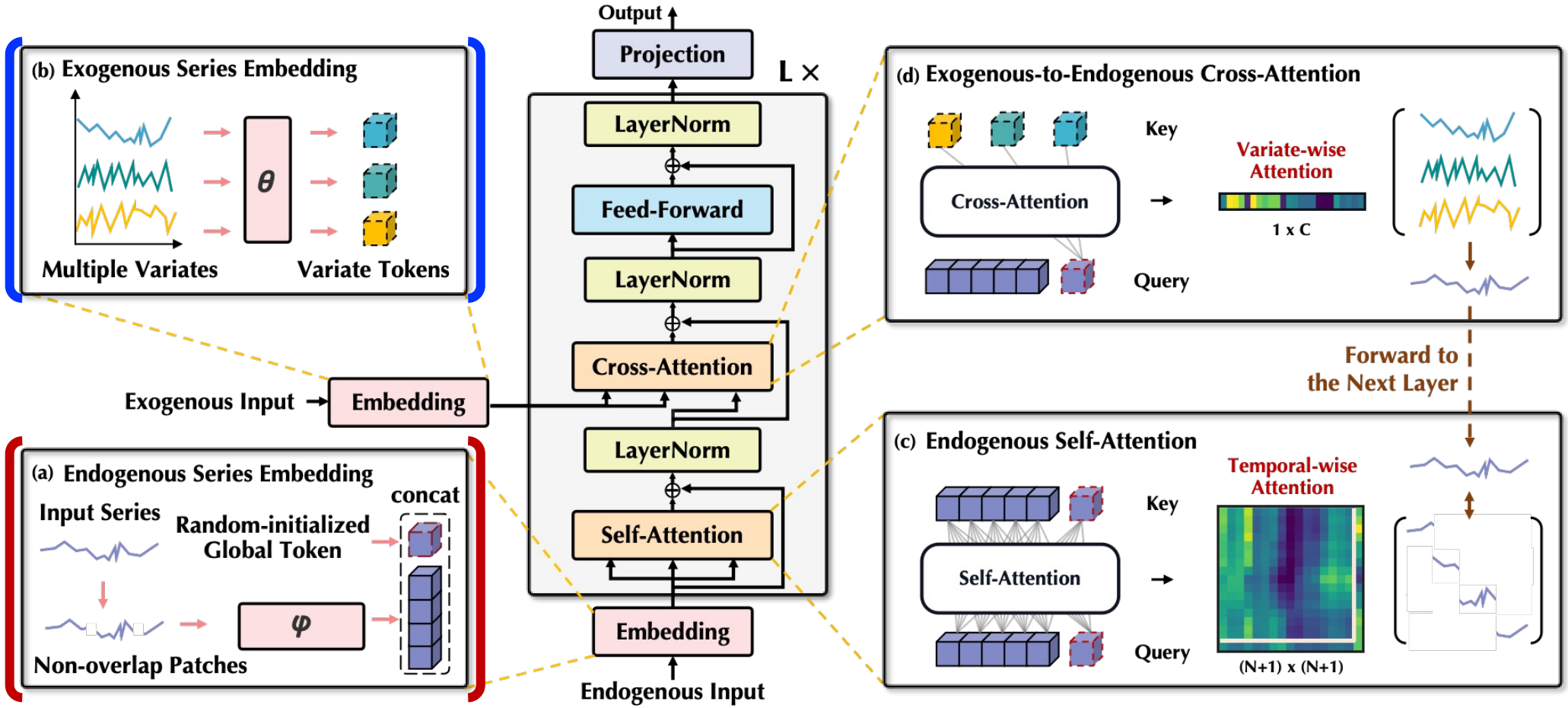


Unique challenge distinct from previous models

- Most of the existing Transformer-based approaches treat all the variables equally or ignore exogenous information, **lacking a special design of exogenous series**.
- Previous forecasters who design models for exogenous variables overlook the **complex nature of these variables**.

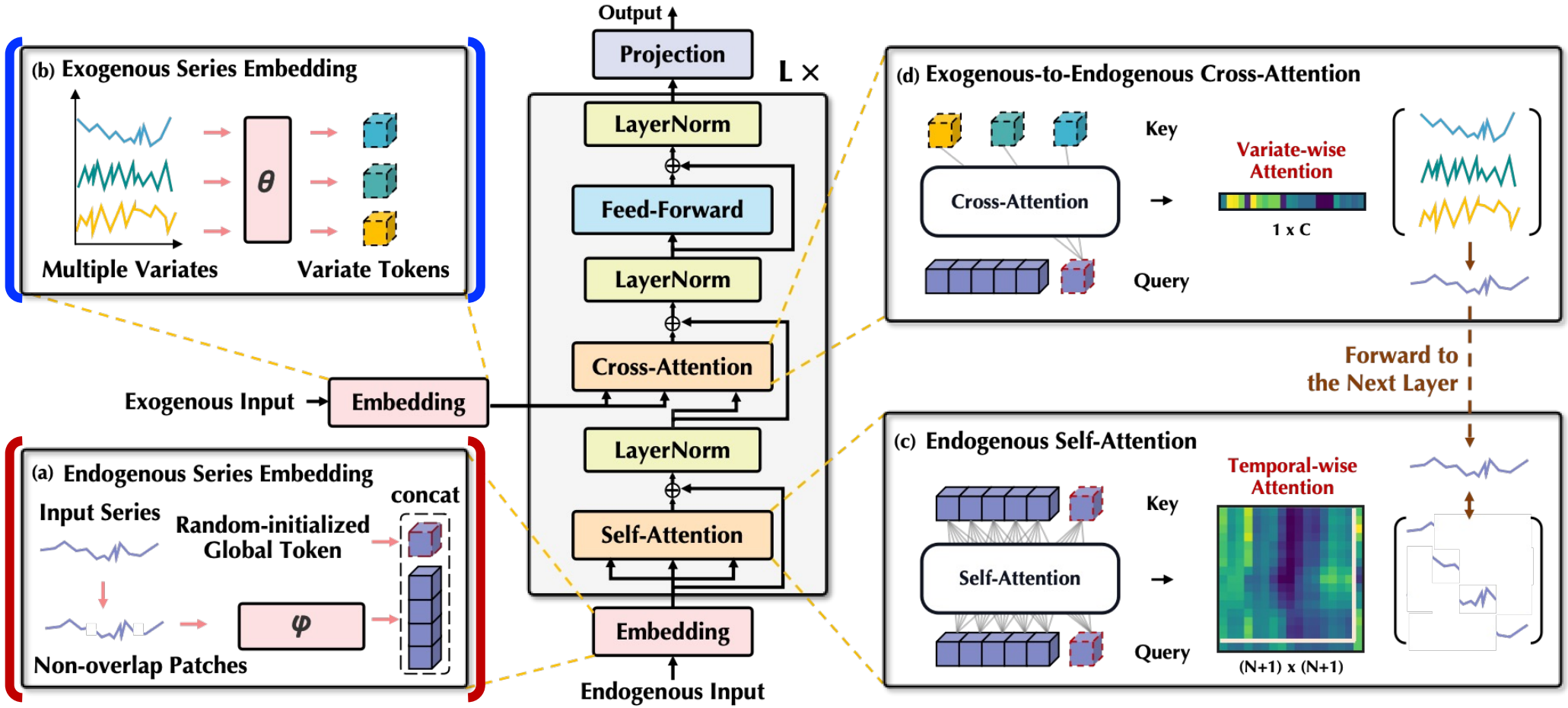
Methods	**TimeXer**	iTran. [23]	PatchTST [28]	Cross. [43]	Auto. [37]	TFT [16]	NBEATSx [29]	TiDE [5]
Univariate	✓	✗	✓	✗	✓	✗	✗	✗
Multivariate	✓	✓	✧	✓	✧	✗	✗	✓
Exogenous	✓	✗	✗	✗	✗	✓	✓	✓

Overall architecture



Overall architecture

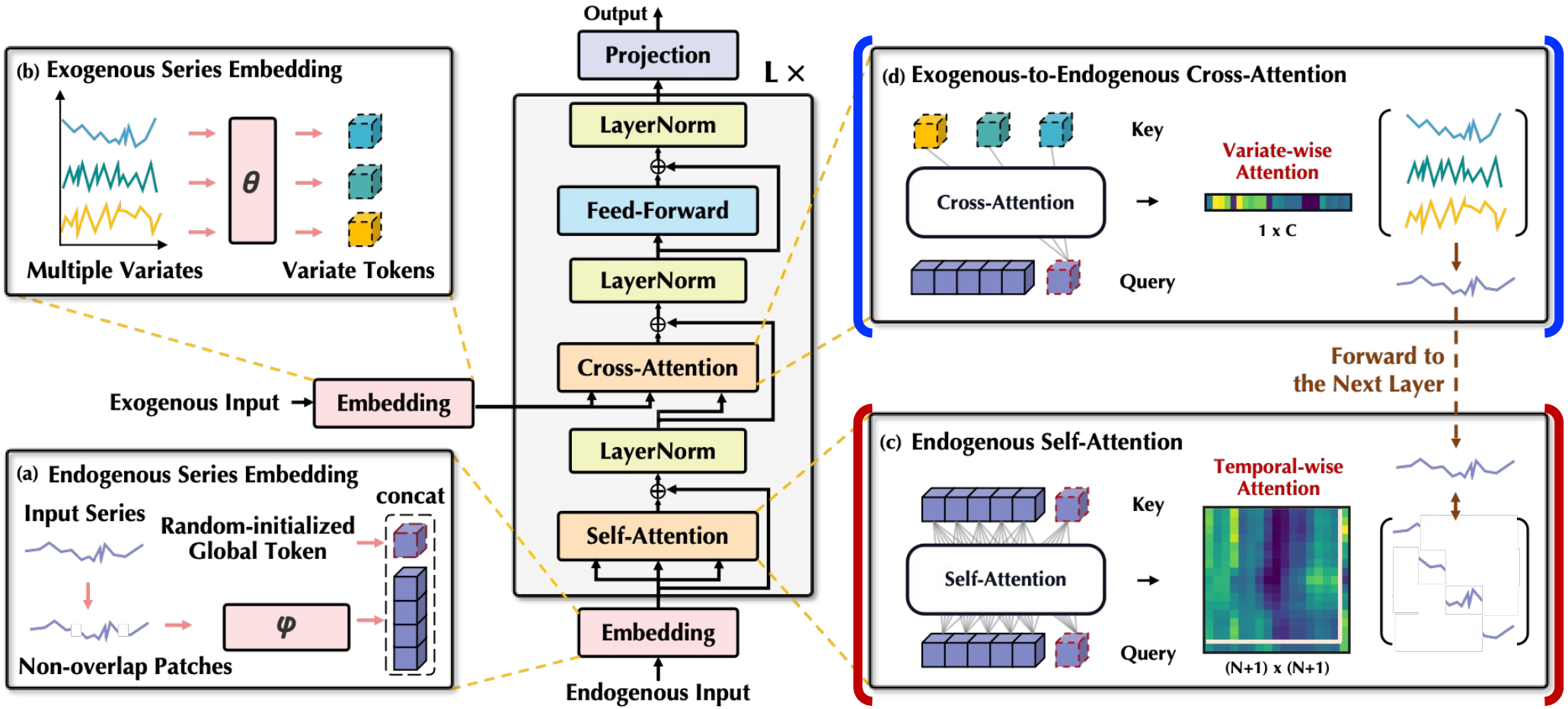
✓ **Exogenous** time series are embedded in **variate Tokens**.



✓ **Endogenous** time series are embedded at the patch level.

Overall architecture

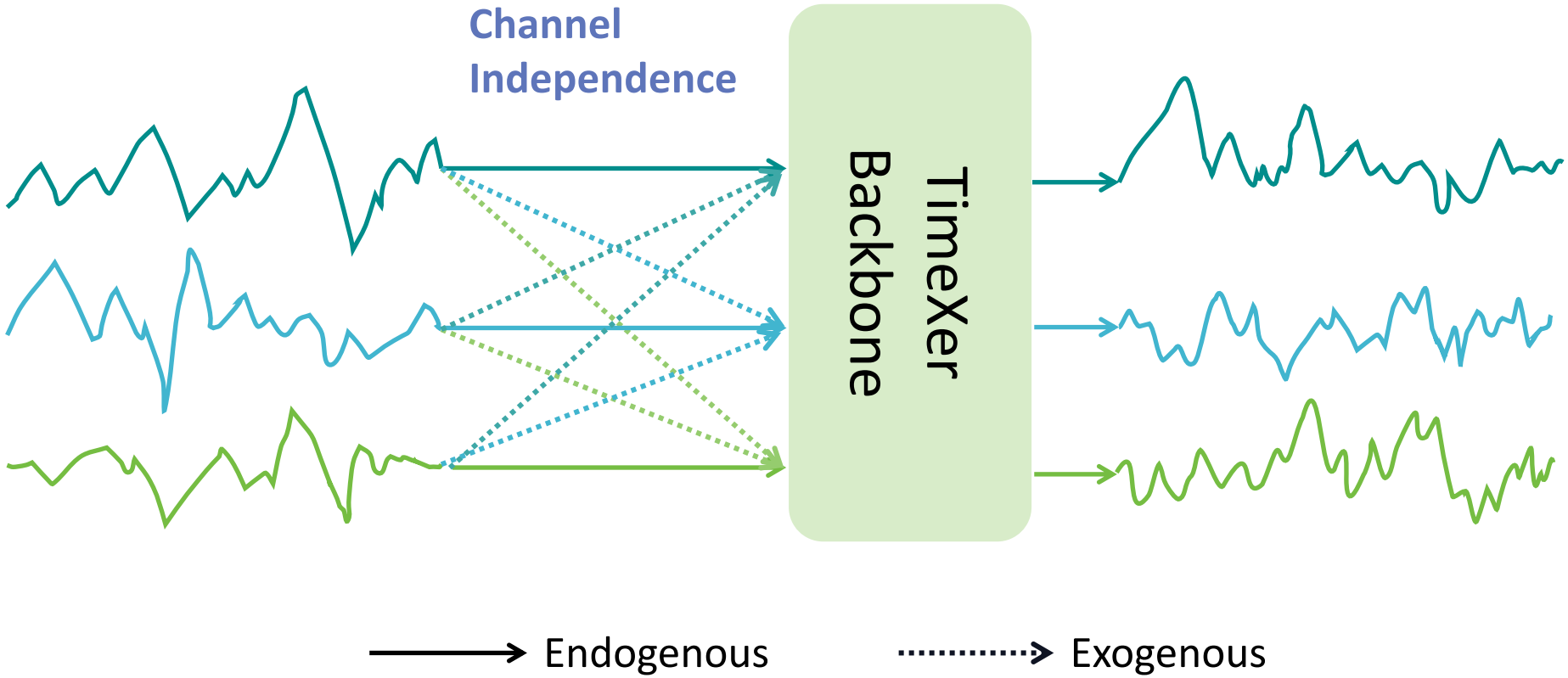
✓ **Cross-Attention** to incorporate exogenous information



✓ **Self-Attention** to capture temporal correlations among endogenous information.

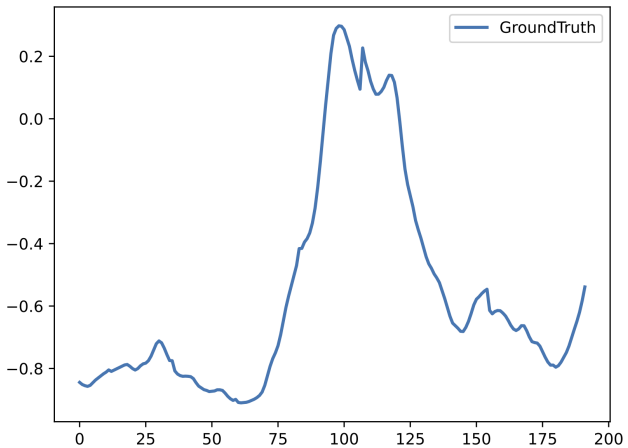
Parallel computing for multivariate forecasting

- **Extending to multivariate forecasting:** Each variable is treated as the endogenous one, with the others being exogenous, and uses **a shared TimeXer backbone**.

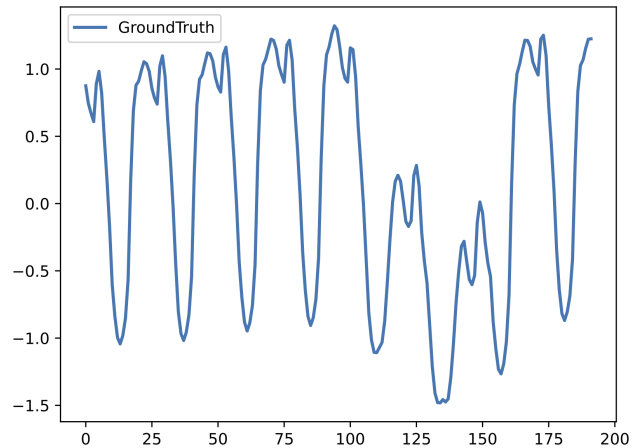


Some showcases

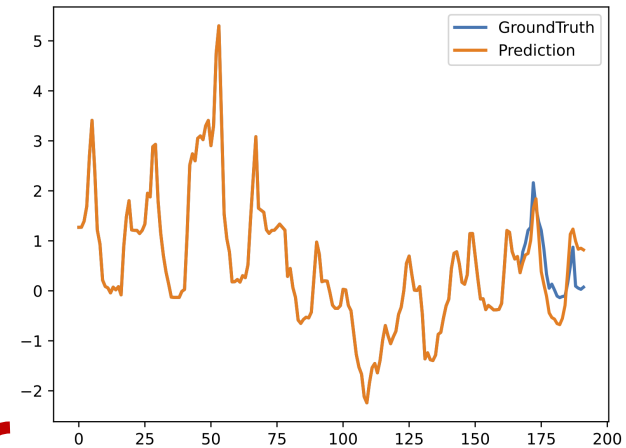
(a) Wind power (Exogenous)



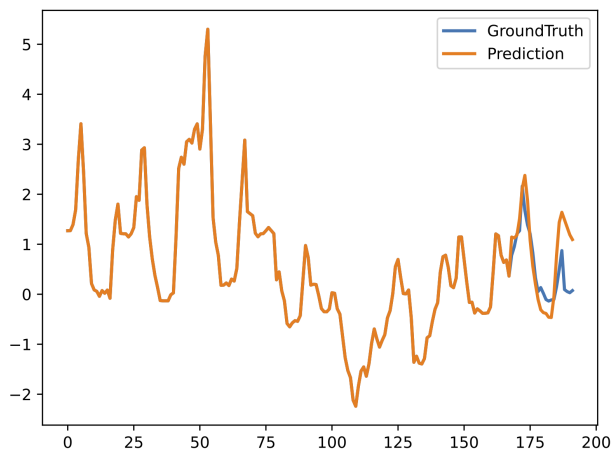
(b) Ampirion zonal load (Exogenous)



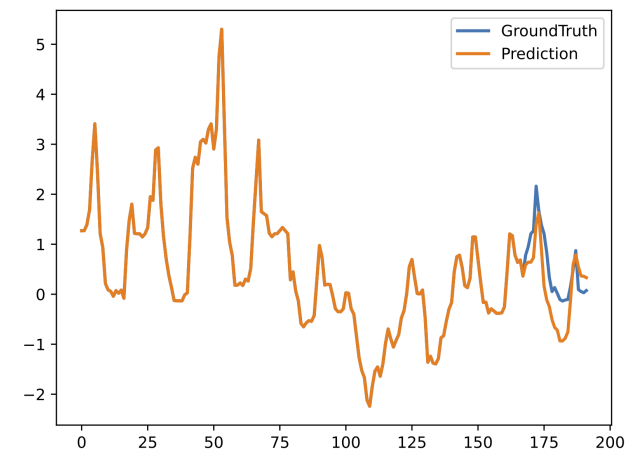
(c) TimeXer (PastEx, PastEn)



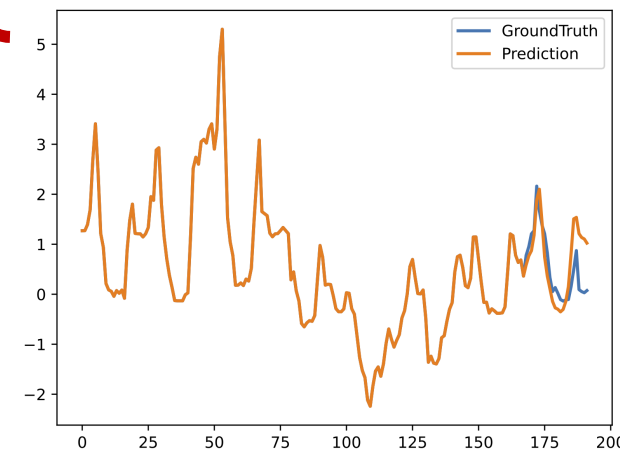
(d) TimeXer (PastEx, None)



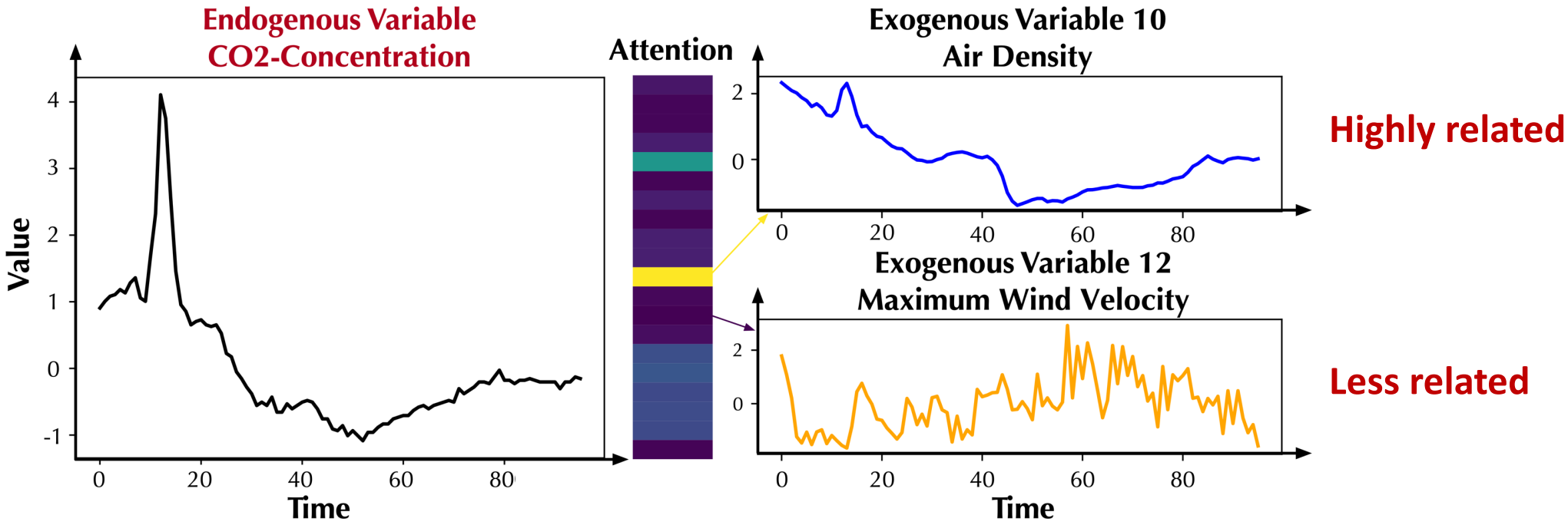
(e) TimeXer (None, PastEn)



(f) TimeXer (Past+FutureEx, PastEN)



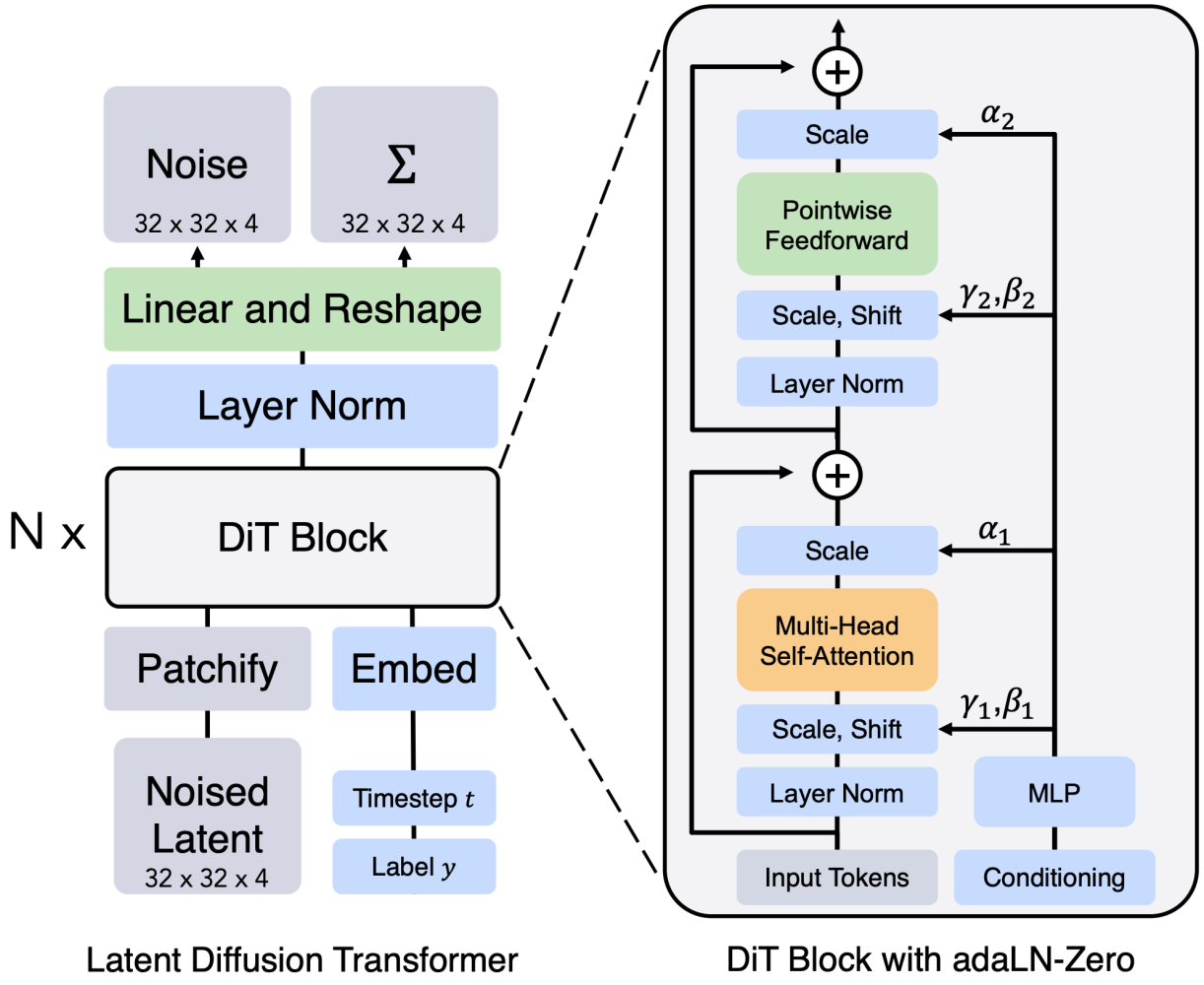
Attention Map analysis



TimeXer can distinguish different exogenous variables, resulting in a more focused and interpretable attention map. (can be used to identify factors)

* An attention map can reflect the correlation among different variables

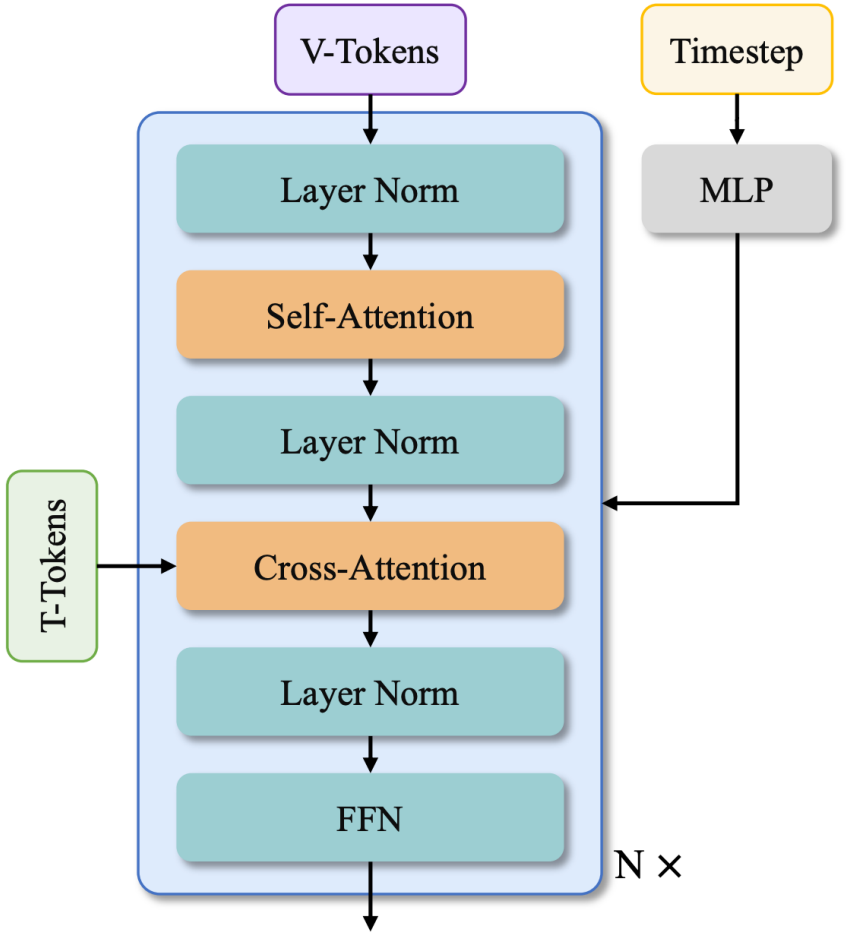
The modern way to do so…



Latent Diffusion Transformer

DiT Block with adaLN-Zero

DiT [Peebles et al., ICCV 2023]



Wan 2.1 [Wan Team, Alibaba Group, arXiv 2025]

Lifted Framework in Dynamical Systems (Time Series)

$p(\mathbf{x})$ ---------- Vanilla Time Series Forecasting

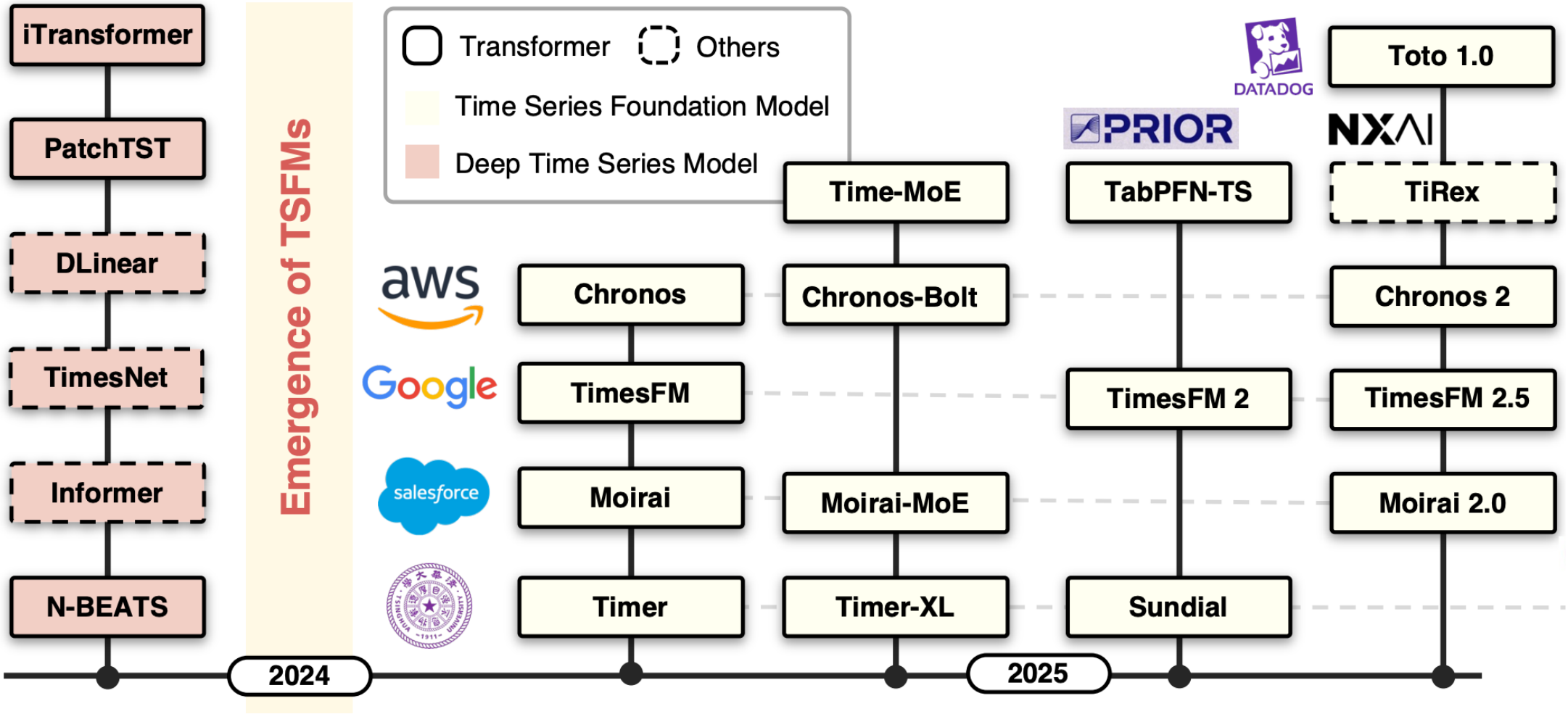
$p(\mathbf{x}, \mathbf{x}_{\text{long-term}})$ ------ Long-term Forecasting — Autoformer

$p(\mathbf{x}, \text{ex})$ ---------- Forecasting with Exogenous Variables — TimeXer

$p(\mathbf{x}|\mathbf{z})$ ---------- **Large-scale Pre-training** — Some Discussion

* Omit the shared conditional variables, such as past observations.

Part 3. Time Series Foundation Models



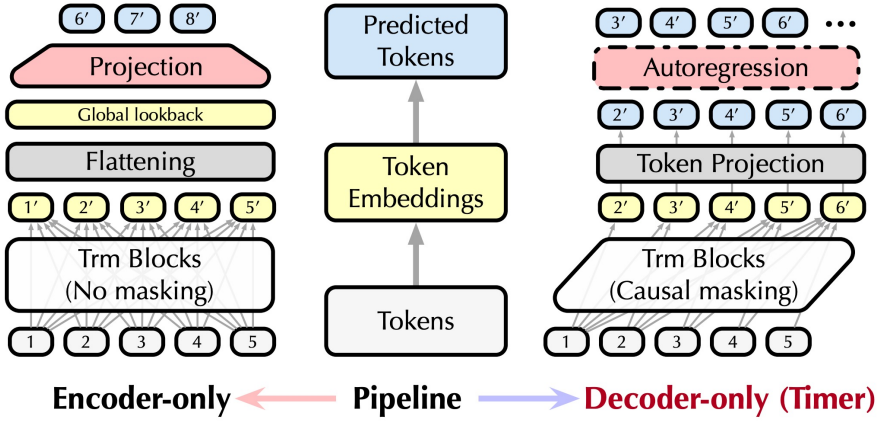
Liu et al., Timer-S1: A Billion-Scale Time Series Foundation Model with Serial Scaling. arXiv 2026

A large design space

Method	**Sundial (Ours)**	Time-MoE (2024b)	Timer-XL (2024a)	Moirai (2024)	MOMENT (2024)	LLMTime (2024)	Chronos (2024)	Lag-Llama (2023)	TimesFM (2023b)
Architecture	Decoder	Decoder	Decoder	Encoder	Encoder	Decoder	EncDec	Decoder	Decoder
Model Size	32M 128M 444M	113M 453M 2.4B	84M	14M 91M 311M	40M 125M 385M	-	46M 200M 710M	200M	17M 70M 200M
Pre-training Scale	1032B	300B	260B	231B	1.13B	-	84B	0.36B	100B
Token Level	Patch	Point	Patch	Patch	Patch	Point	Point	Point	Patch
Tokenization	Continuous	Continuous	Continuous	Continuous	Continuous	Discrete	Discrete	Continuous	Continuous
Context Length	$\leq$2880	$\leq$4096	$\leq$2880	$\leq$5000	= 512	-	$\leq$512	$\leq$1024	$\leq$512
Probabilistic	True	False	False	True	False	True	True	True	False

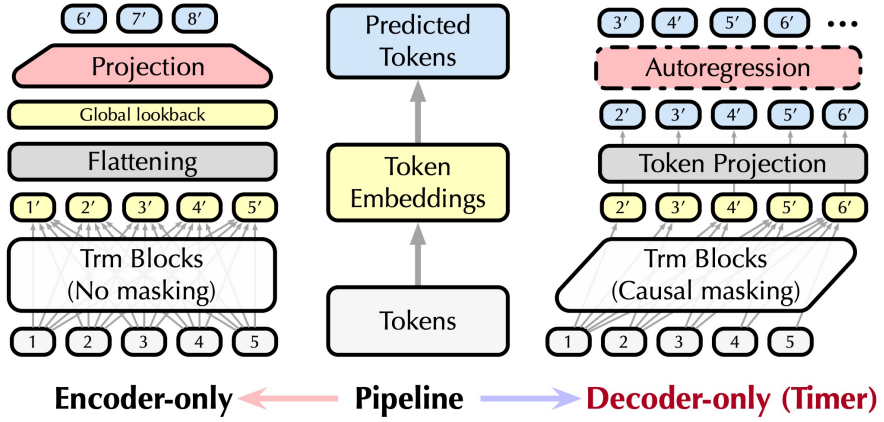
Liu et al., Sundial: A Family of Highly Capable Time Series Foundation Models. ICML 2025

Some early explorations

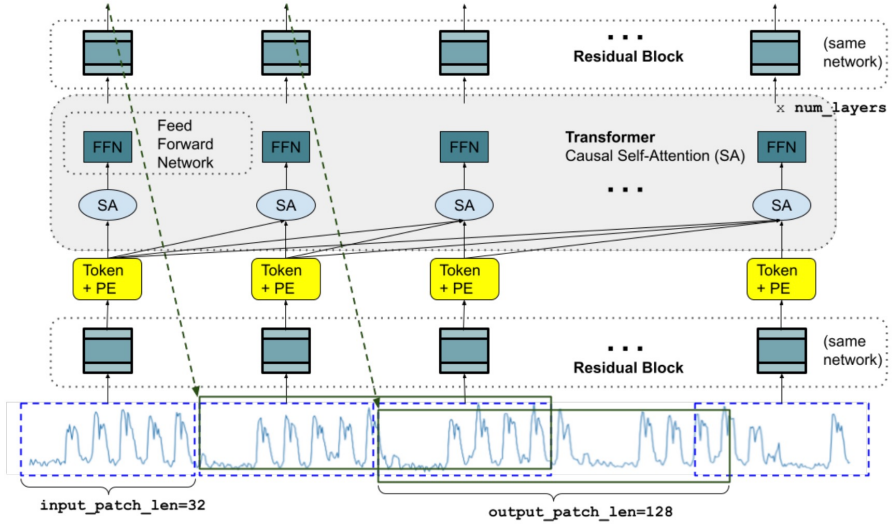


[Timer, ICML 2024] Tsinghua University

Some early explorations

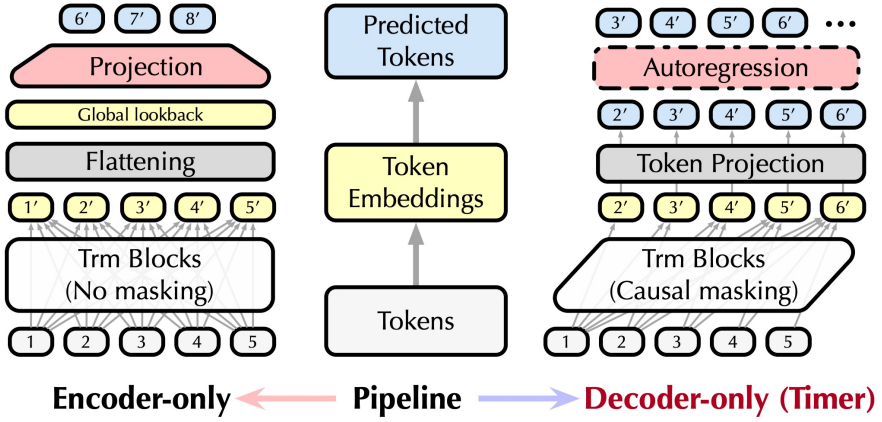


[Timer, ICML 2024] Tsinghua University

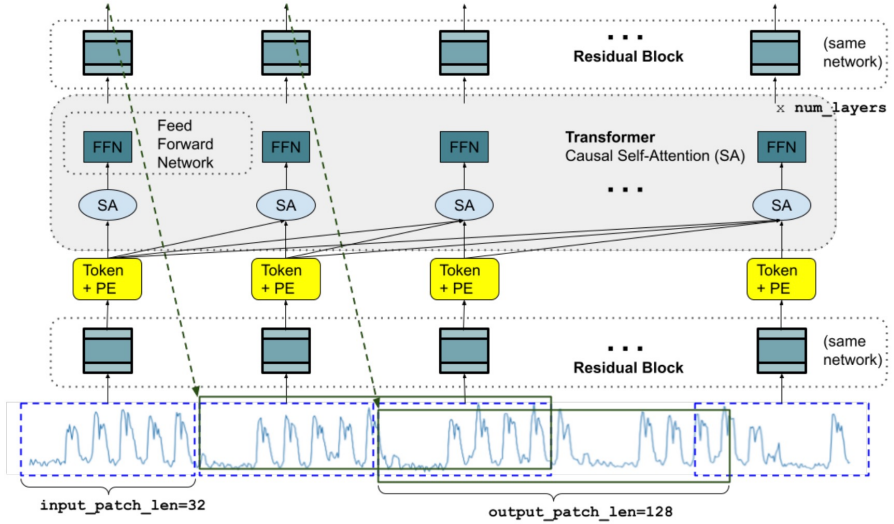


[TimesFM, ICML 2024] Google

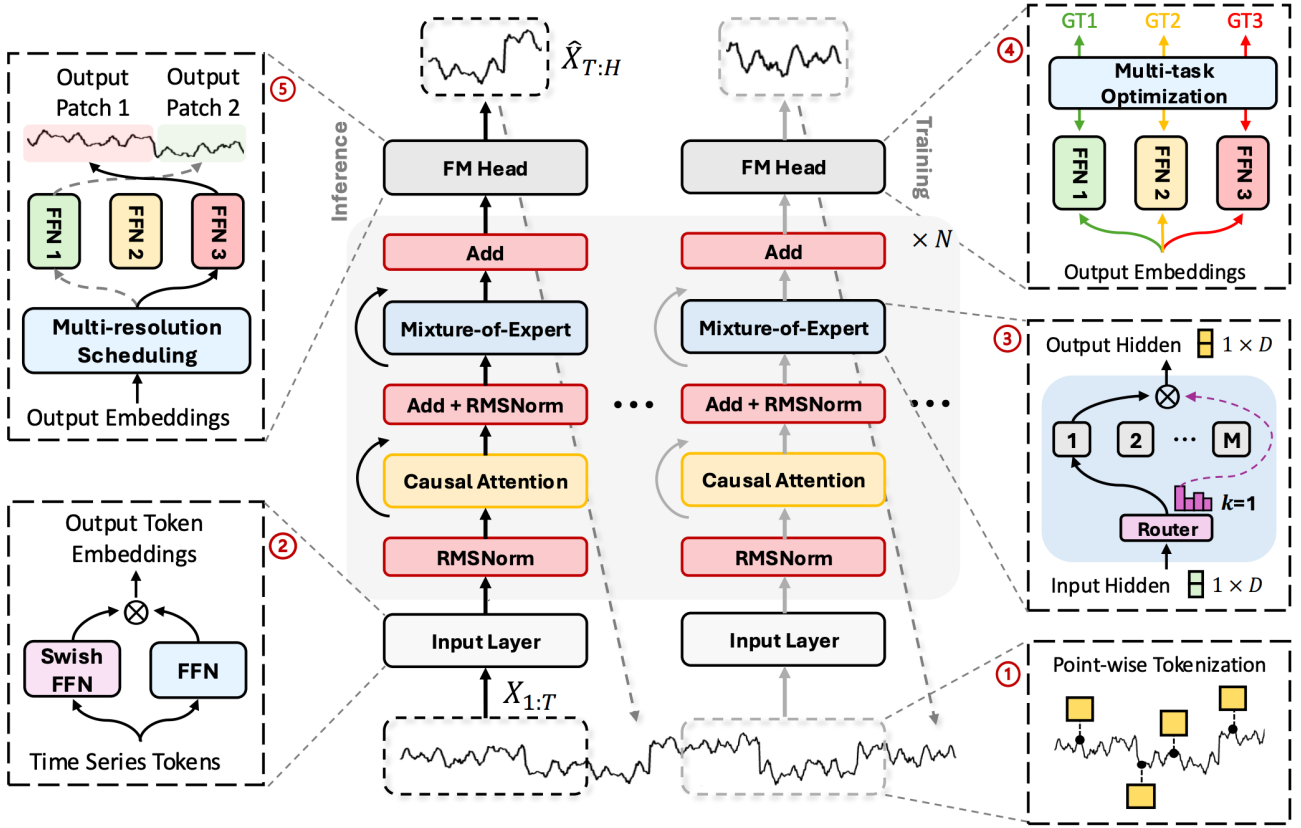
Some early explorations



[Timer, ICML 2024] Tsinghua University



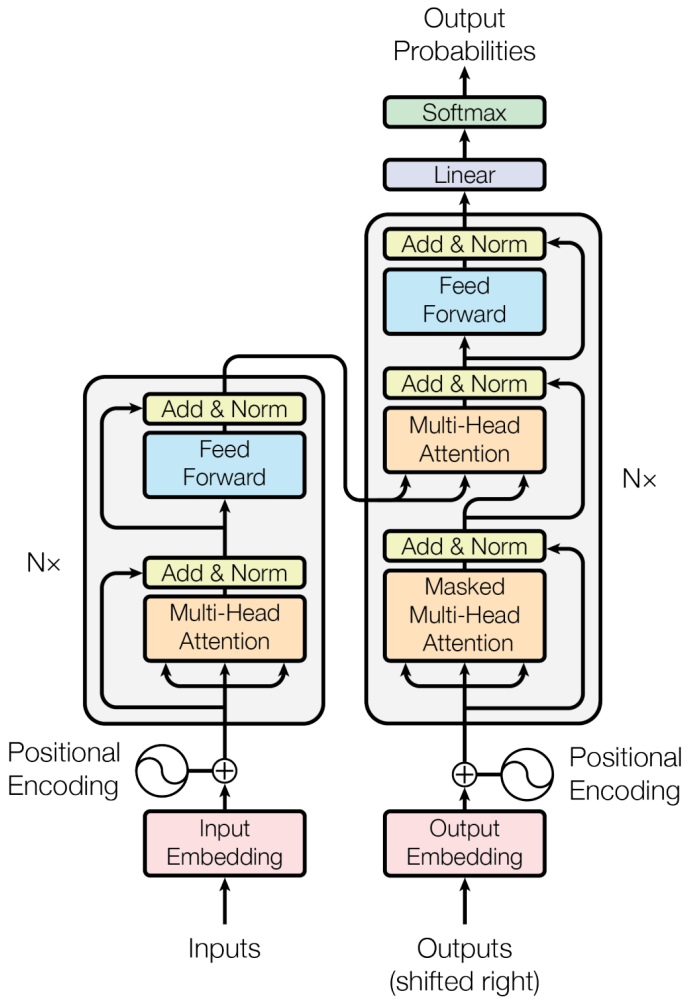
[TimesFM, ICML 2024] Google



[Time-MOE, ICLR 2025] Bytedance

2.4B model trained from 300B time points

Recap Intelligence: A close look at LLMs



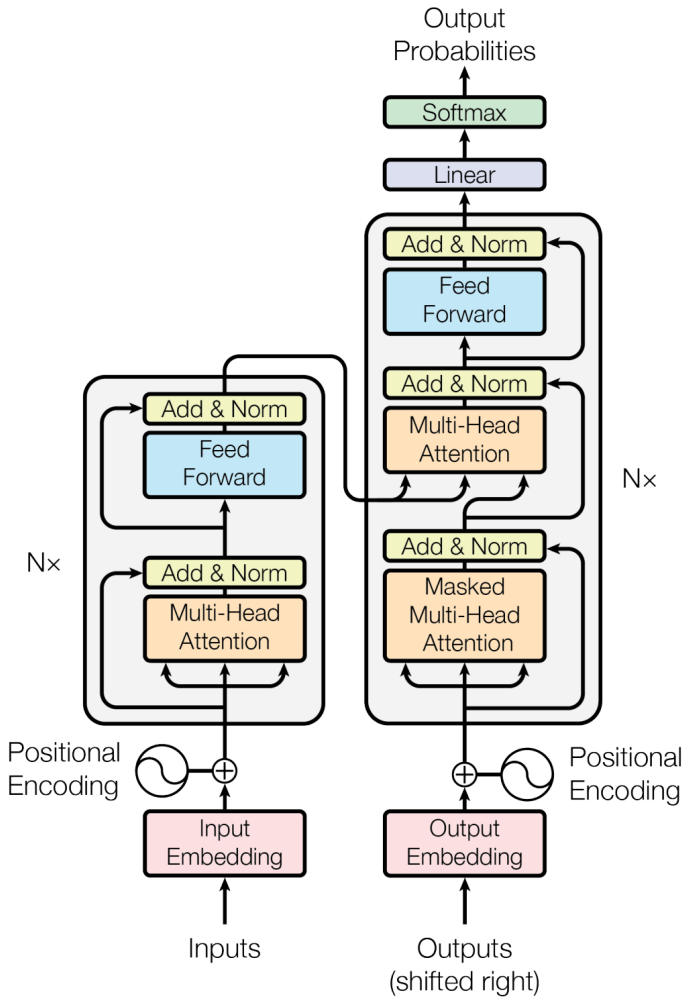
Why does a Transformer activate “intelligence”?

1. Architectural aspect

- Encoder
- Decoder

Vaswani et al., Attention Is All You Need. NeurIPS 2017

A close look at LLMs



Why does a Transformer activate “intelligence”?

1. Architectural aspect

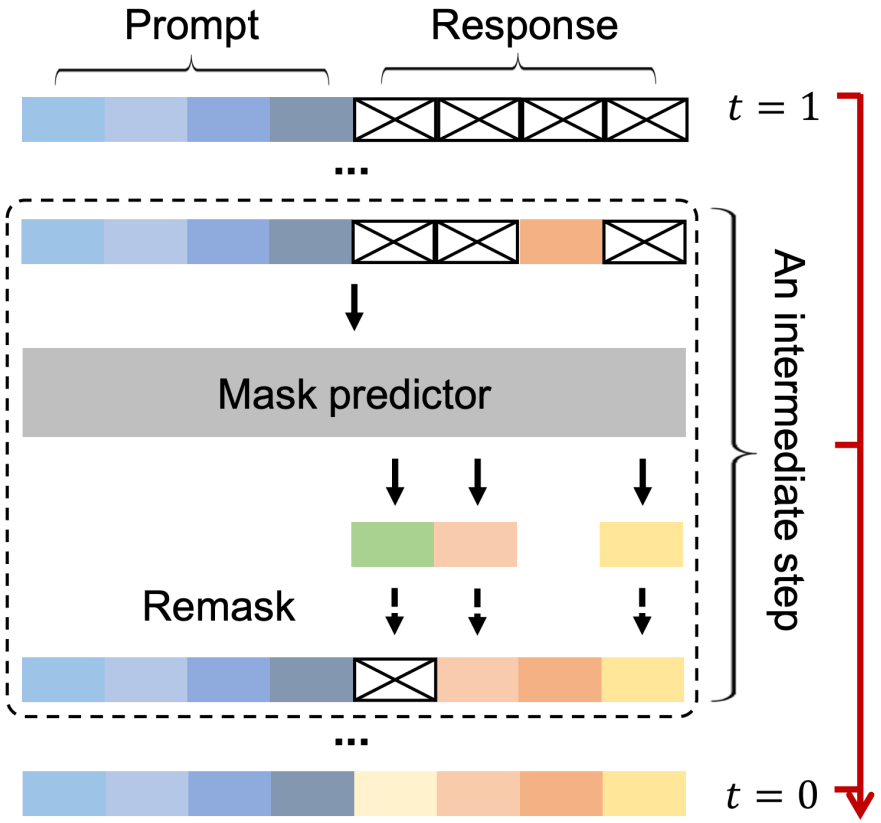
- Encoder
- Decoder

The decoder-only architecture.

The next-token prediction task.

Zhu et al., Physics of Language Models. https://physics.allen-zhu.com/

A close look at LLMs: Recent evidence



Why does a Transformer activate “intelligence”?

An encoder-only diffusion model can also be a powerful LLM.

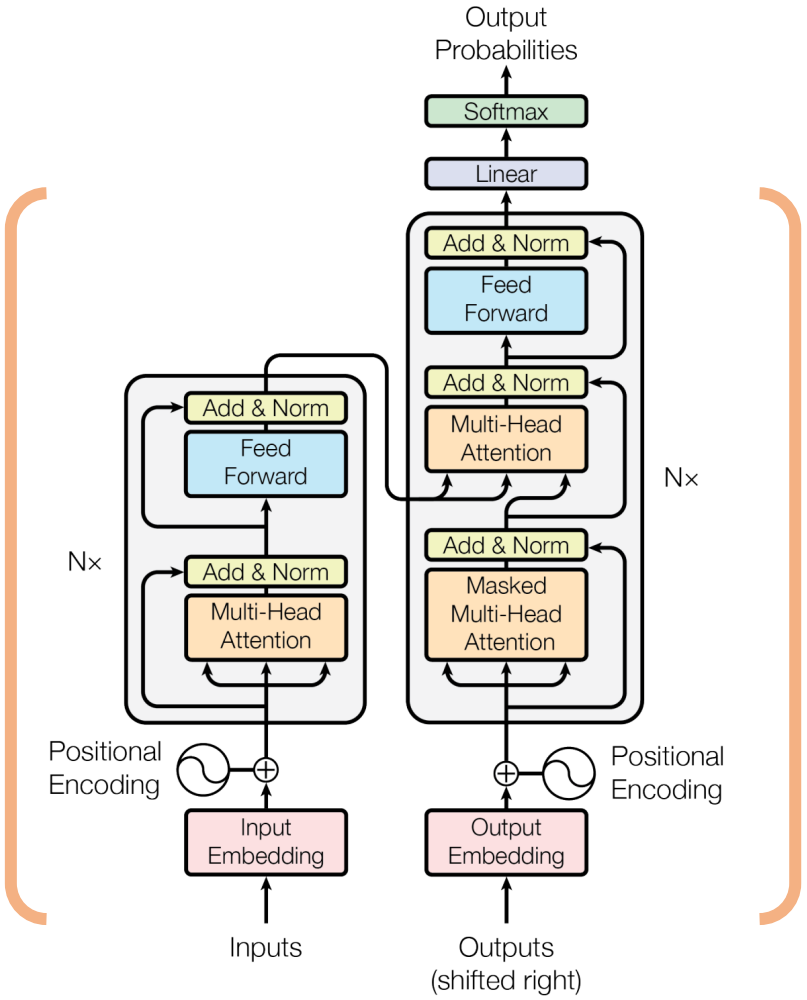
~~**The decoder-only architecture.**~~

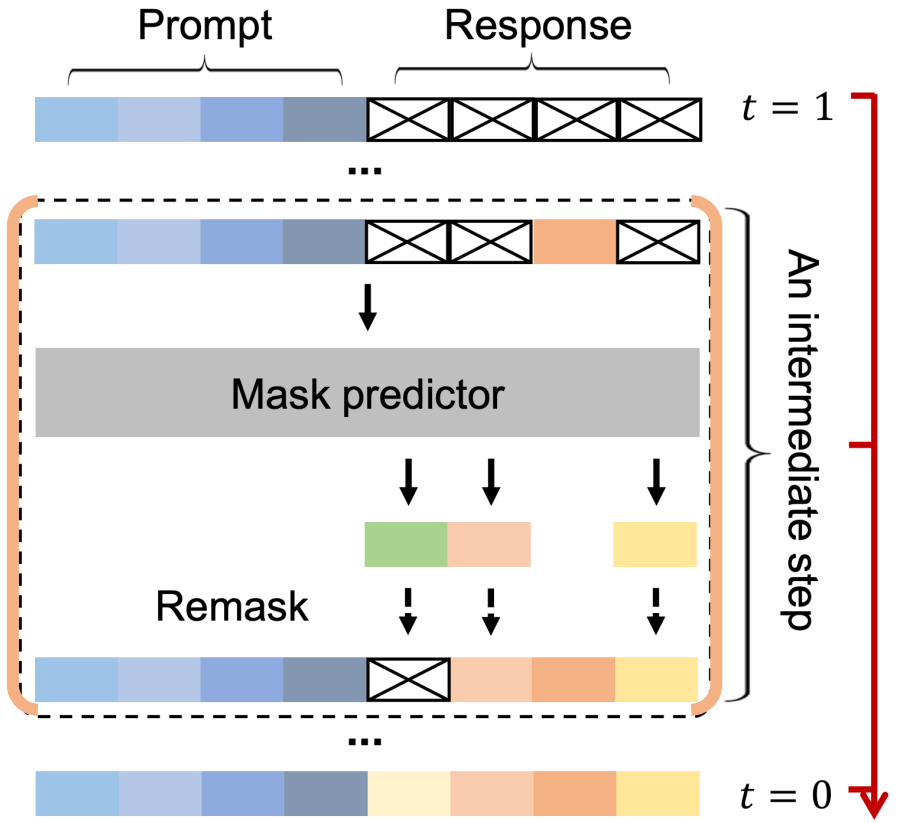
The next-token prediction task.

(correct but not principal)

Nie et al., Large Language Diffusion Models. NeurIPS 2025

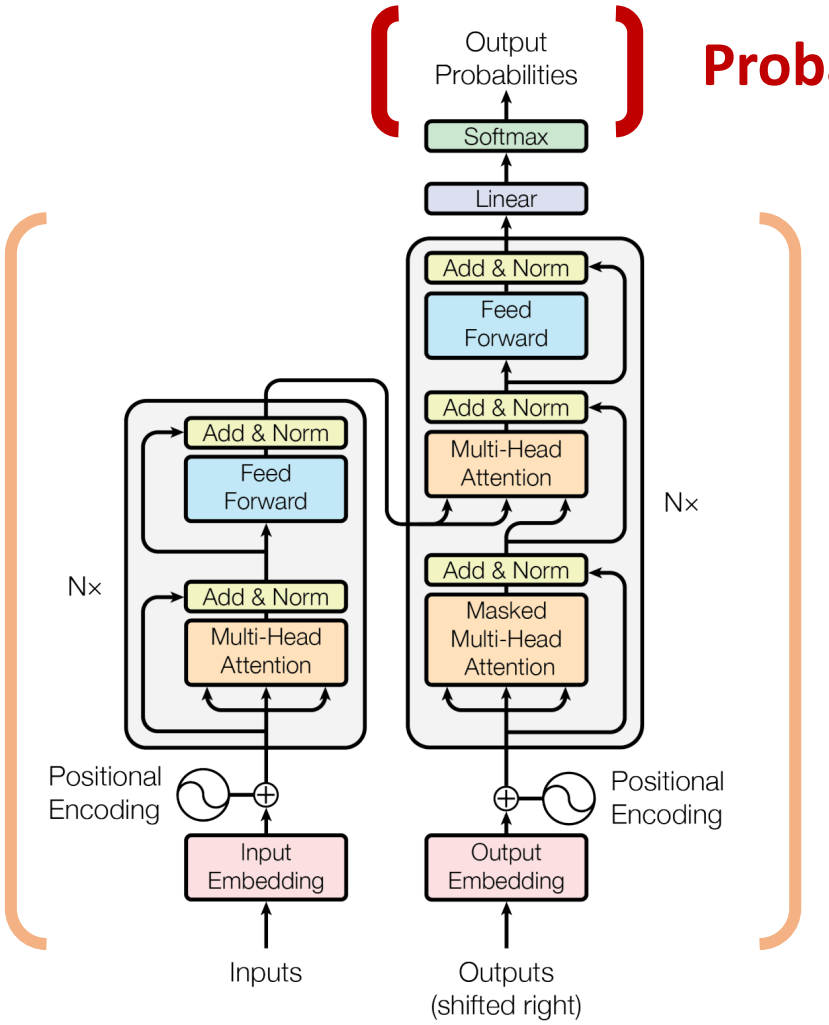
A close look at LLMs: Beyond architecture



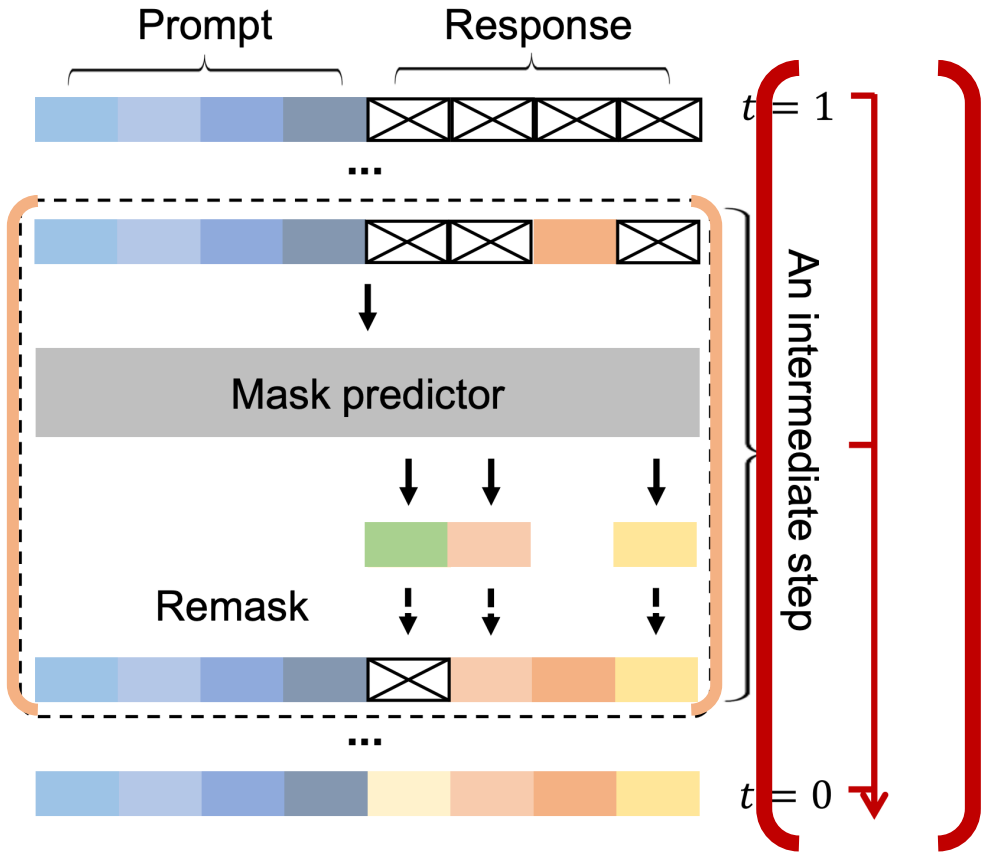


Encoding part: Project data into a unified representation space

A close look at LLMs: Beyond architecture

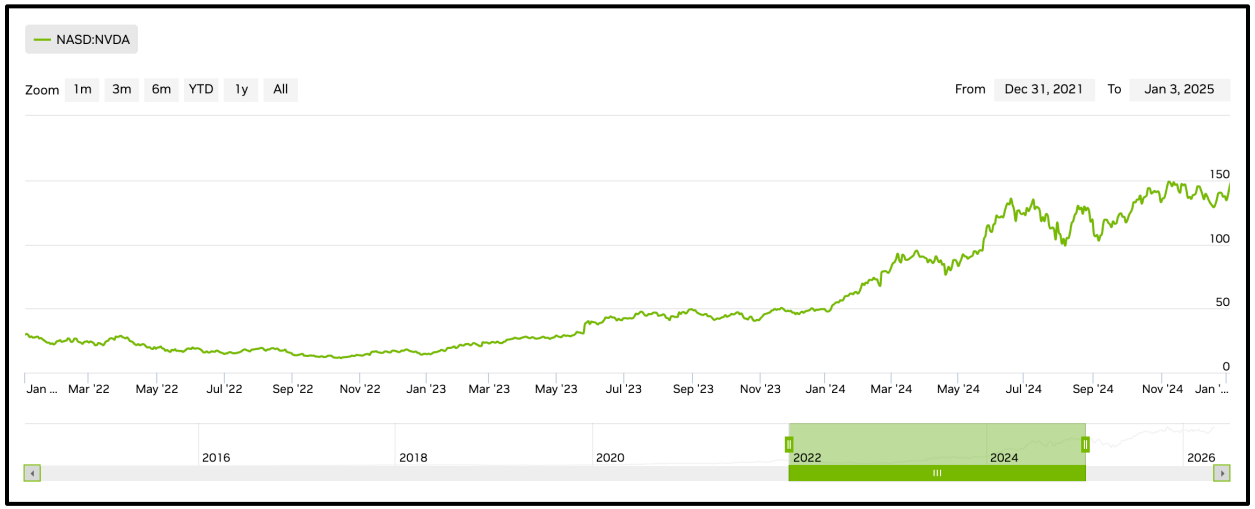


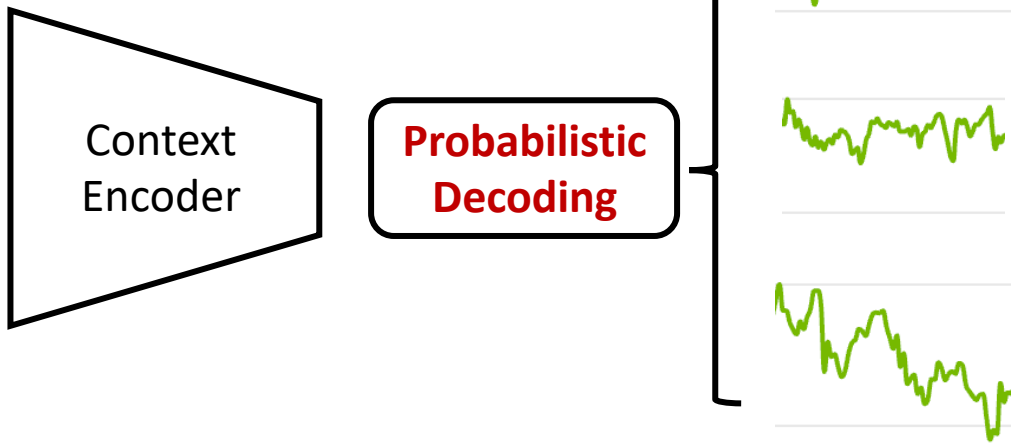
Probabilistic Decoding



Encoding part: Project data into a unified representation space

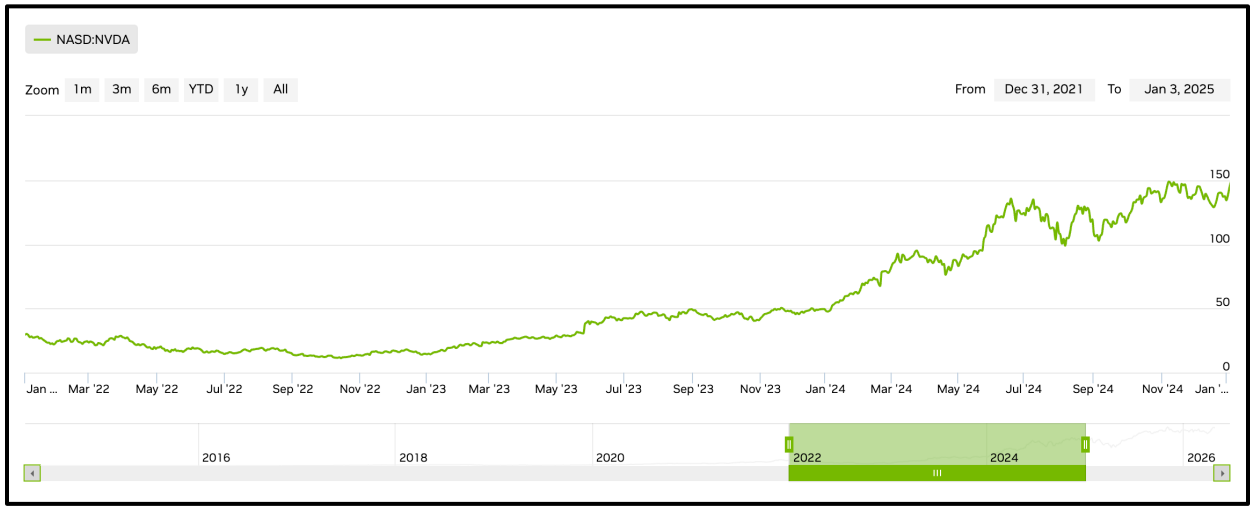
Intelligence: Encoding + Probabilistic decoding

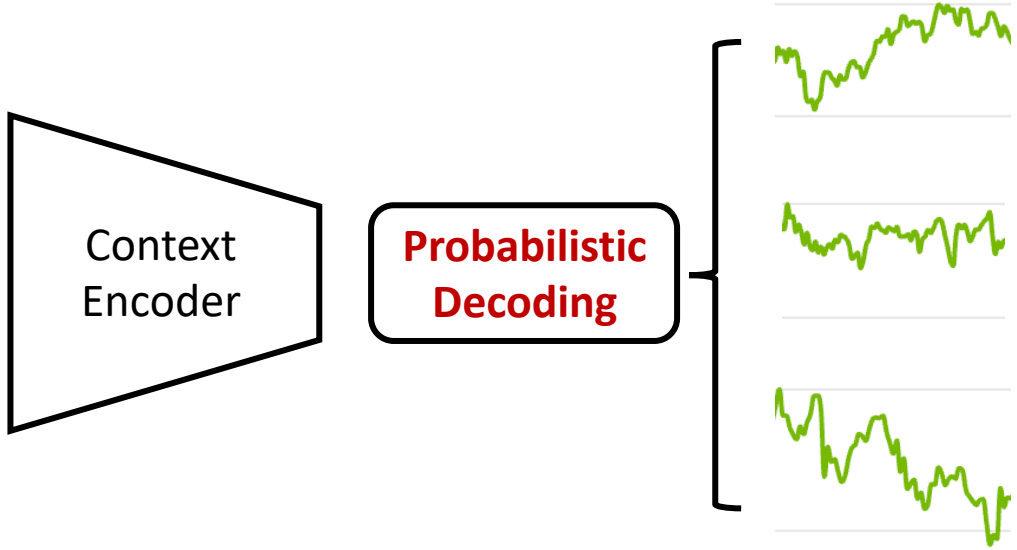


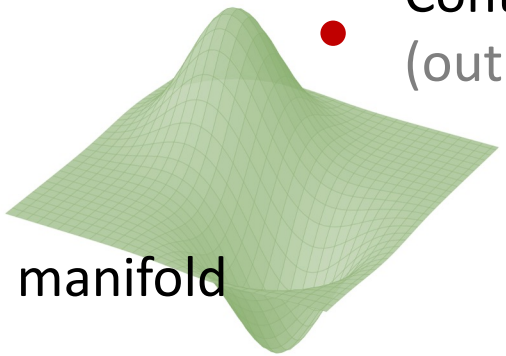


https://investor.nvidia.com/stock-info/stock-quote-and-chart/default.aspx

Intelligence: Encoding + Probabilistic decoding





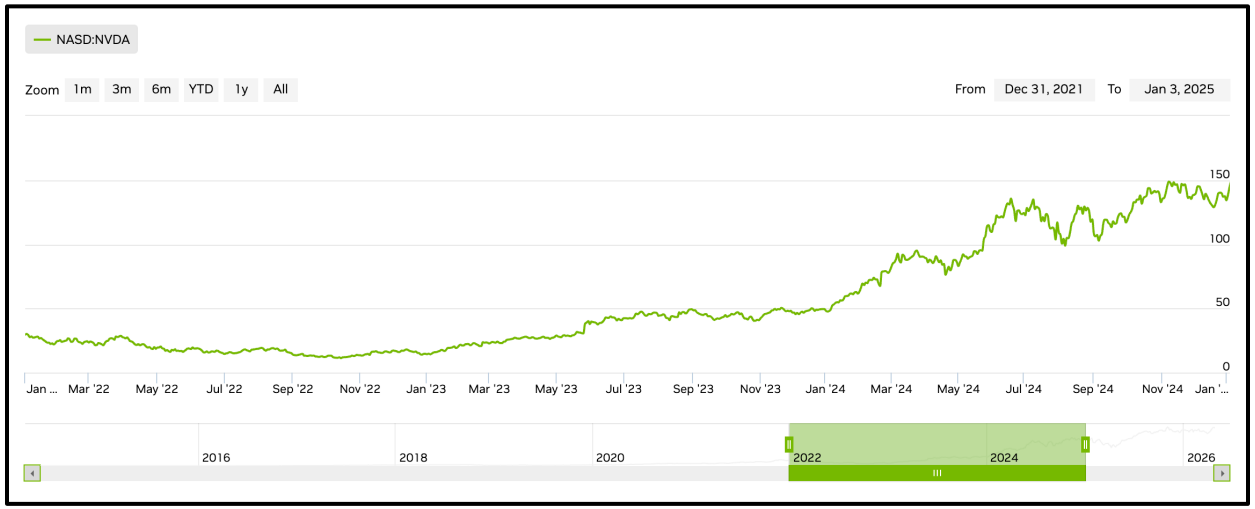


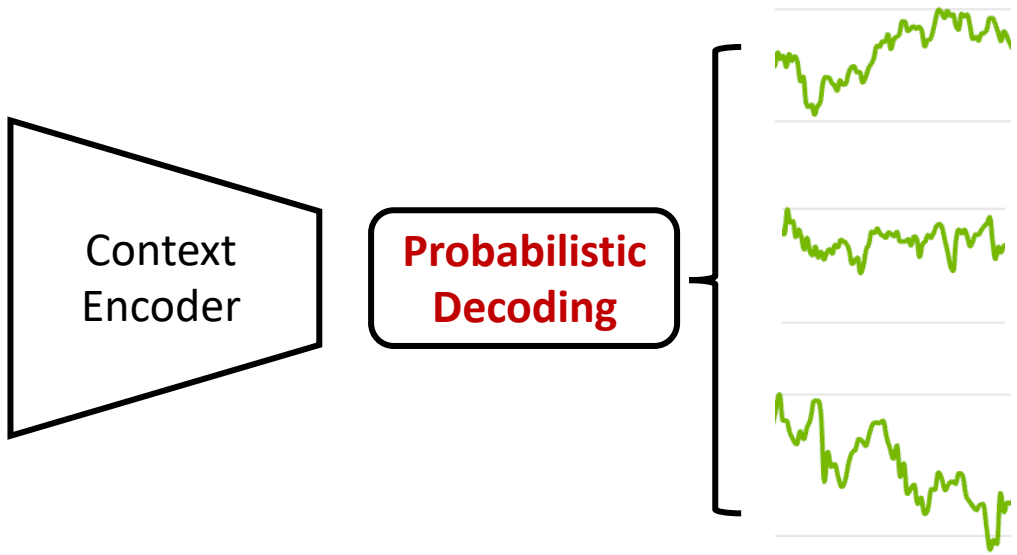
Context feature
(out of the manifold due to imperfect information)

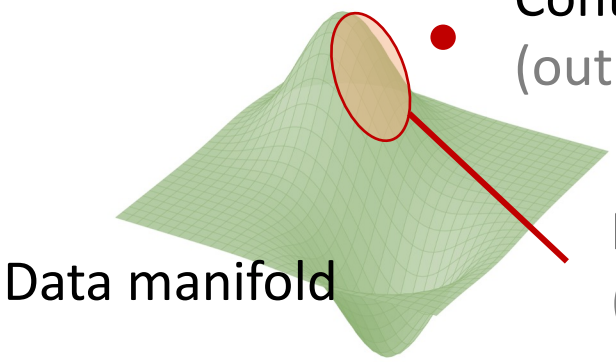
Data manifold

https://investor.nvidia.com/stock-info/stock-quote-and-chart/default.aspx

Intelligence: Encoding + Probabilistic decoding







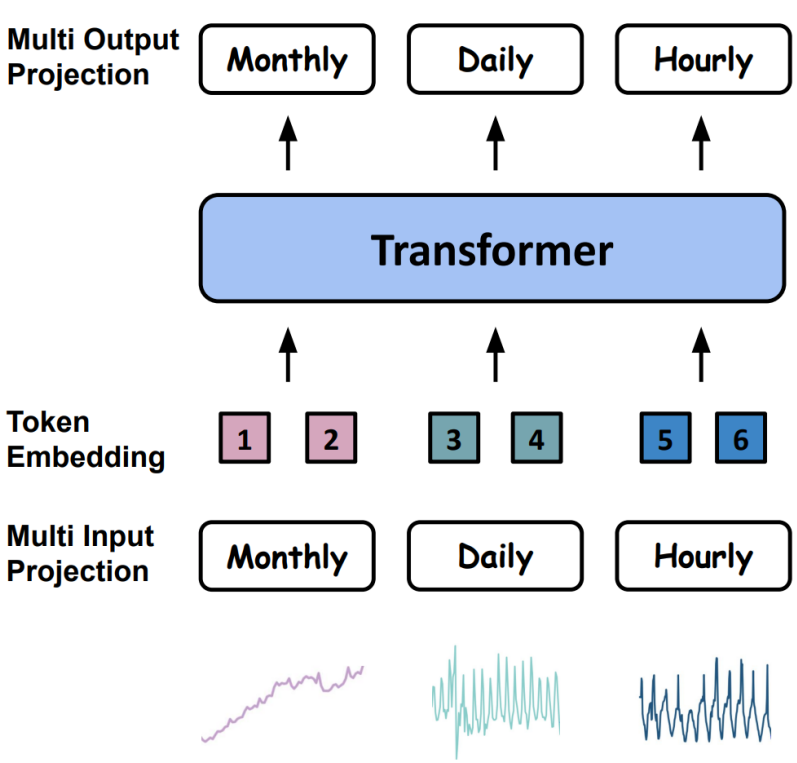
Context feature
(out of the manifold due to imperfect information)

Probabilistic Decoding: adopt **sampling** to project back to the manifold
(sample a word in LLM or denoise an image)

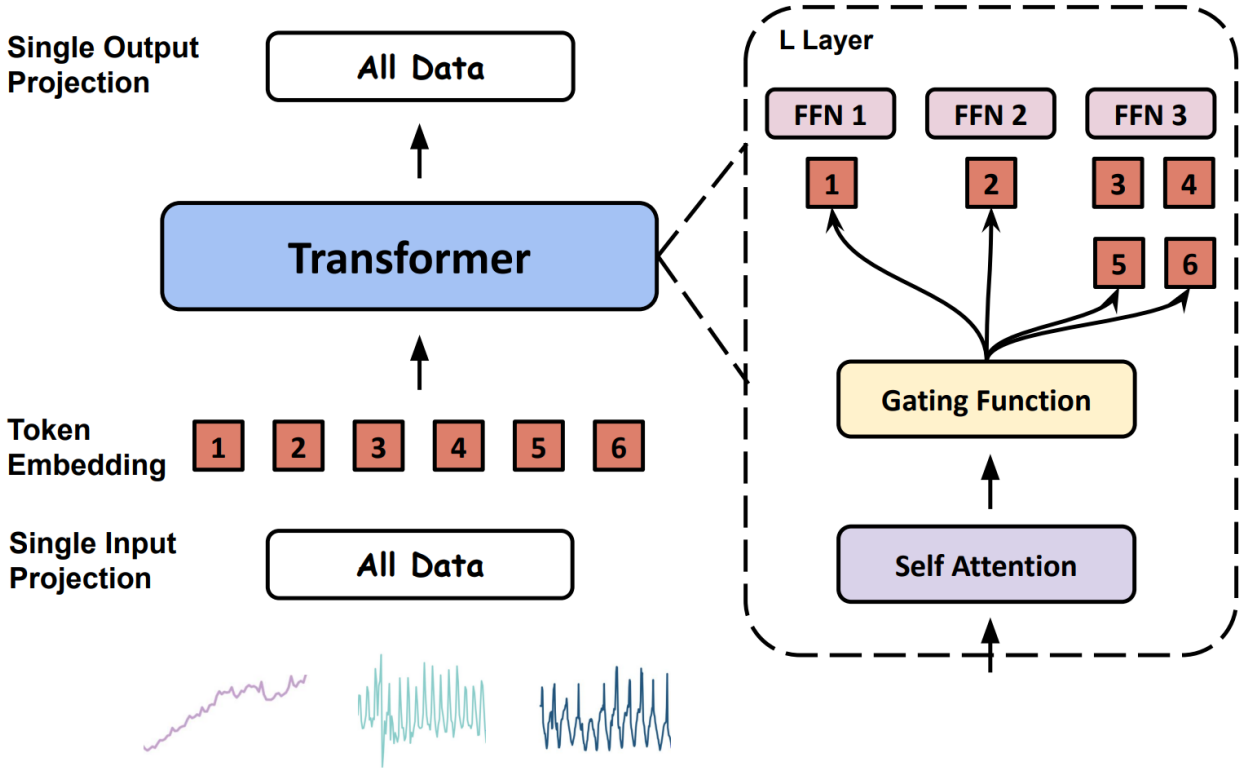
https://investor.nvidia.com/stock-info/stock-quote-and-chart/default.aspx

Mainstream players 1 - Moirai / Moirai-MOE

Mixed distribution parameters $\mathcal{L}_{\text{pred}} = -\log p(\boldsymbol{x}_{t+1}|\hat{\phi}),\ \hat{\phi} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-l+1:t})$

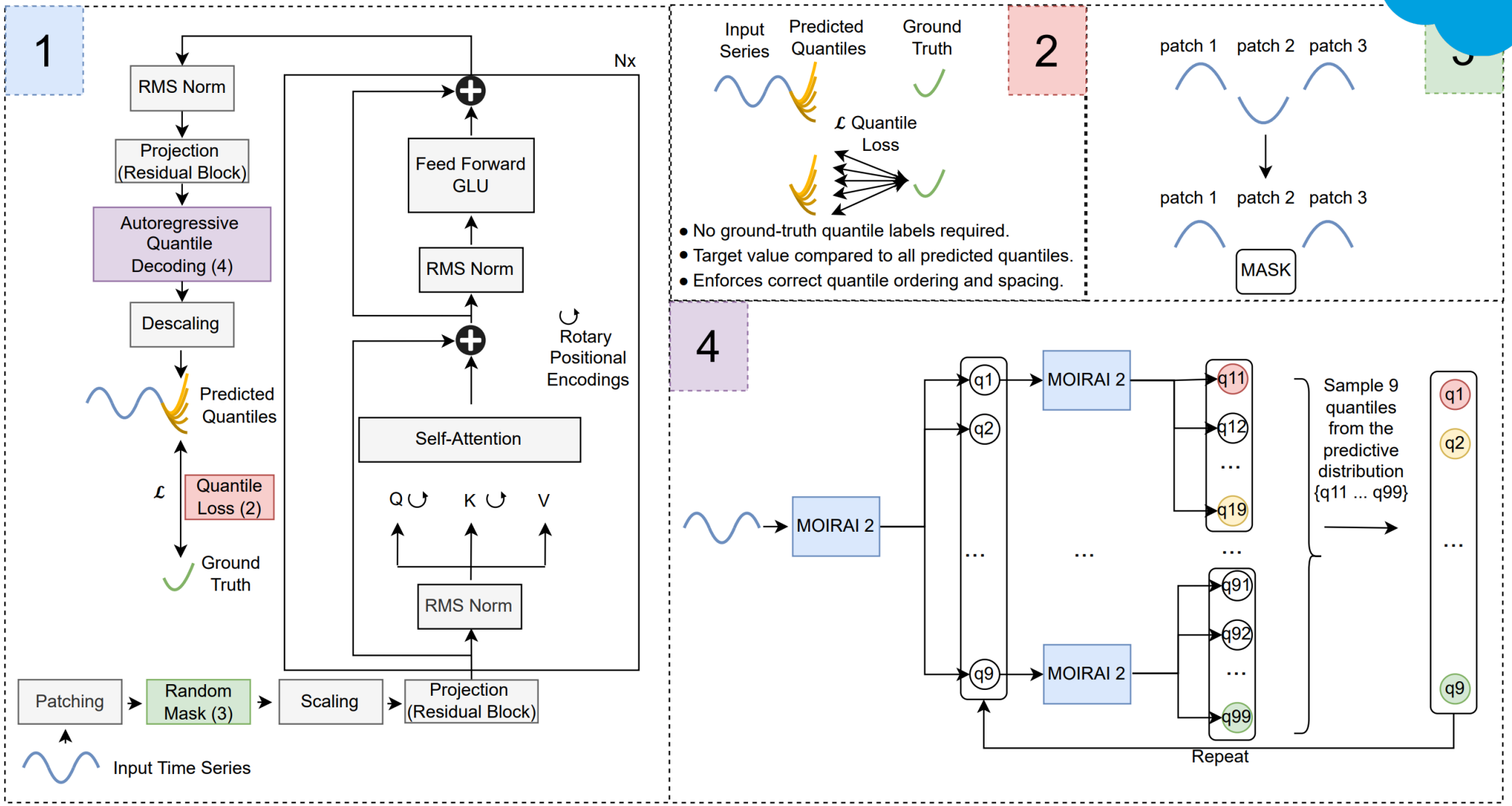


[Moirai, ICML 2024]



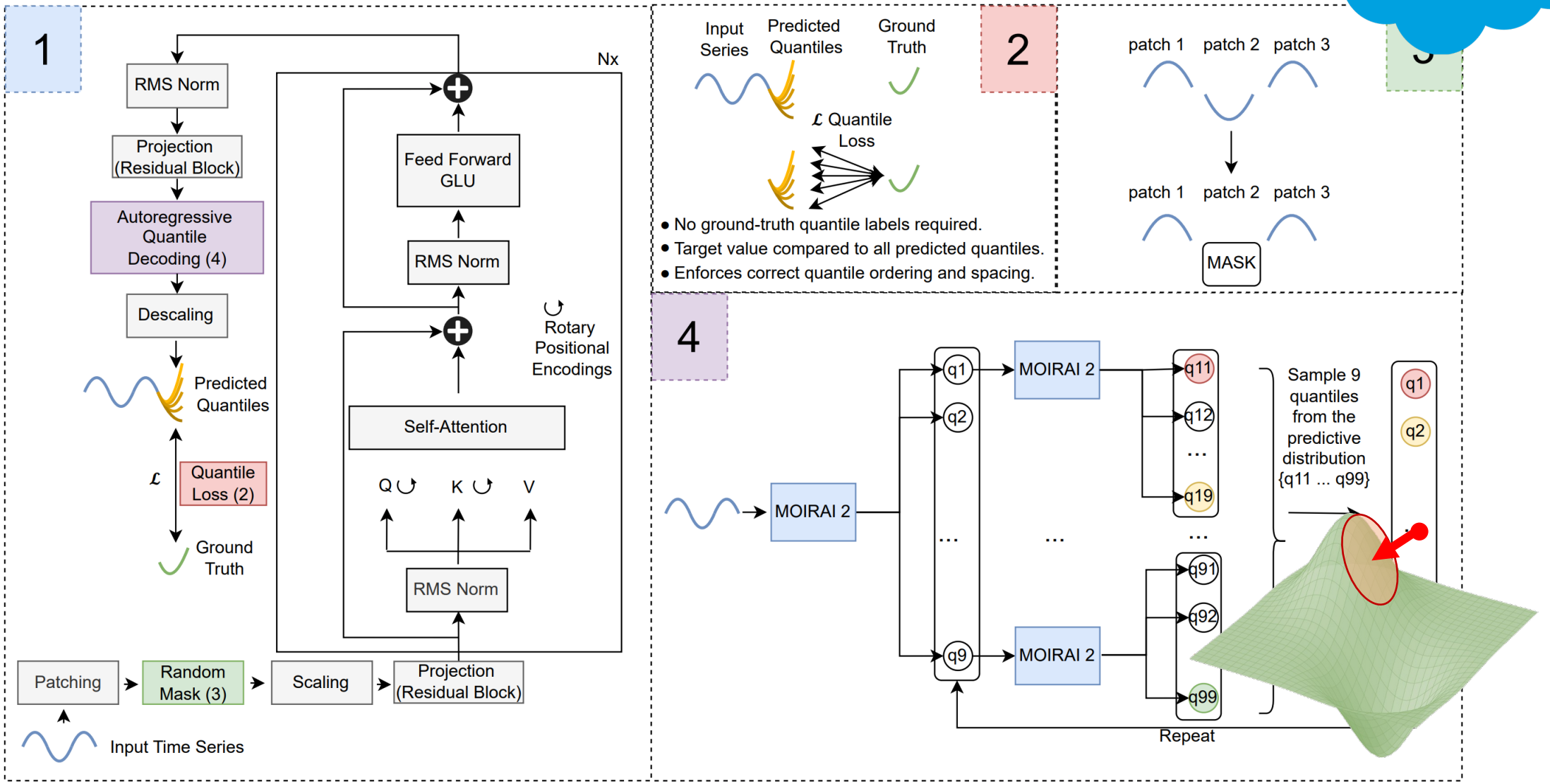
[Moirai-MOE, arXiv 2024]

Mainstream players 1 - Moirai 2.0



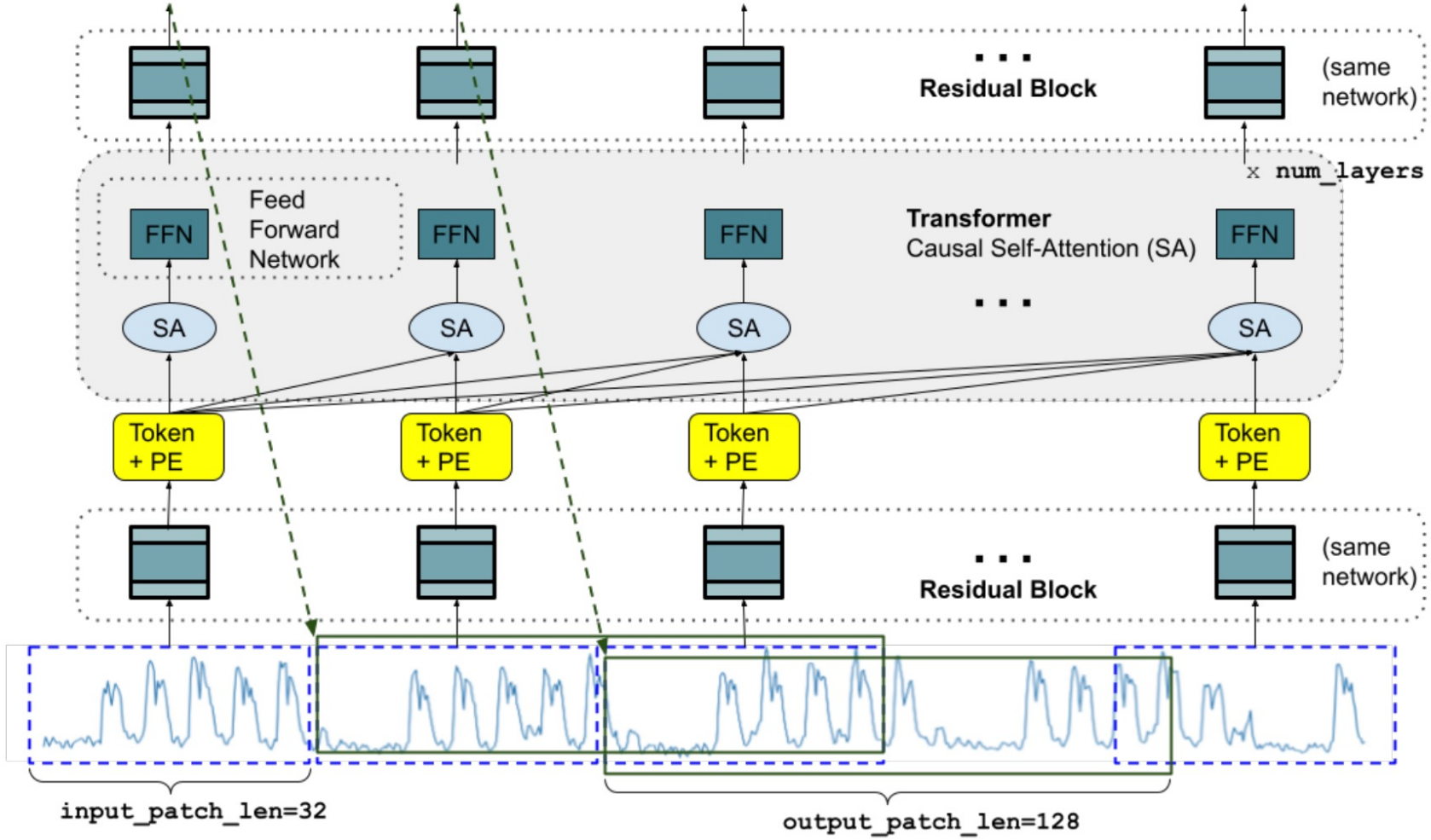
[Moirai 2.0, arXiv 2026]

Mainstream players 1 - Moirai 2.0



[Moirai 2.0, arXiv 2026]

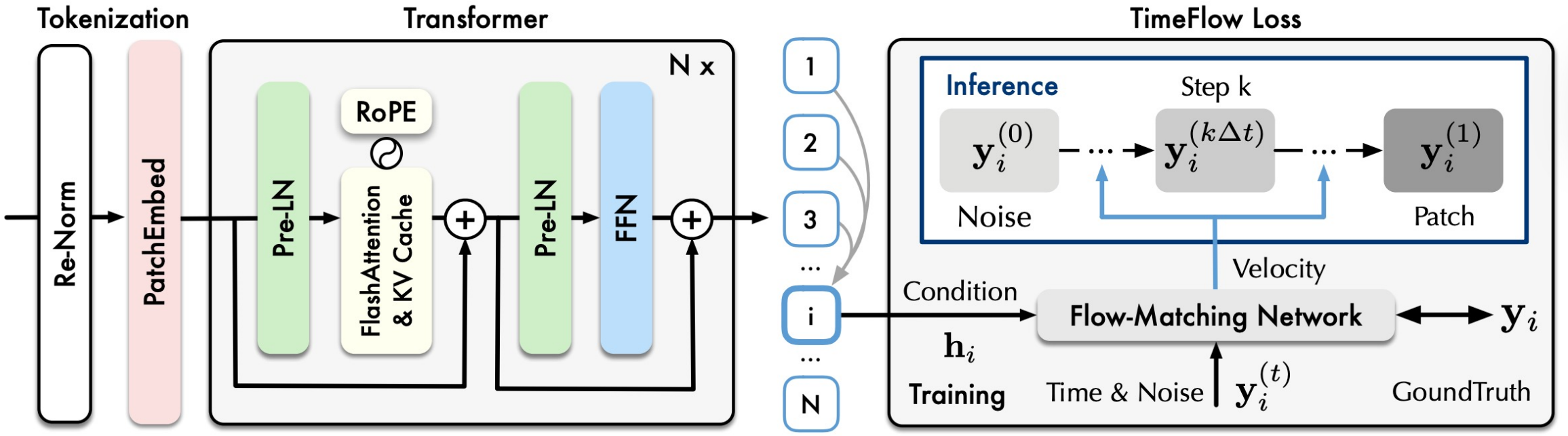
Mainstream players 2 - TimesFM 2.0 / 2.5



[TimesFM, ICML 2024] Google

Keep scaling up training data and context length

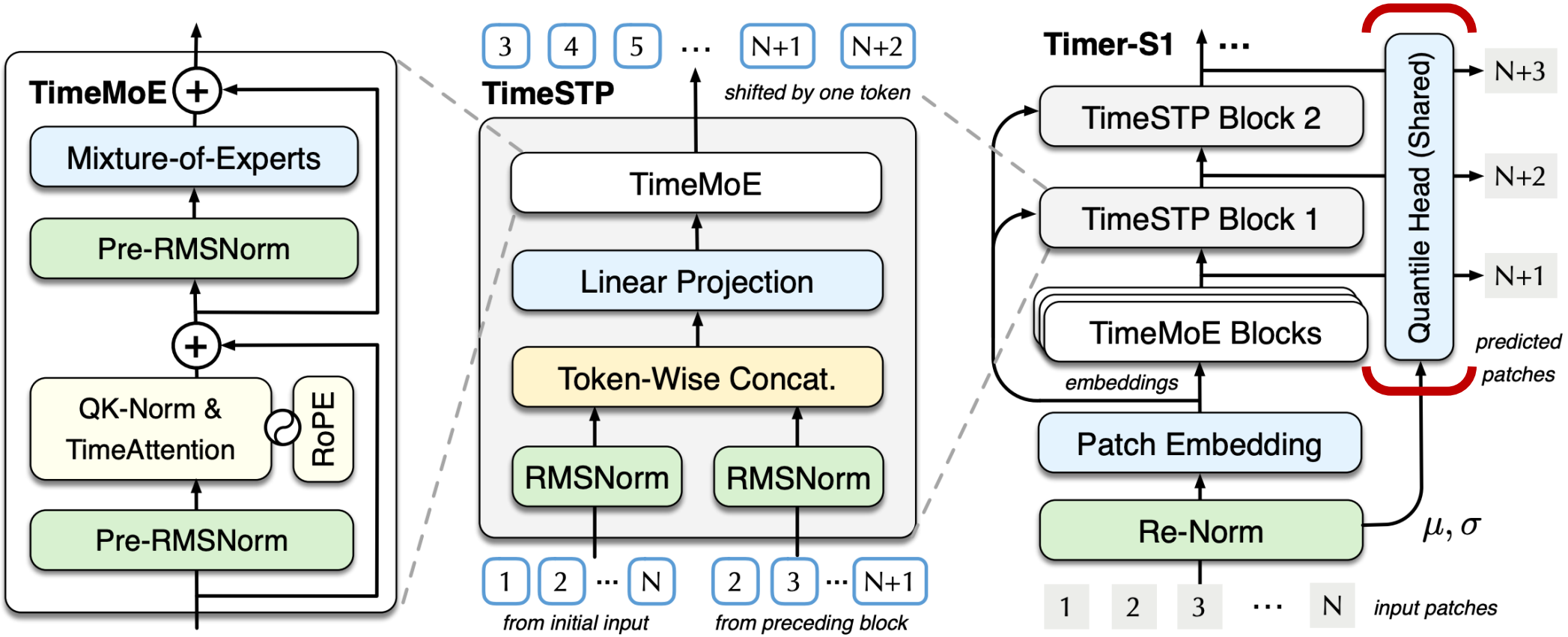
Mainstream players 3 - Sundial



[Sundial, ICML 2025] Tsinghua University

400M model trained from 1T time points

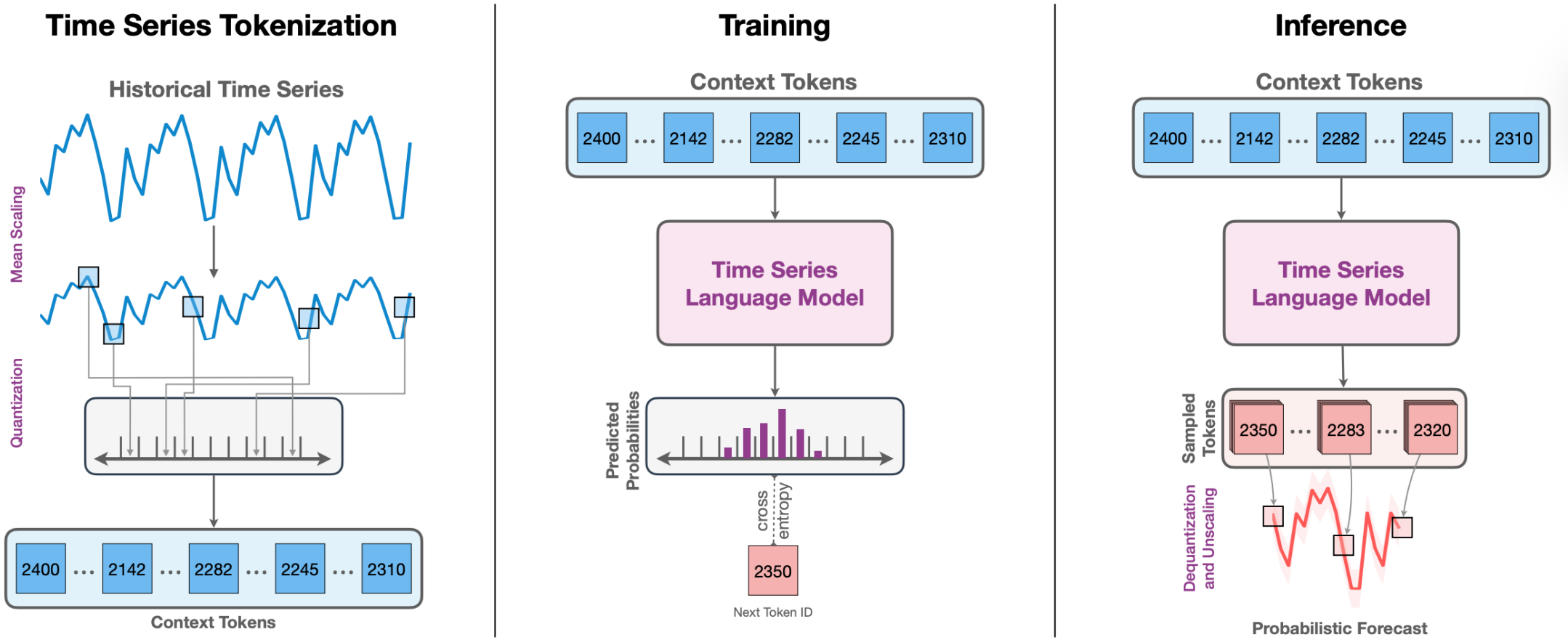
Mainstream players 3 - Timer-S1



[Timer-S1, arXiv 2026] Tsinghua University + Bytedance

8B model from 1T time points

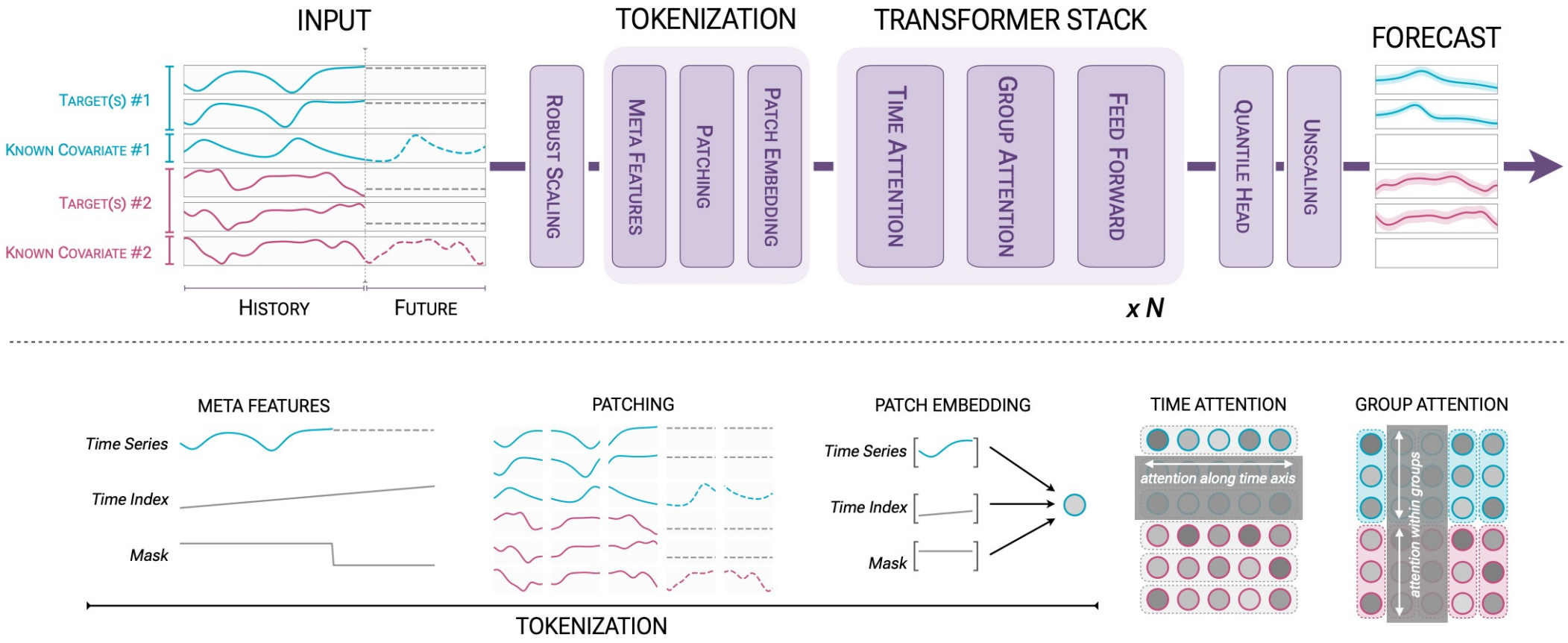
Mainstream players 4 - Chronos



[Chronos, TMLR 2024] Amazon

710M model trained from 84B time points

Mainstream players 4 - Chronos-2

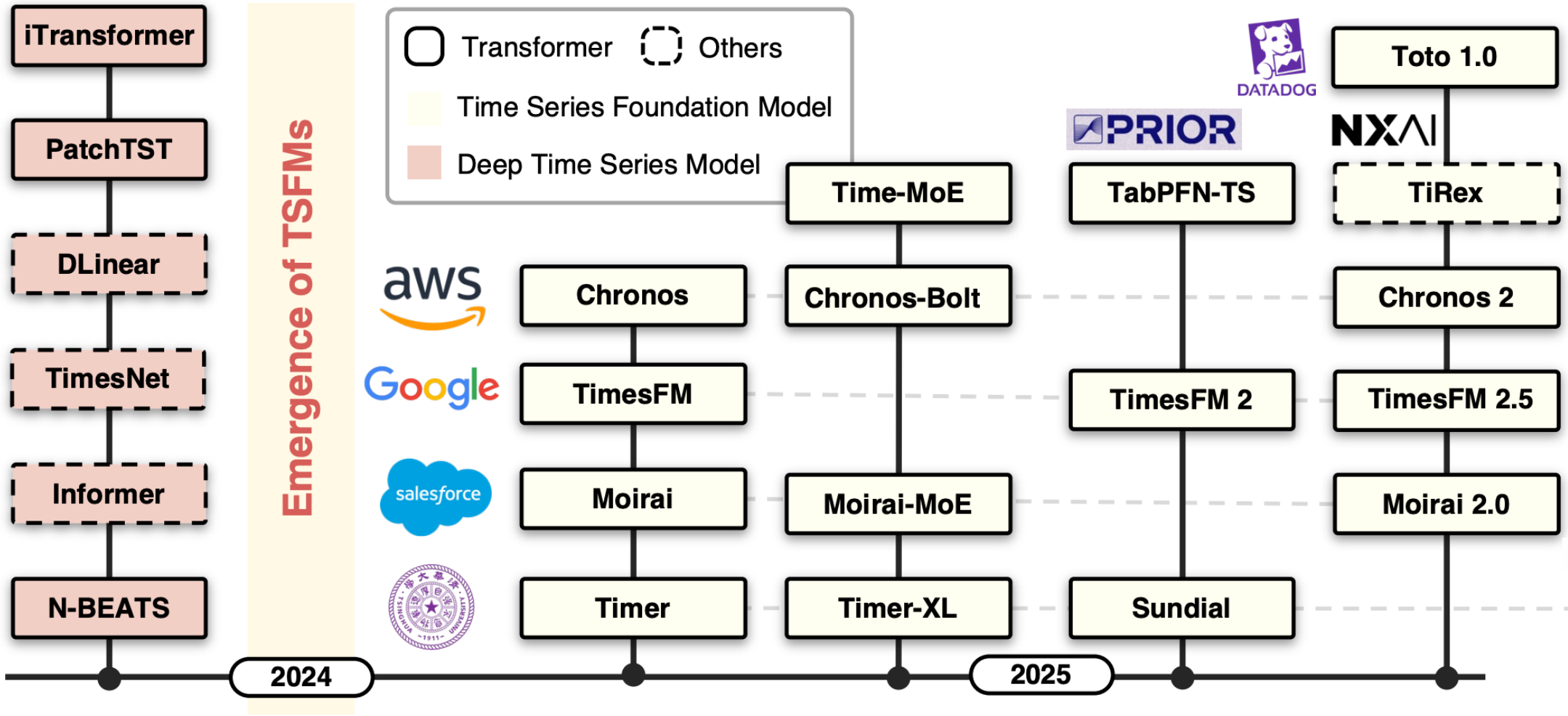


[Chronos-2, arXiv 2025] Amazon

120M model trained from synthetic + real data

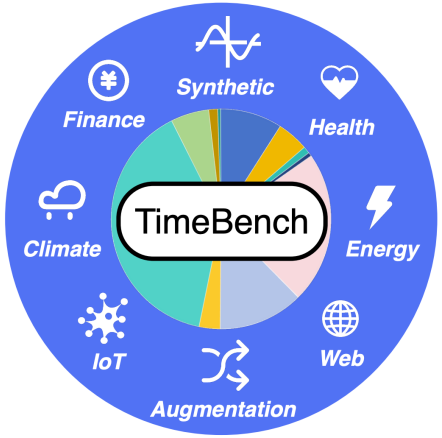
Which model is in the right way in my opinion

A probabilistic decoding head; Quantile prediction is only a metric, not a target.



Does the TSFM exist or not?

$p(\mathbf{x}|\mathbf{z})$ Pre-training is to learn the underlying structure of data, namely, $\mathbf{z}$.



[Sundial, ICML 2025]

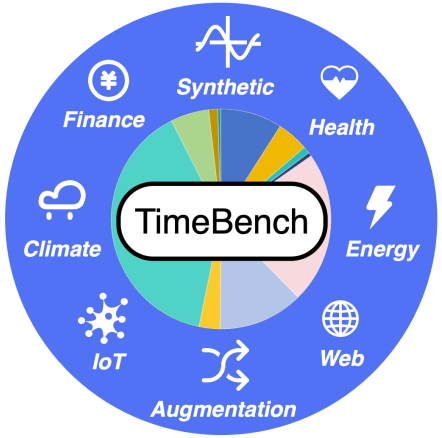
Modeling everything = Modeling nothing.

Does the TSFM exist or not?

$p(\mathbf{x}|\mathbf{z})$ Pre-training is to learn the underlying structure of data, namely, $\mathbf{z}$.

Modeling everything = Modeling nothing.

$$p(\mathbf{x}|\mathbf{z}, \underbrace{\text{domain}}_{\text{Optional}})$$



[Sundial, ICML 2025]

Modeling domain-specific knowledge.
At least the domain info should be given.

The “right” way of TSFM in my opinion

- ✓ A **diffusion model**.

In general, a **diffusion model** is easier to train than an autoregressive model.

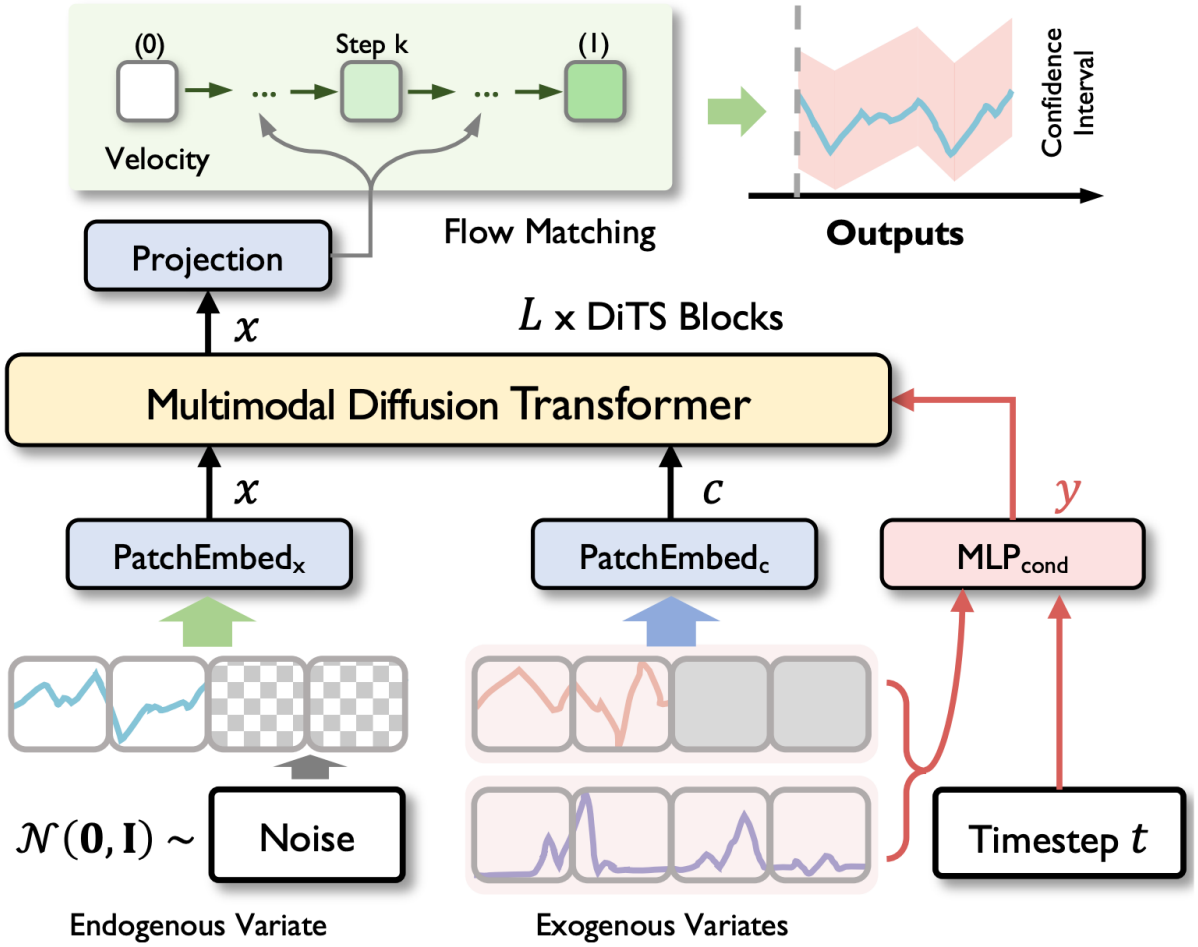
- ✓ A **conditional generation** framework;
- ✓ Incorporate additional information, such as **covariate and language** instructions.

The “right” way of TSFM in my opinion

- ✓ A **diffusion model**.

In general, a **diffusion model** is easier to train than an autoregressive model.

- ✓ A **conditional generation** framework;
- ✓ Incorporate additional information, such as **covariate and language** instructions.



Zhang et al., DiTS: Multimodal Diffusion Transformers Are Time Series Forecasters. arXiv 2026 (Tsinghua University)

Lifted Framework in Dynamical Systems (Time Series)

$p(\mathbf{x})$ ---------- **Vanilla Time Series Forecasting**

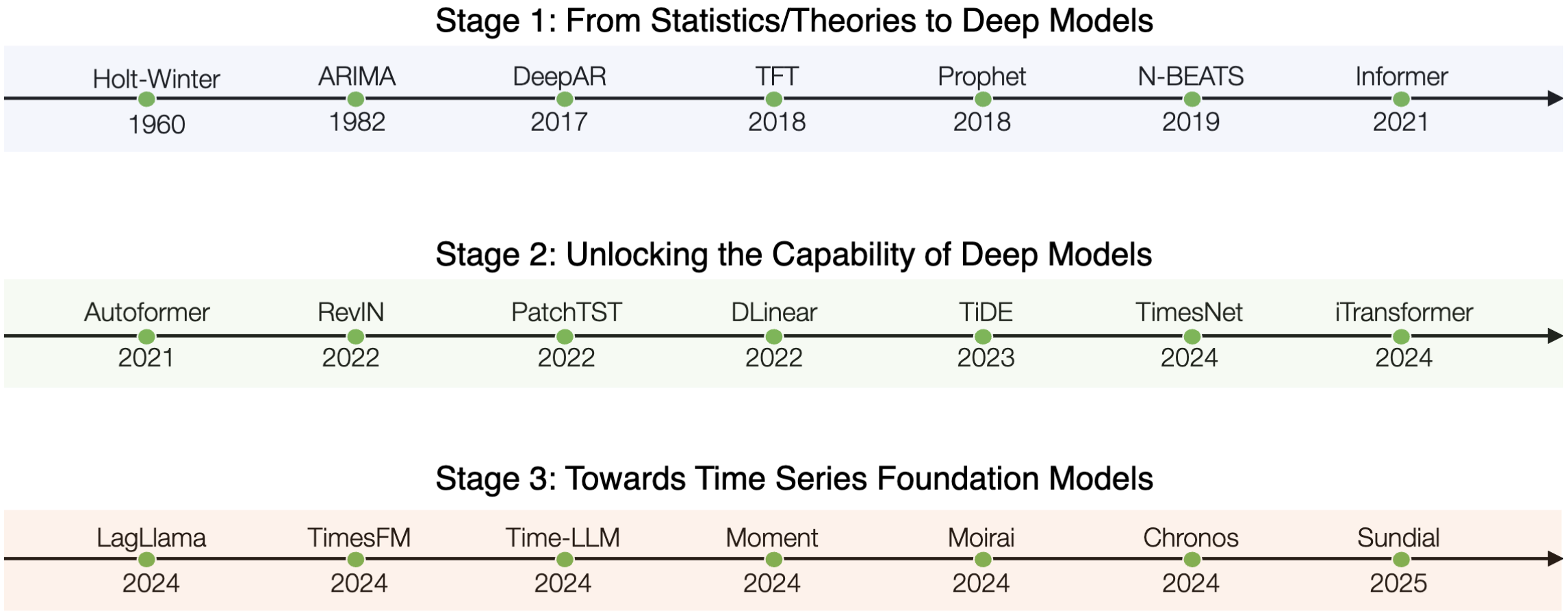
$p(\mathbf{x}, \mathbf{x}_{\text{long-term}})$ ------ **Long-term Forecasting** — Autoformer

$p(\mathbf{x}, \text{ex})$ ---------- **Forecasting with Exogenous Variables** — TimeXer

$p(\mathbf{x}|\mathbf{z})$ ---------- **Large-scale Pre-training** — Some Discussion

* Omit the shared conditional variables, such as past observations.

What's next of time series forecasting



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

https://cloud.tsinghua.edu.cn/f/8d526337afde465e87c9/

Multimodal model for world simulator

IDEAS AI

Spatial Intelligence Is AI's Next Frontier

ADD TIME ON GOOGLE

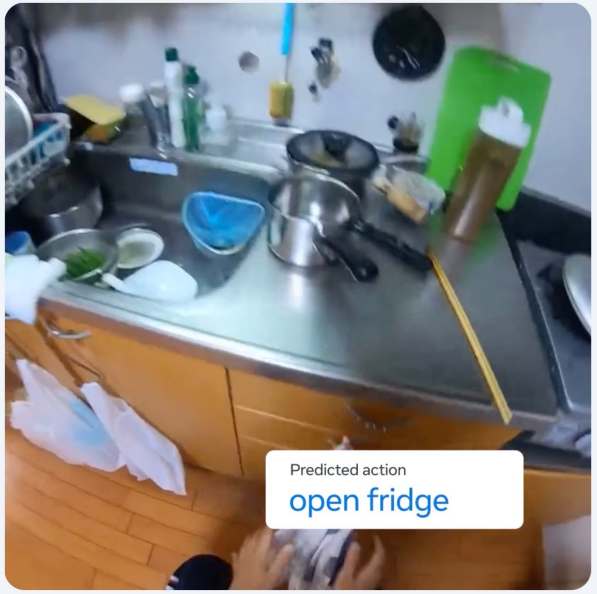
by Fei-Fei Li

Li is co-director of Stanford's Human-Centered AI Institute and co-founder CEO of World Labs

DEC 11, 2025 7:52 AM ET

Language **+ 3D scene**

By Prof. FeiFei Li



Unlock world understanding

V-JEPA 2 delivers exceptional motion understanding as well as leading visual reasoning capabilities when combined with language modeling.

Anticipate what's next

V-JEPA 2 can make predictions about how the world will evolve, setting a new state-of-the-art in anticipating actions from contextual cues.

Language **+ Interaction**

By Prof. Yann LeCun

Takeaways

1. Intelligence = Encoding + Probabilistic decoding

Encoding is not the whole part of intelligence. After “encoding the information”, we need a probabilistic decoding to “release” the compressed intelligence.

2. The lifting framework for prediction

Time series forecasting should embrace the “lifted framework”; otherwise, it will suffer from the performance bottleneck (identified by the accuracy law).

3. The “right” way to construct time series foundation models: Multimodal model

Going beyond the time series native space, which is less informative, try to incorporate domain-specific information as conditions and more modality data.

Acknowledgement

Mingsheng Long

Jianmin Wang

Wojciech Matusik

Jiaxiang Dong

Yong Liu

Tengge Hu

Yuxuan Wang

Hang Zhou

Yuezhou Ma

Haoran Zhang

World Simulators

Toward Intelligent Dynamical System Simulation

Haixu Wu

MIT CSAIL

May 11, 2026