Unified Training of Universal Time Series Forecasting Transformers

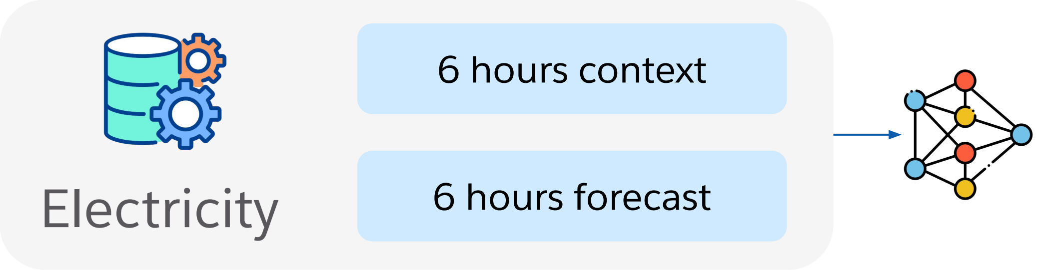
Gerald Woo[1,2], Chenghao Liu[1], Akshat Kumar[2], Caiming Xiong[1], Silvio Savarese[1], Doyen Sahoo[1]

[1] Salesforce AI Research

[2] Singapore Management University

Existing Deep Forecasting Paradigm

One-model-per-(dataset, context length, prediction length)

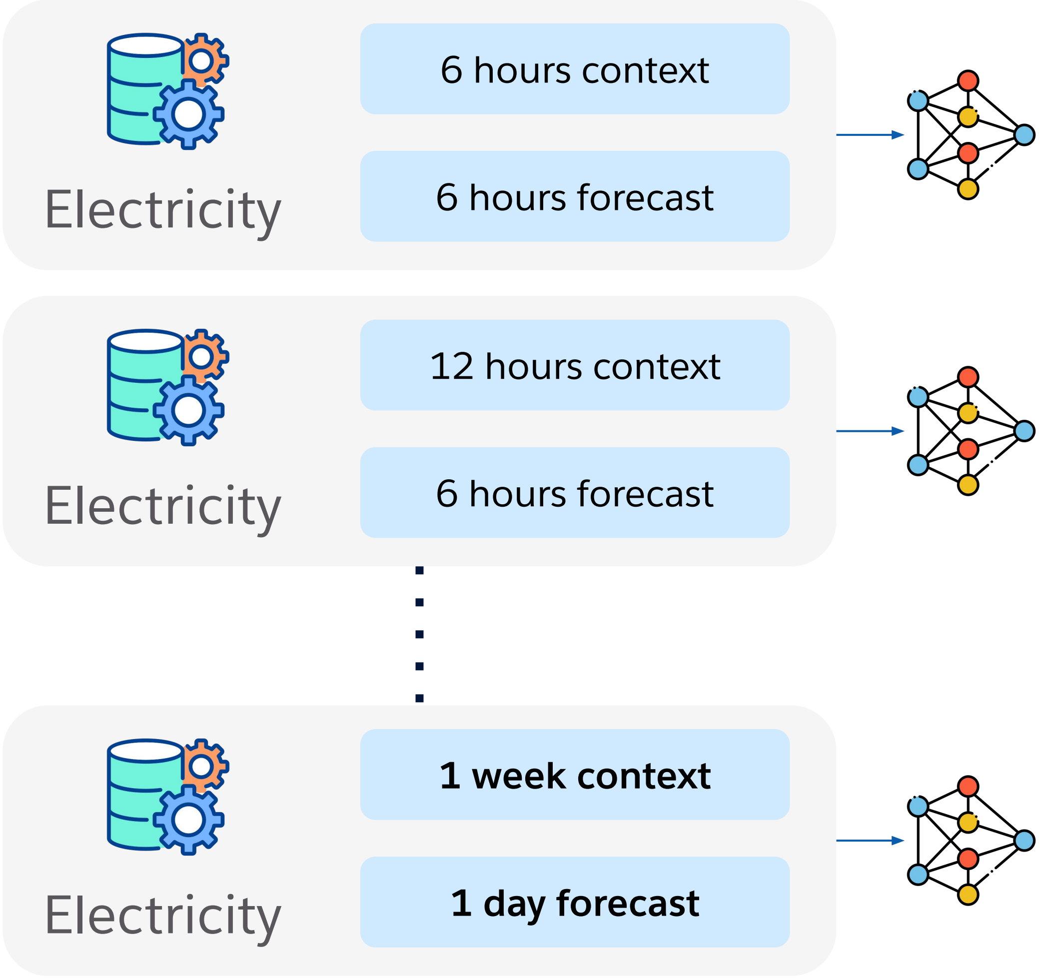


Existing Deep Forecasting Paradigm

One-model-per-(dataset, context length, prediction length)

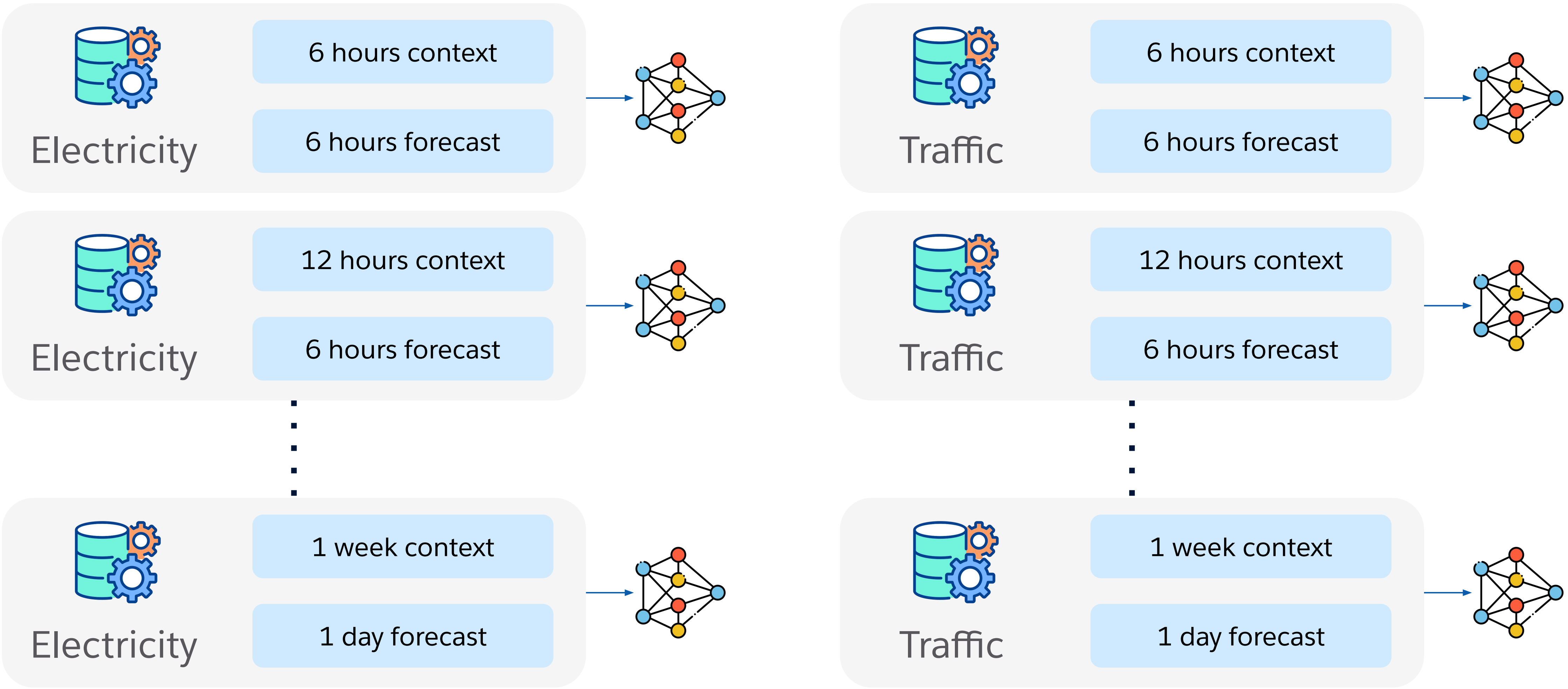
Existing Deep Forecasting Paradigm

One-model-per-(dataset, context length, prediction length)

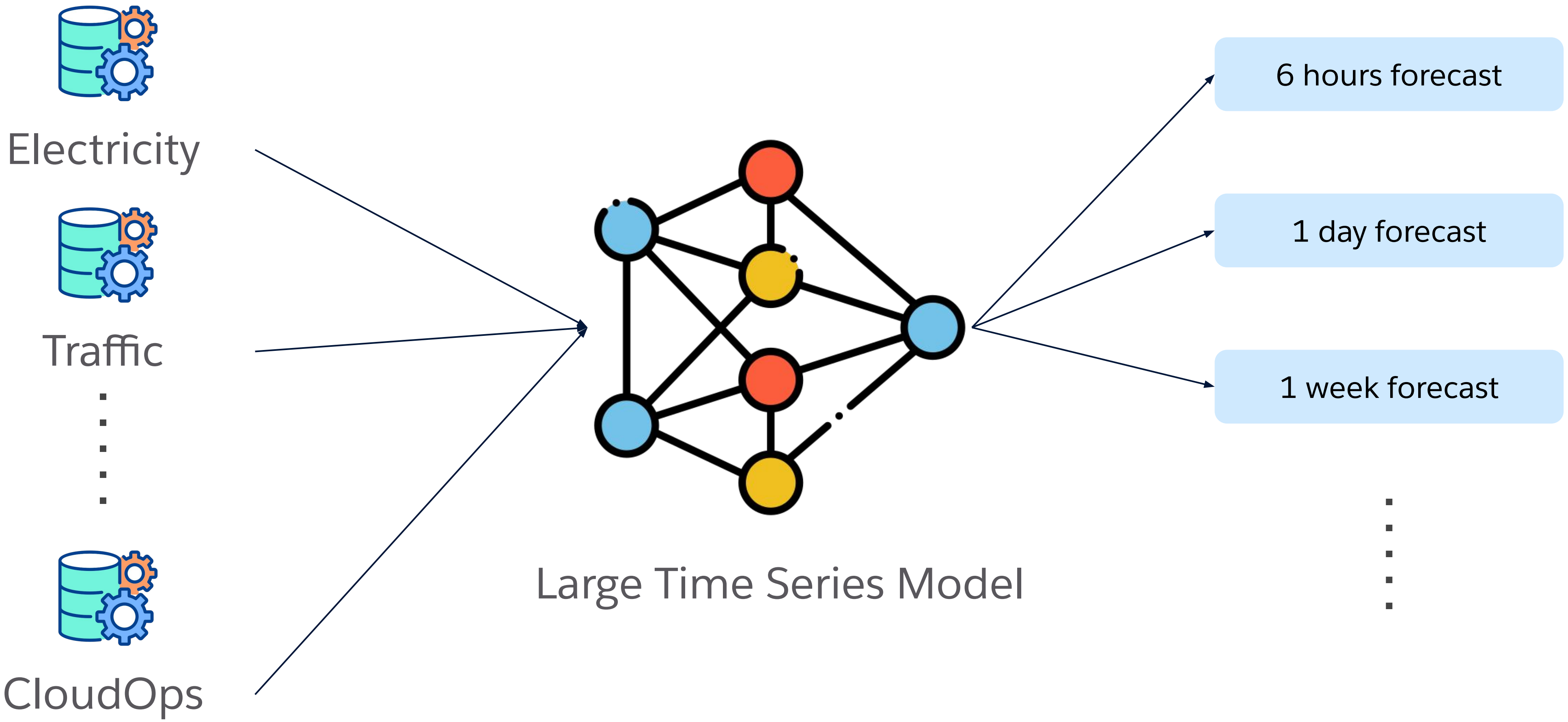


Existing Deep Forecasting Paradigm

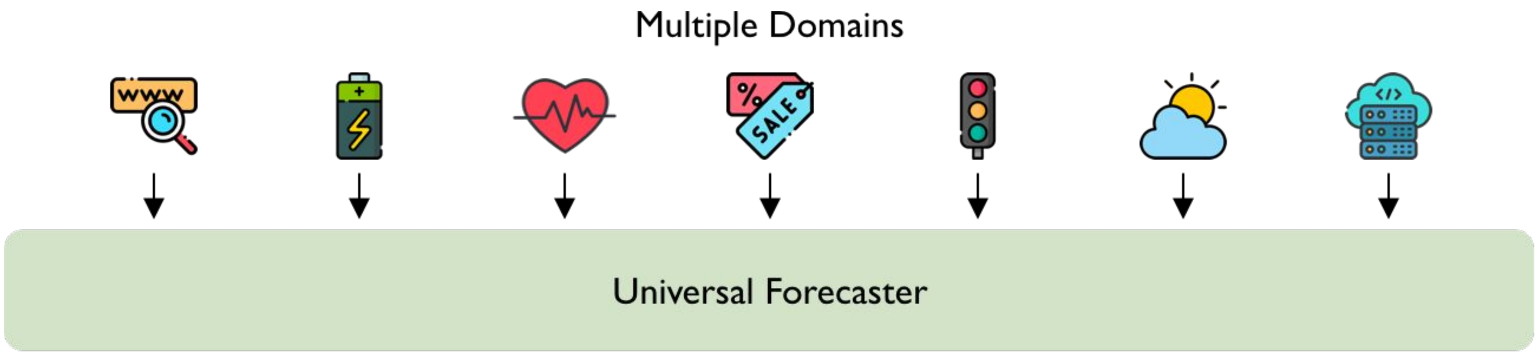
One-model-per-(dataset, context length, prediction length)



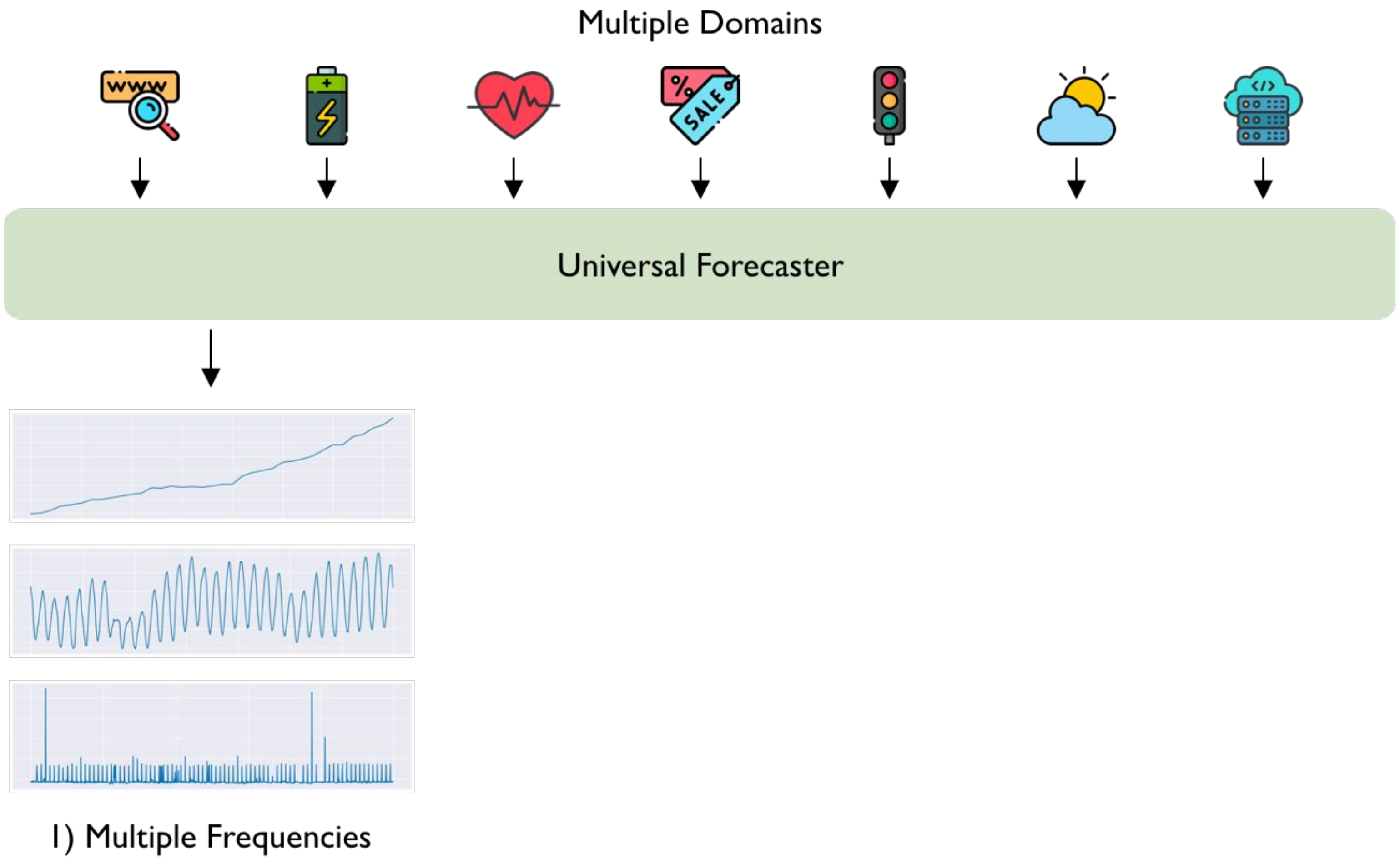
Universal Forecasting



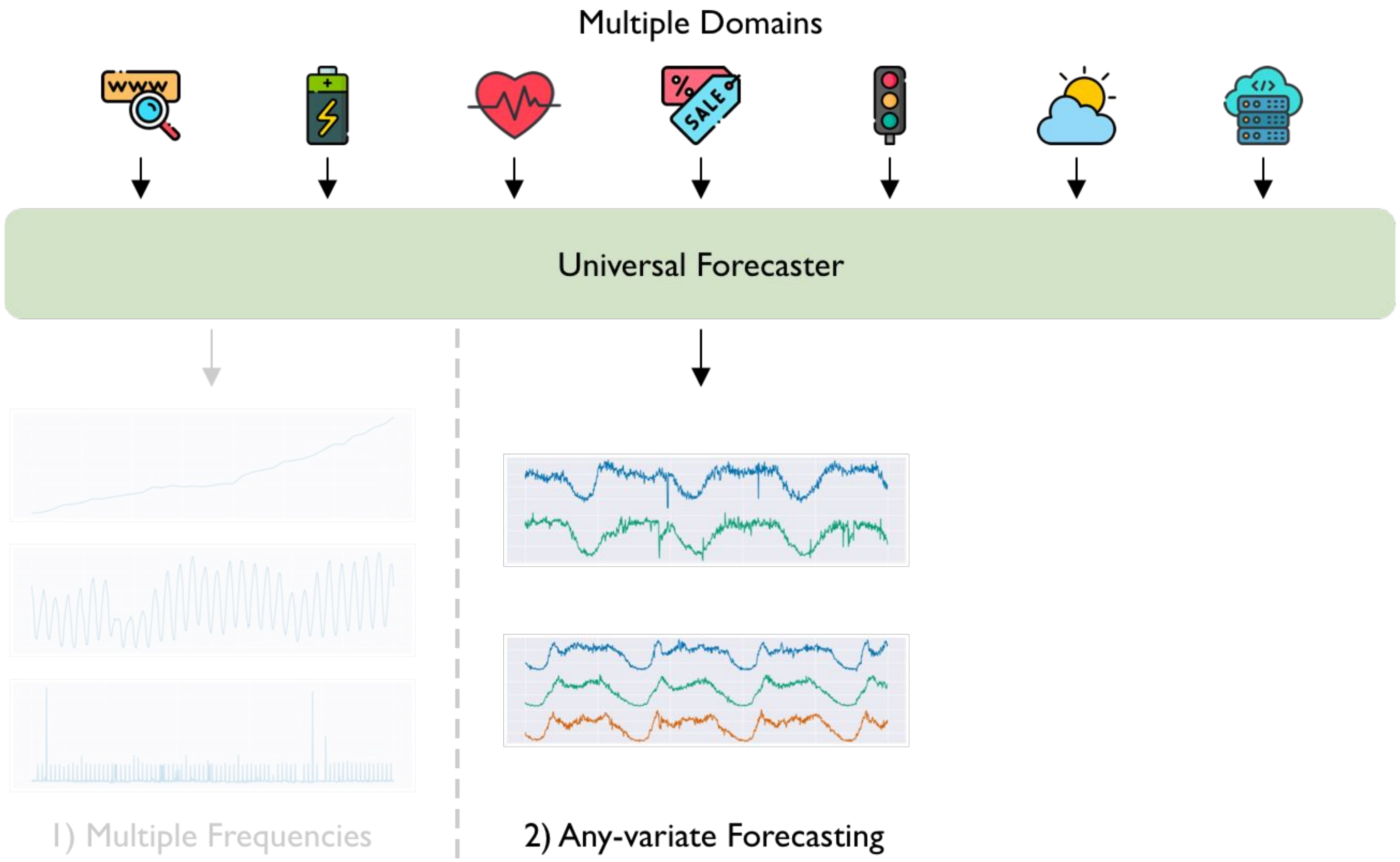
Challenges to Universal Forecasting



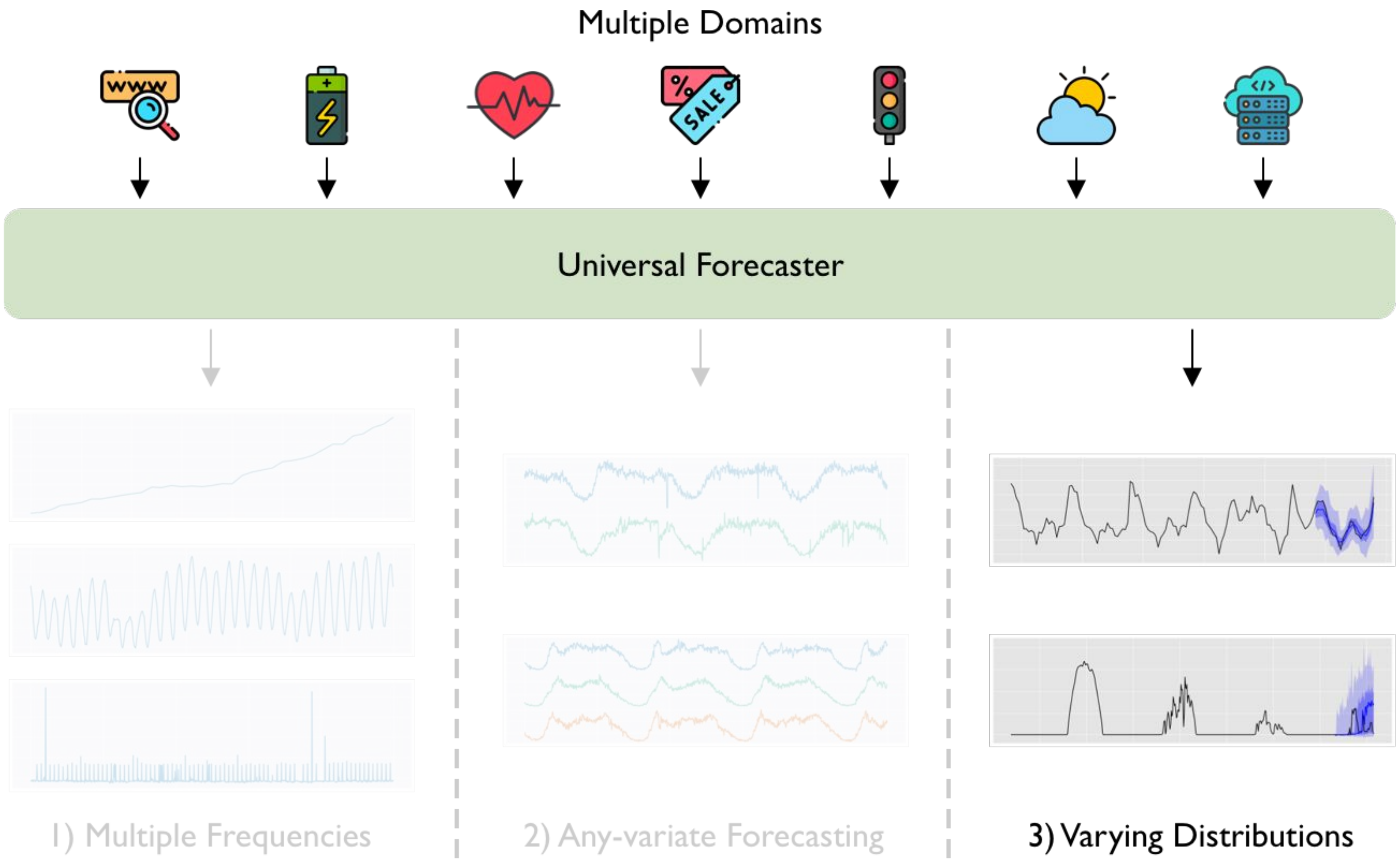
Challenges to Universal Forecasting



Challenges to Universal Forecasting

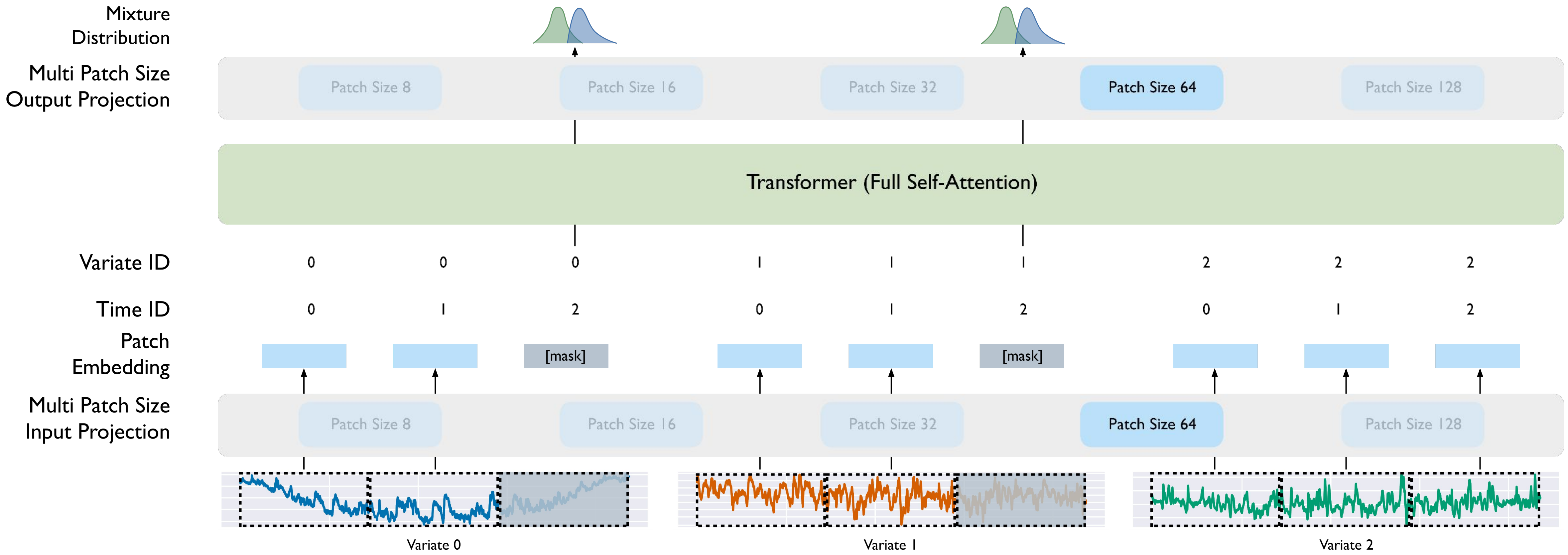


Challenges to Universal Forecasting



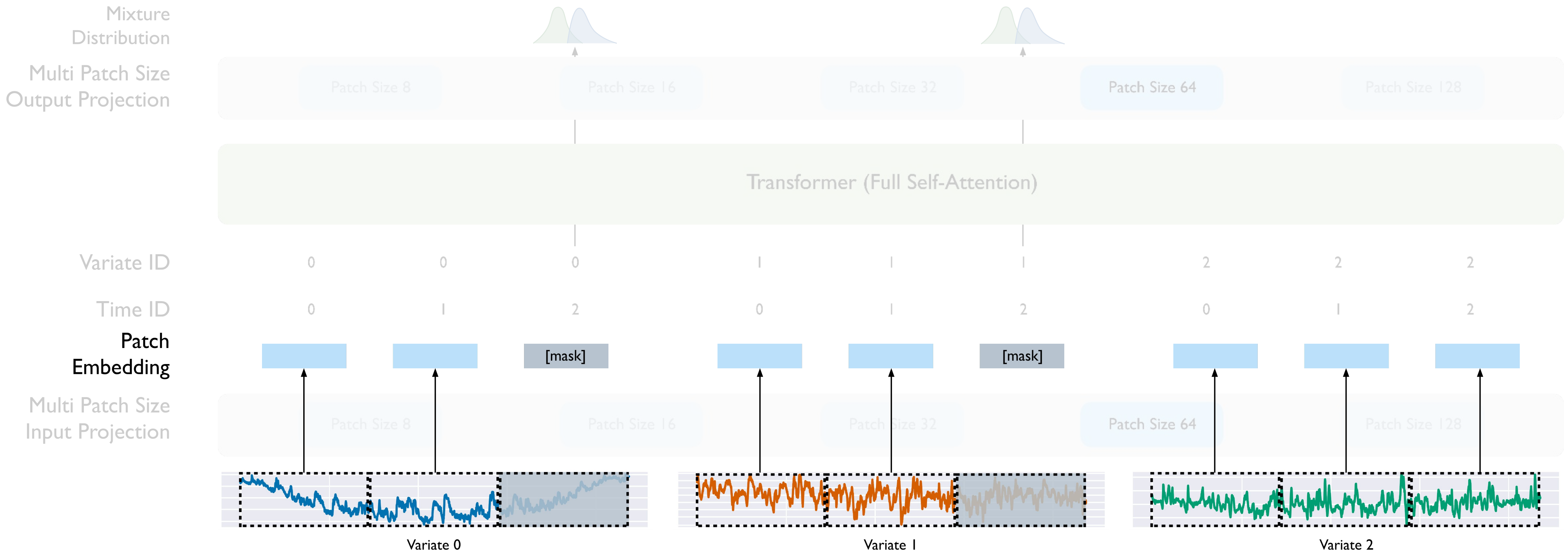
MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

Overall architecture



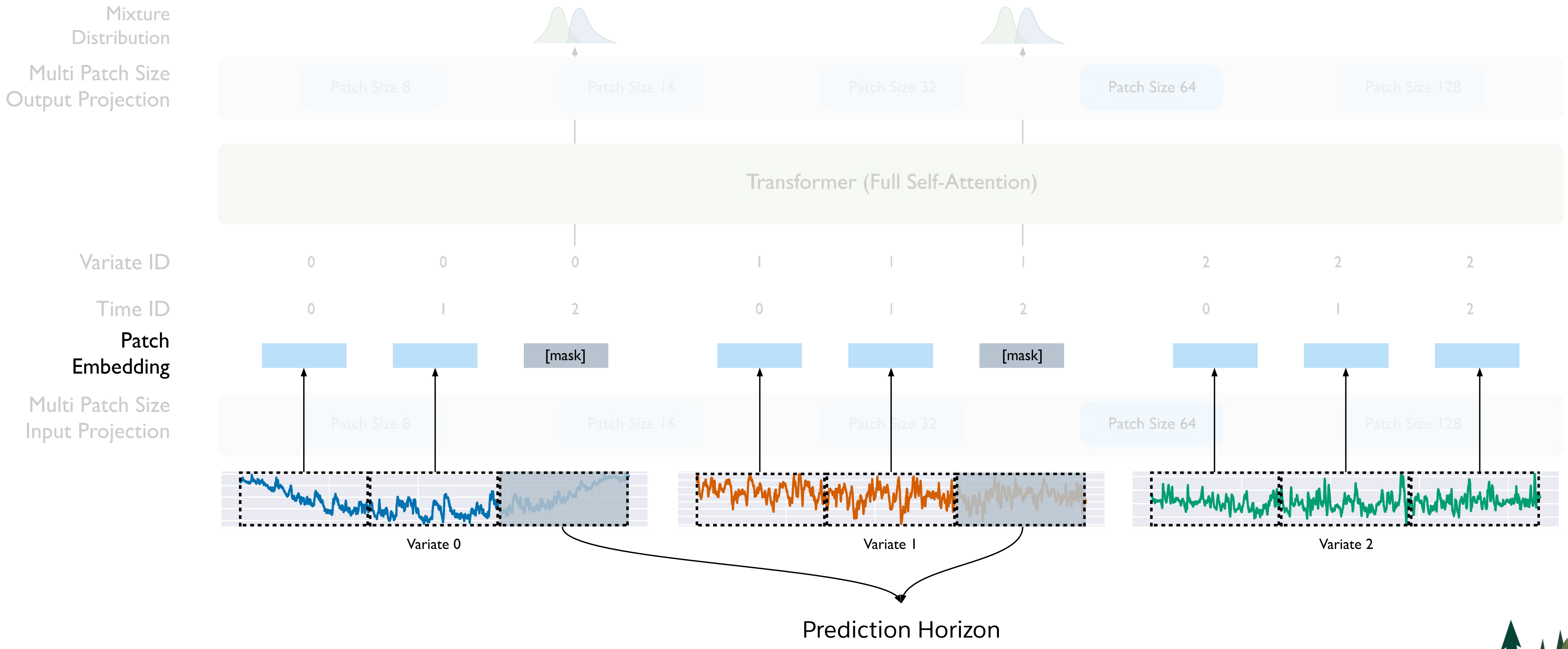
MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

Overall architecture



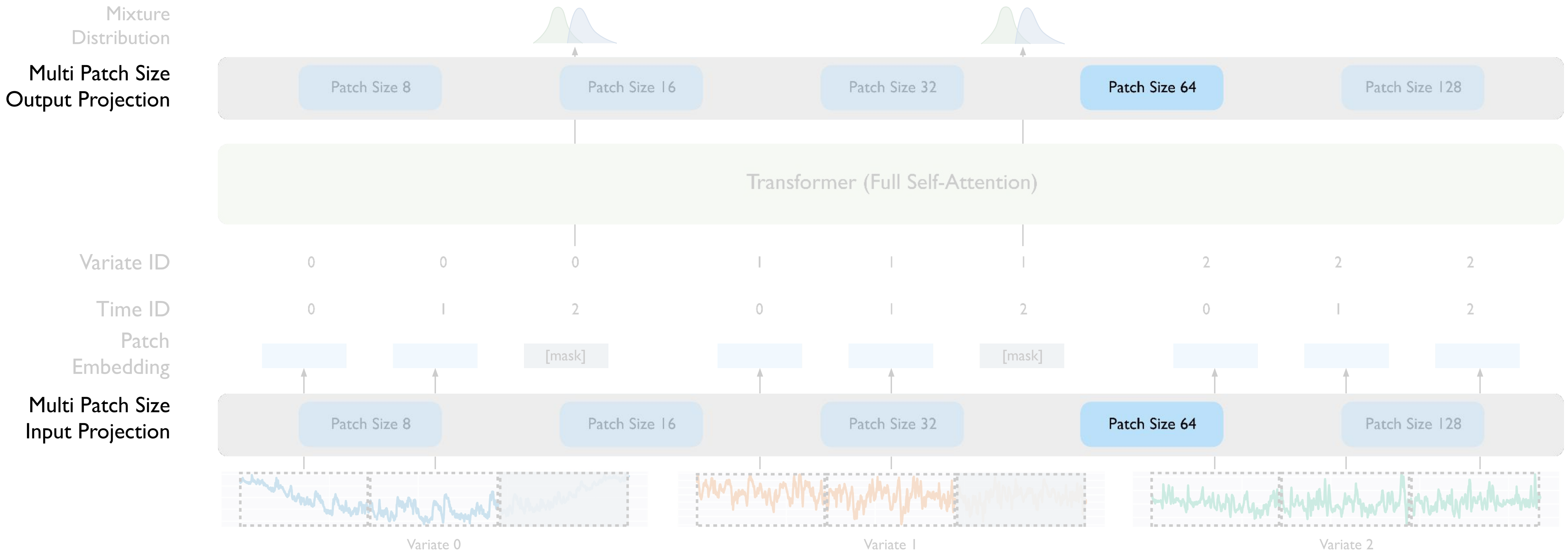
MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

Overall architecture



MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

1) Multi Input/Output Patch Projections

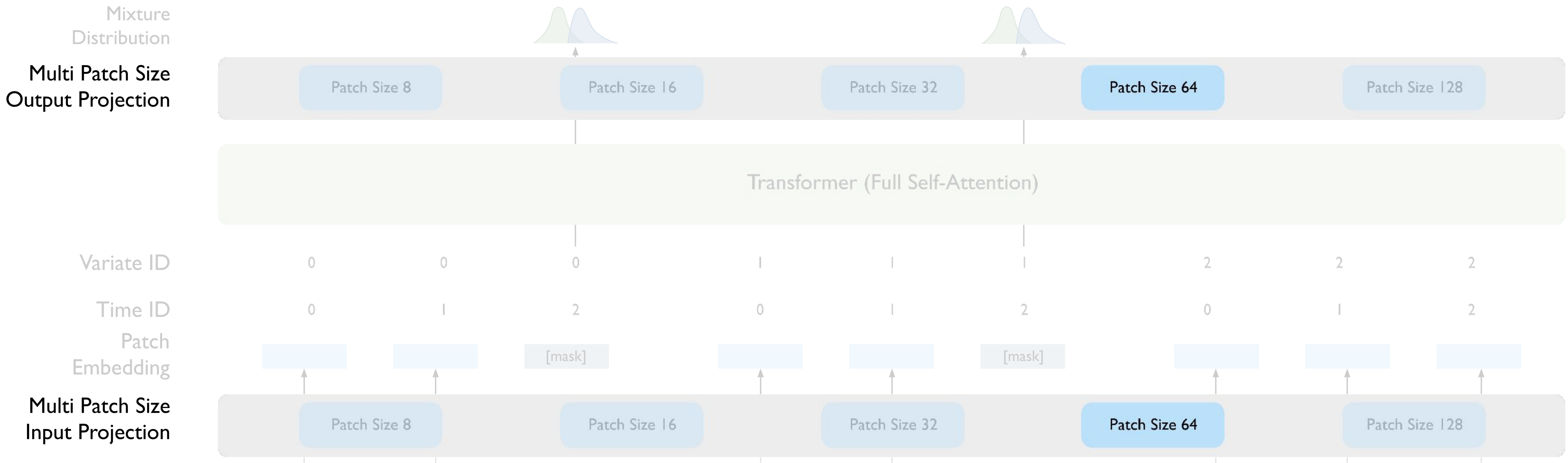


MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

SMU SINGAPORE MANAGEMENT UNIVERSITY

salesforce

1) Multi Input/Output Patch Projections

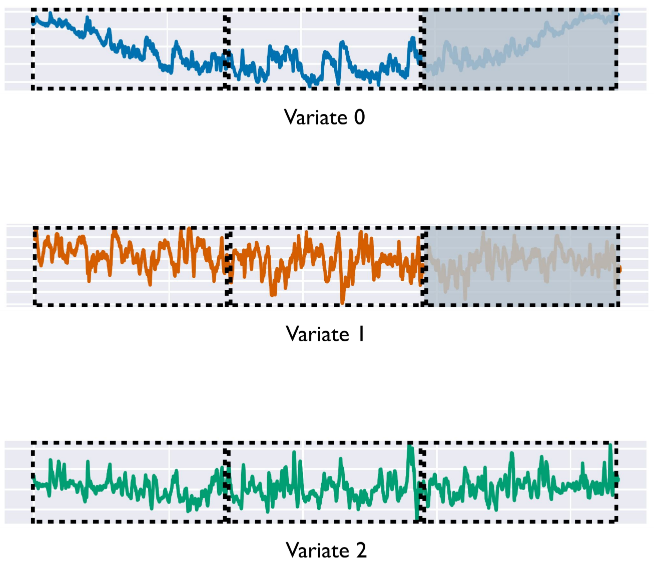


- Frequency-based patch size mapping
  - Low frequency → smaller patch size
  - High frequency → larger patch size

- Yearly, Quarterly: 8
- Monthly: 8, 16, 32
- Weekly, Daily: 16, 32
- Hourly: 32, 64
- Minute-level: 32, 64, 128
- Second-level: 64, 128

MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

2) Any-variate Attention

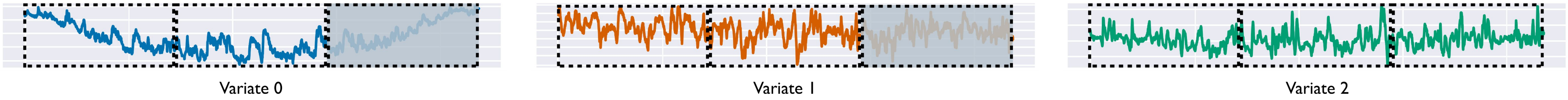


MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

SMU SINGAPORE MANAGEMENT UNIVERSITY

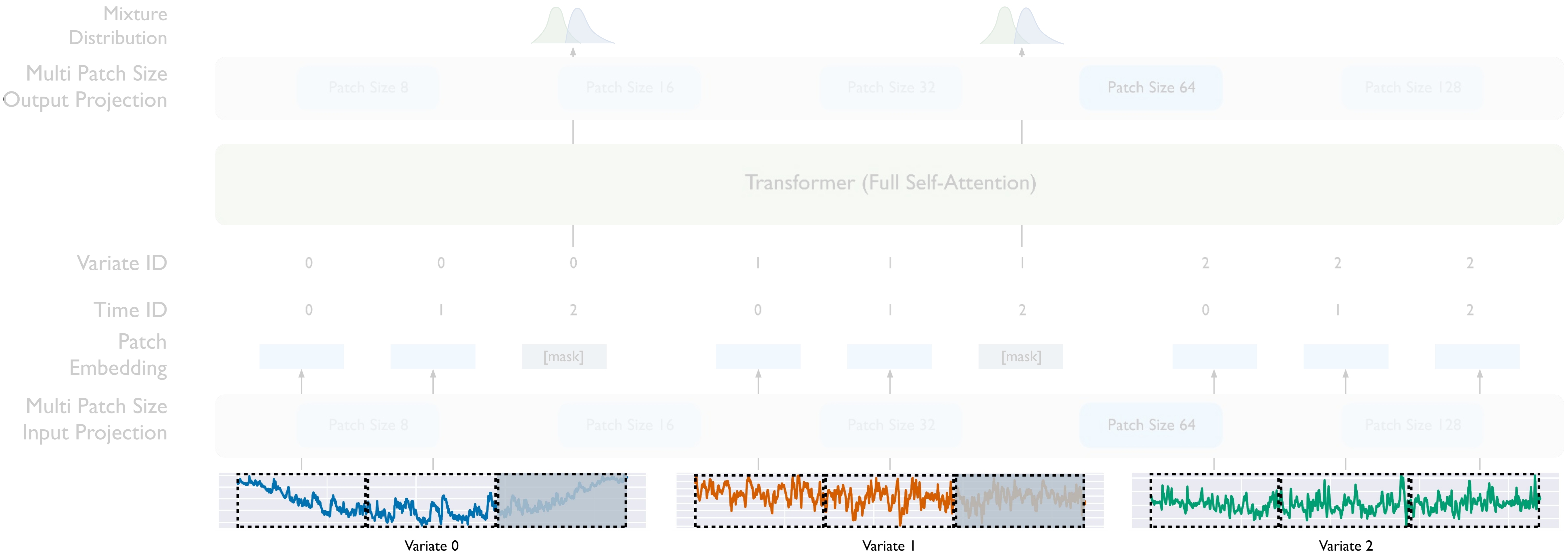
salesforce

2) Any-variate Attention



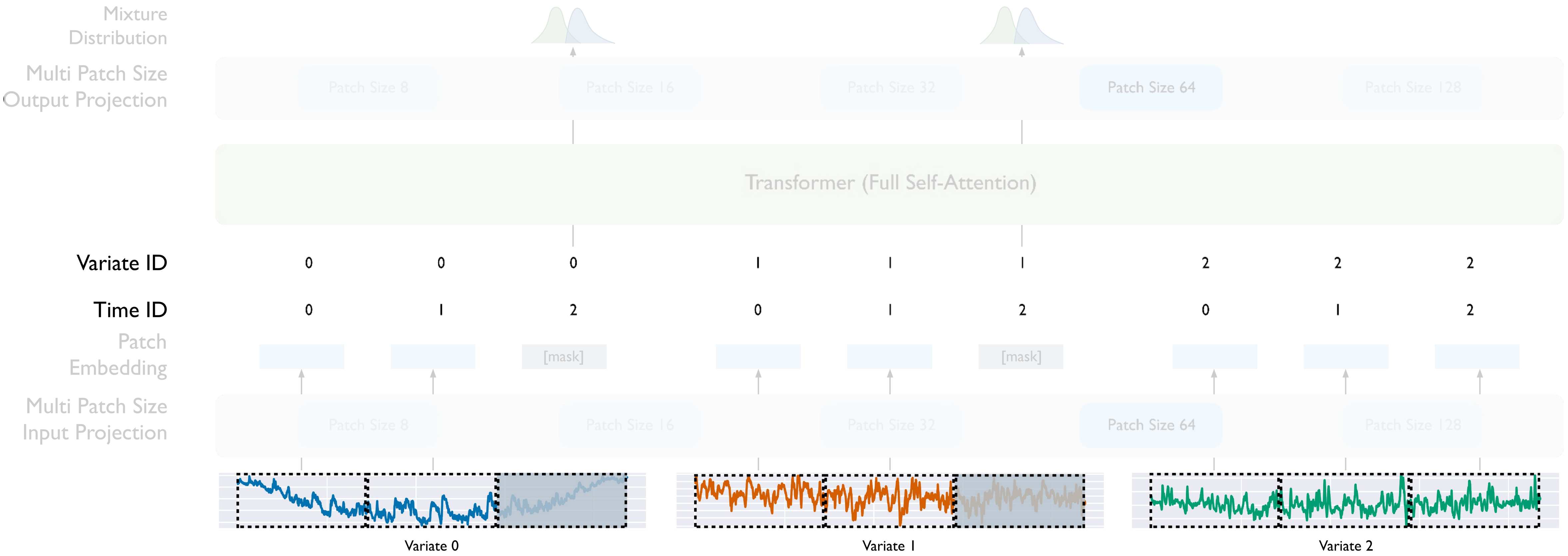
MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

2) Any-variate Attention



MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

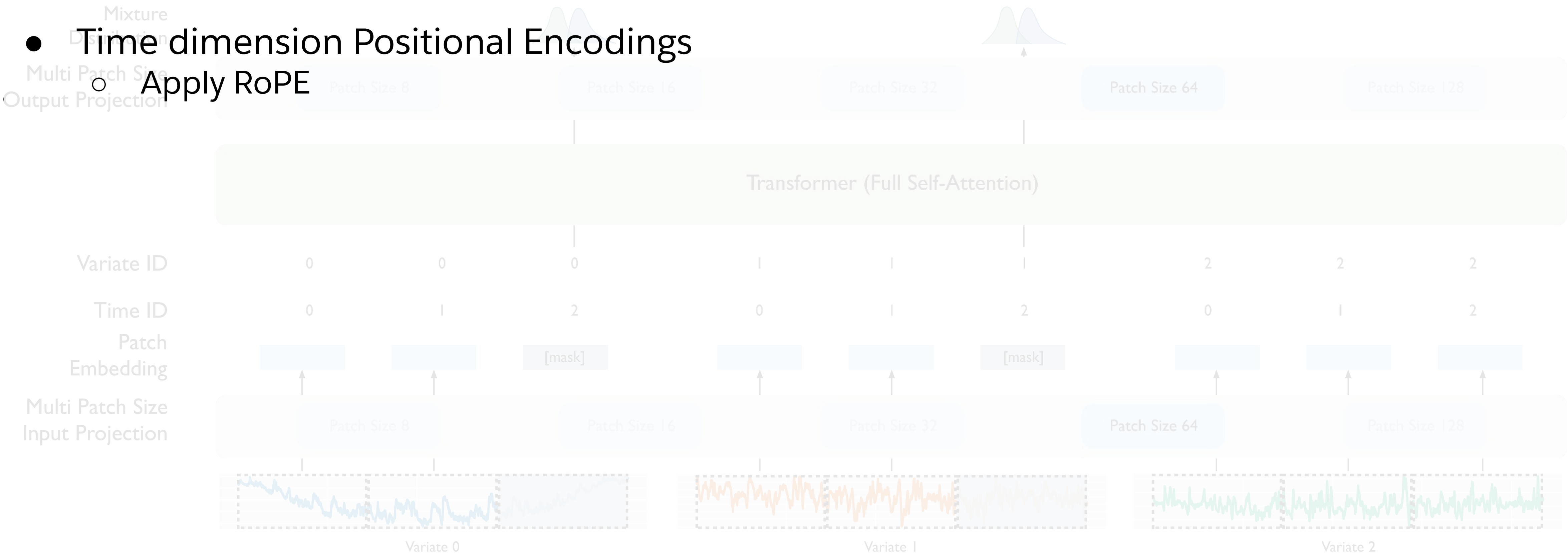
2) Any-variate Attention



MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

2) Any-variate Attention

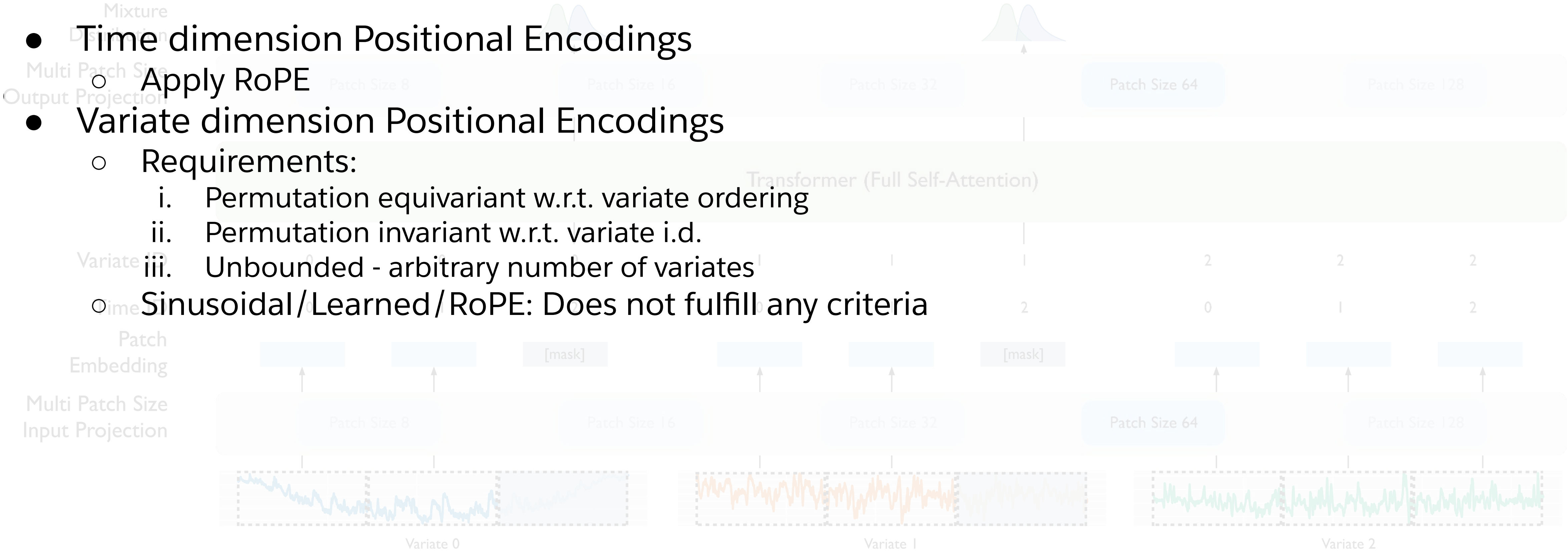
- Time dimension Positional Encodings
  - Apply RoPE



MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

2) Any-variate Attention

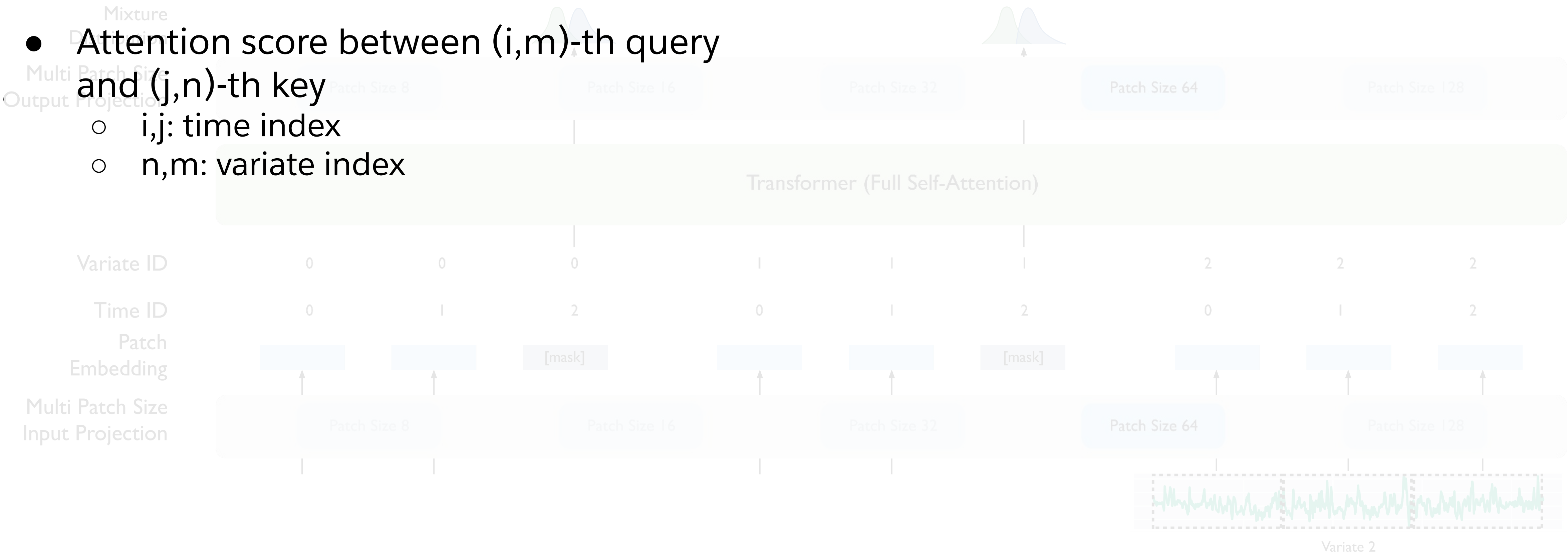
- Time dimension Positional Encodings
  - Apply RoPE
- Variate dimension Positional Encodings
  - Requirements:
    1. Permutation equivariant w.r.t. variate ordering
    2. Permutation invariant w.r.t. variate i.d.
    3. Unbounded - arbitrary number of variates
  - Sinusoidal/Learned/RoPE: Does not fulfill any criteria



MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

2) Any-variate Attention

- Attention score between (i,m)-th query and (j,n)-th key
  - i,j: time index
  - n,m: variate index



MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

SMU SINGAPORE MANAGEMENT UNIVERSITY

salesforce

2) Any-variate Attention

- Attention score between (i,m)-th query and (j,n)-th key
  - i,j: time index
  - n,m: variate index

$$E_{ij,mn} = \overbrace{(\boldsymbol{W}^Q \boldsymbol{x}_{i,m})^T \boldsymbol{R}_{i-j} (\boldsymbol{W}^K \boldsymbol{x}_{j,n})}^{\text{Rotary matrix (RoPE)}}$$

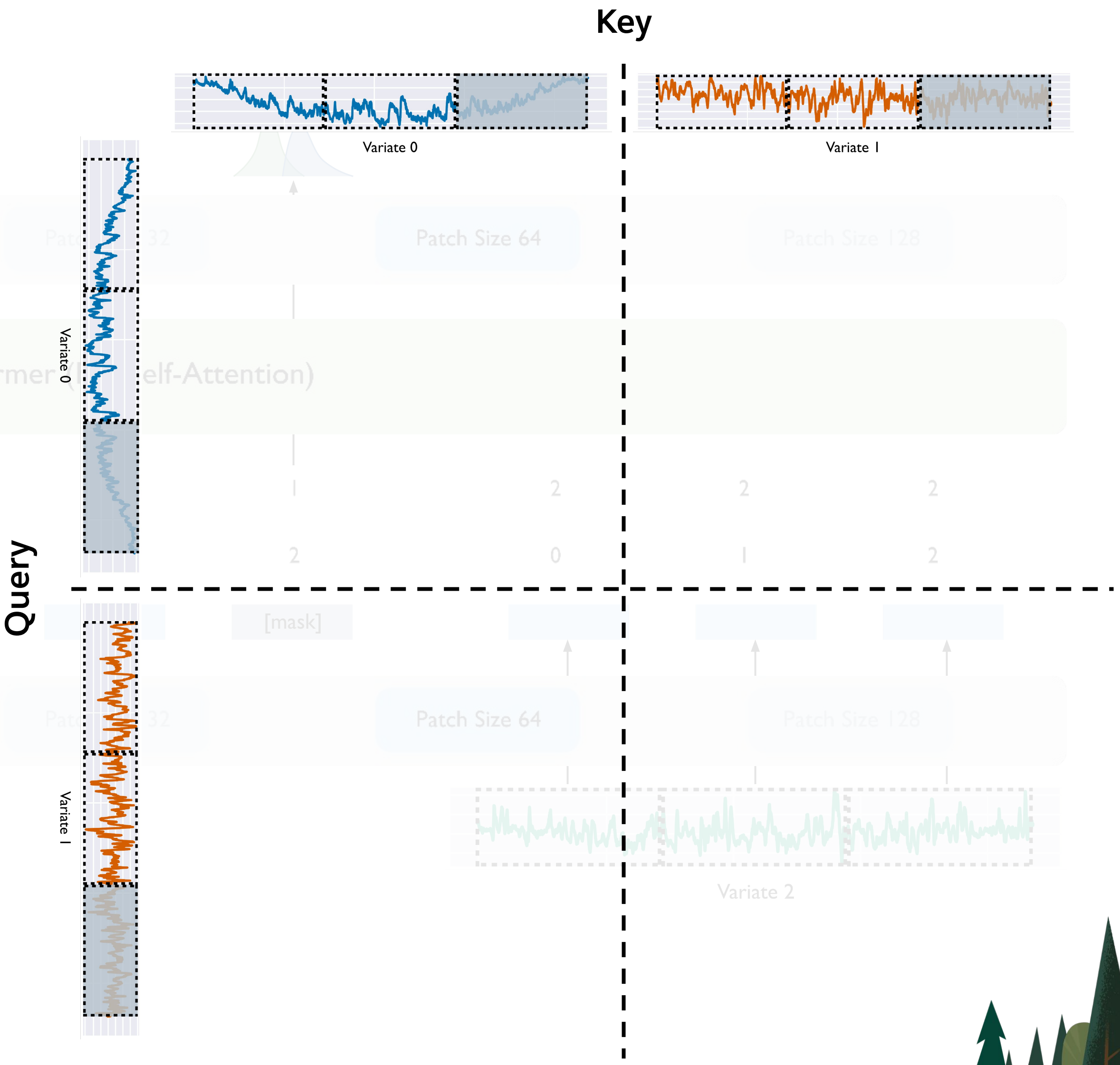
$$A_{ij,mn} = \frac{\exp\{E_{ij,mn}\}}{\sum_{k,o} \exp\{E_{ik,mo}\}},$$

MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

2) Any-variate Attention

- Attention score between (i,m)-th query and (j,n)-th key
  - i,j: time index
  - n,m: variate index

$$E_{ij,mn} = (W^Q x_{i,m})^T R_{i-j} (W^K x_{j,n}) \underbrace{+ u^{(1)} * \mathbb{1}_{\{m=n\}} + u^{(2)} * \mathbb{1}_{\{m \neq n\}}}_{\text{Binary attention bias}},$$



MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

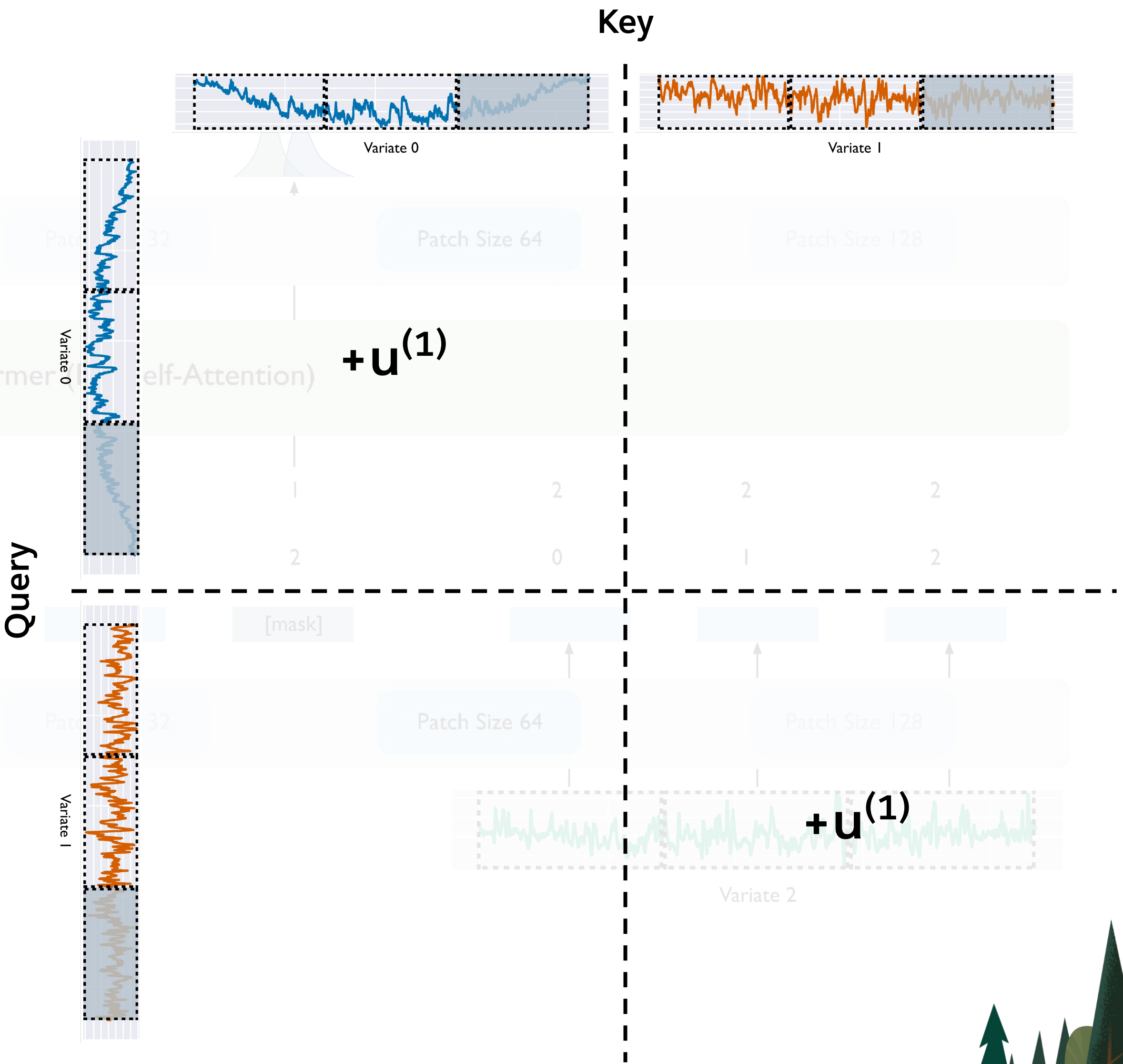
SMU SINGAPORE MANAGEMENT UNIVERSITY

salesforce

2) Any-variate Attention

- Attention score between (i,m)-th query and (j,n)-th key
  - i,j: time index
  - n,m: variate index

$$E_{ij,mn} = (\boldsymbol{W}^Q \boldsymbol{x}_{i,m})^T \boldsymbol{R}_{i-j} (\boldsymbol{W}^K \boldsymbol{x}_{j,n}) \underbrace{\boxed{+ u^{(1)} * \mathbb{1}_{\{m=n\}}} + u^{(2)} * \mathbb{1}_{\{m \neq n\}}}_{\text{Binary attention bias}},$$



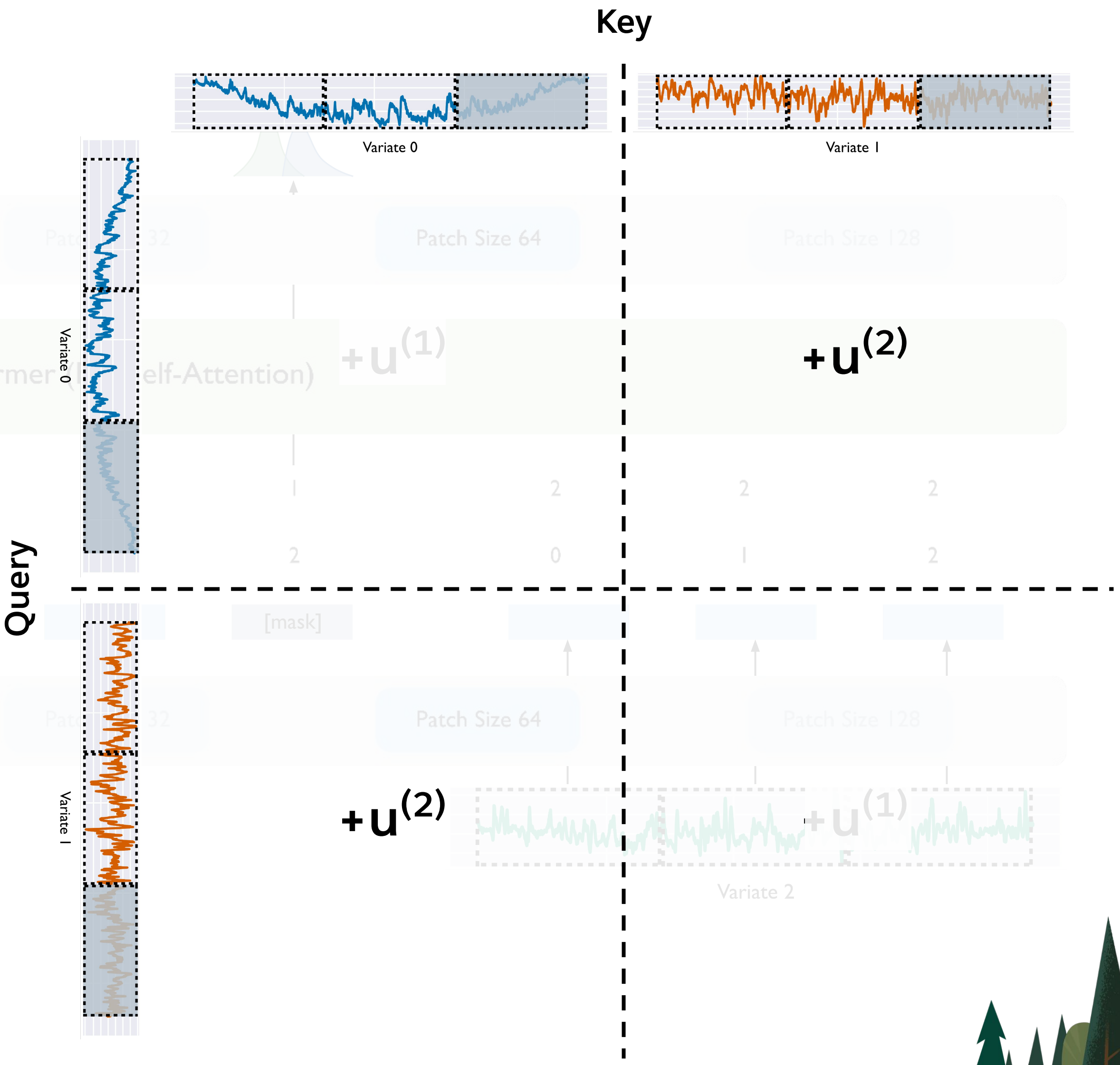
MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

2) Any-variate Attention

- Attention score between (i,m)-th query and (j,n)-th key
  - i,j: time index
  - n,m: variate index

$$E_{ij,mn} = \overbrace{(W^Q x_{i,m})^T R_{i-j} (W^K x_{j,n})}^{\text{Rotary matrix (RoPE)}} \underbrace{+ u^{(1)} * \mathbb{1}_{\{m=n\}} \boxed{+ u^{(2)} * \mathbb{1}_{\{m \neq n\}}}}_{\text{Binary attention bias}},$$

$$A_{ij,mn} = \frac{\exp\{E_{ij,mn}\}}{\sum_{k,o} \exp\{E_{ik,mo}\}},$$



MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

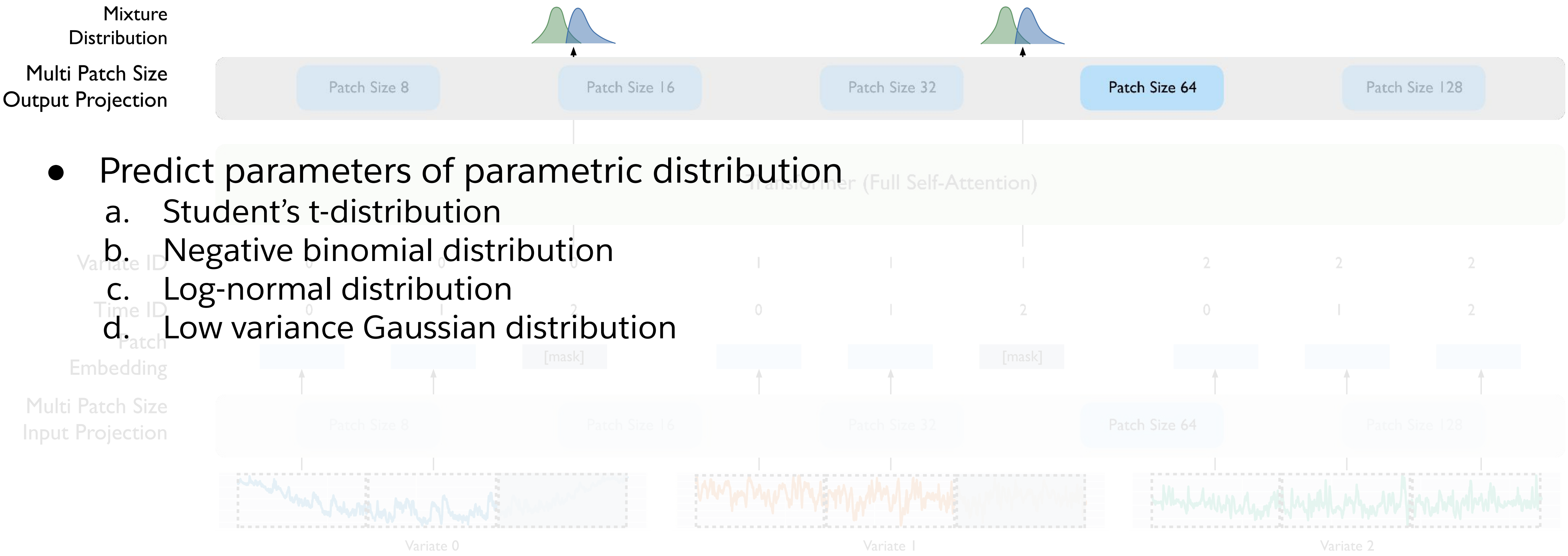
3) Mixture Distribution

MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

SMU SINGAPORE MANAGEMENT UNIVERSITY

salesforce

3) Mixture Distribution



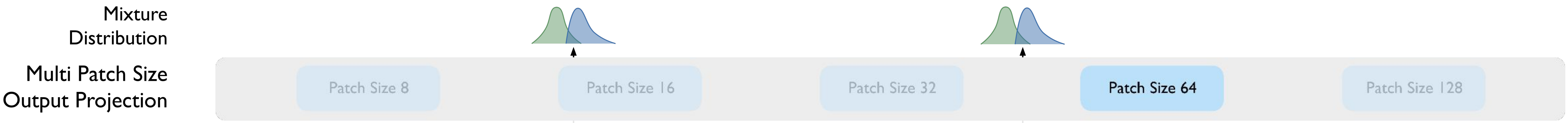
- Predict parameters of parametric distribution
    a. Student’s t-distribution
    b. Negative binomial distribution
    c. Log-normal distribution
    d. Low variance Gaussian distribution

MOIRAI: Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer

SMU SINGAPORE MANAGEMENT UNIVERSITY

salesforce

3) Mixture Distribution



- Predict parameters of parametric distribution
  a. Student’s t-distribution
  b. Negative binomial distribution
  c. Log-normal distribution
  d. Low variance Gaussian distribution

$$p(\mathbf{Y}_{t:t+h}|\hat{\phi}) = \sum_{i=1}^{c} w_i p_i(\mathbf{Y}_{t:t+h}|\hat{\phi}_i)$$

From the output projection layer

Pre-training Datasets for Time Series

Existing Work

- Comparison between prior work on pre-training for time series forecasting

	Any-variate (Zero-shot)	Probabilistic Forecasting	Flexible Distribution	Pre-training Data (Size)	Open-source
MOIRAI	✓	✓	✓	LOTSA (> 27B)	✓
TimeGPT-1	✓	✓	✗	Unknown (100B)	✗
ForecastPFN	✗	✗	-	Synthetic Data (60M)	✓
Lag-Llama	✗	✓	✗	Monash (< 1B)	✓
TimesFM	✗	✗	-	Wiki + Trends + Others (> 100B)	✓
TTM	✗	✗	-	Monash (< 1B)	✓
LLMTime	✗	✓	✓	Web-scale Text	✓

Large-scale Open Time Series Archive

Some key statistics

Table 2. Key statistics of LOTSA by domain.

	Energy	Transport	Climate	CloudOps	Web	Sales	Nature	Econ/Fin	Healthcare
# Datasets	30	23	6	3	3	6	5	23	6
# Obs.	16,358,600,896	4,900,453,419	4,188,011,890	1,518,268,292	428,082,373	197,984,339	28,547,647	24,919,596	1,594,281
%	59.17%	17.73%	15.15%	5.49%	1.55%	0.72%	0.09%	0.10%	0.01%

Table 3. Key statistics of LOTSA by frequency.

	Yearly	Quarterly	Monthly	Weekly	Daily	(Multi) Hourly	(Multi) Minute-level	(Multi) Second-level
# Datasets	4	5	10	7	21	31	25	2
# Obs.	873,297	2,312,027	11,040,648	18,481,871	709,017,118	19,875,993,973	7,013,949,430	14,794,369
%	0.003%	0.008%	0.040%	0.067%	2.565%	71.893%	25.370%	0.054%

Other Training Details

- Data distribution
  - Cap sampling % of extremely large datasets due to imbalance data
- Task distribution
  - Randomly sample context length, prediction length
  - Randomly subsample multivariate time series
  - Randomly combine aligned univariate time series into multivariate

Table 4. Details of MOIRAI model sizes.

	Layers	d_{model}	d_{ff}	**Heads**	d_{kv}	**Params**
$\text{MOIRAI}_{\text{Small}}$	6	384	1536	6	64	14m
$\text{MOIRAI}_{\text{Base}}$	12	768	3072	12	64	91m
$\text{MOIRAI}_{\text{Large}}$	24	1024	4096	16	64	311m

Experiments

In-distribution Forecasting

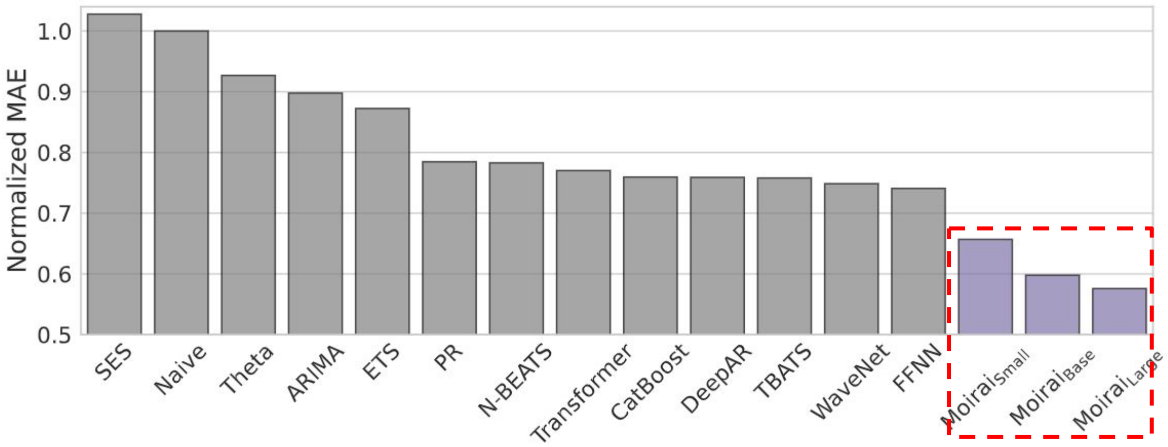


Fig: Aggregate results on the Monash TSF Benchmark.

- In-distribution on the Monash benchmark
    - Results from this figure are aggregated over 29 datasets
- Train region of these datasets are present in our pre-training dataset
- Test region is held-out for evaluation
- Moirai is a **single model**
- Baselines have 1 model per dataset

Experiments

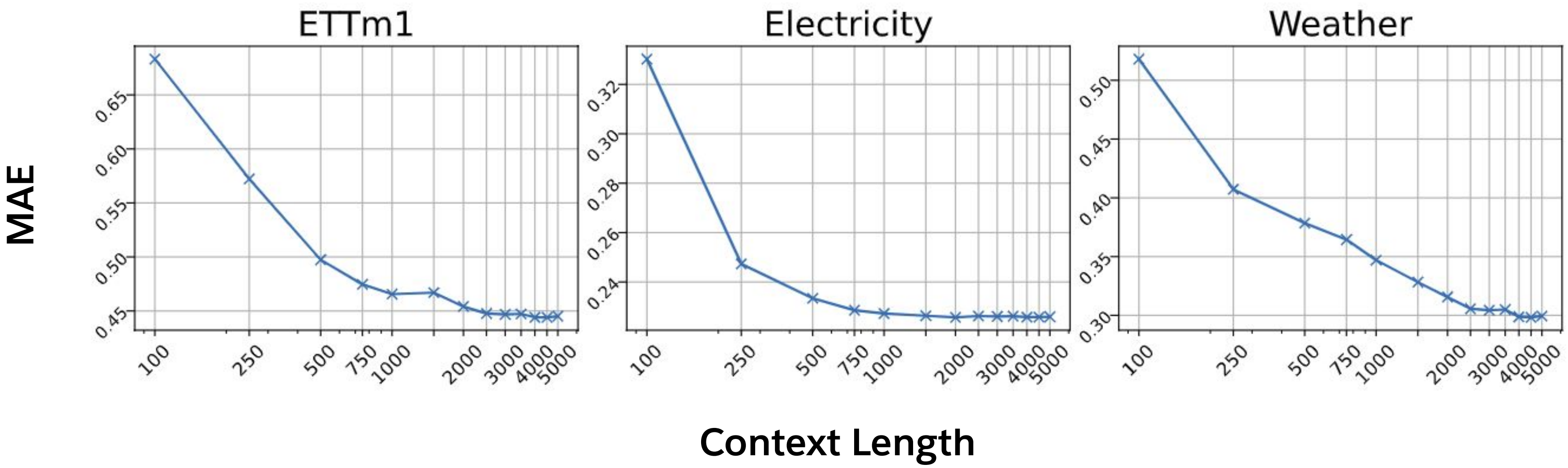
Out-of-distribution / Zero-shot forecasting

Table 5. Probabilistic forecasting results. Best results are highlighted in **bold**, and second best results are <u>underlined</u>. Baseline results are aggregated over five training runs with different seeds, reporting the mean and standard deviation.

		Zero-shot			Full-shot				Baseline	
		MOIRAI$_{\text{Small}}$	MOIRAI$_{\text{Base}}$	MOIRAI$_{\text{Large}}$	PatchTST	TiDE	TFT	DeepAR	AutoARIMA	Seasonal Naive
Electricity	**CRPS**	0.072	0.055	<u>0.050</u>	0.052±0.00	**0.048±0.00**	0.050±0.00	0.065±0.01	0.327	0.070
	MSIS	7.999	6.172	5.875	<u>5.744±0.12</u>	**5.672±0.08**	6.278±0.24	6.893±0.82	29.412	35.251
Solar	**CRPS**	0.471	<u>0.419</u>	**0.406**	0.518±0.09	0.420±0.00	0.446±0.03	0.431±0.01	1.055	0.512
	MSIS	8.425	<u>7.011</u>	**6.250**	8.447±1.59	13.754±0.32	8.057±3.51	11.181±0.67	25.849	48.130
Walmart	**CRPS**	0.103	0.093	0.098	<u>0.082±0.01</u>	**0.077±0.00**	0.087±0.00	0.121±0.00	0.124	0.151
	MSIS	9.371	8.421	8.520	**6.005±0.21**	<u>6.258±0.12</u>	8.718±0.10	12.502±0.03	9.888	49.458
Weather	**CRPS**	0.049	**0.041**	0.051	0.059±0.01	0.054±0.00	<u>0.043±0.00</u>	0.132±0.11	0.252	0.068
	MSIS	5.236	<u>5.136</u>	**4.962**	7.759±0.49	8.095±1.74	7.791±0.44	21.651±17.34	19.805	31.293
Istanbul Traffic	**CRPS**	0.173	0.116	0.112	0.112±0.00	0.110±0.01	<u>0.110±0.01</u>	**0.108±0.00**	0.589	0.257
	MSIS	5.937	4.461	4.277	**3.813±0.09**	4.752±0.17	<u>4.057±0.44</u>	4.094±0.31	16.317	45.473
Turkey Power	**CRPS**	0.048	0.040	**0.036**	0.054±0.01	0.046±0.01	<u>0.039±0.00</u>	0.066±0.02	0.116	0.085
	MSIS	7.127	<u>6.766</u>	**6.341**	8.978±0.51	8.579±0.52	7.943±0.31	13.520±1.17	14.863	36.256

Experiments

Analysis on context length



- Plot of performance (MAE) against context length
- Prediction length 96, patch size 32
- Increasing context length does not hurt performance

Conclusion

- Moirai
  - Modifications to the Transformer architecture for Universal Forecasting
    - Multi in/output patch size projections
    - Any-variate Attention mechanism
    - Mixture distribution predictions

https://github.com/SalesforceAIResearch/uni2ts

Conclusion

- Moirai
  - Modifications to the Transformer architecture for Universal Forecasting
    - Multi in/output patch size projections
    - Any-variate Attention mechanism
    - Mixture distribution predictions
- LOTSA data
  - Largest collection of open-data for pre-training time series forecasting models
  - 27B obs (231B including number of variates per time series)

https://github.com/SalesforceAIResearch/uni2ts

Conclusion

- Moirai
  - Modifications to the Transformer architecture for Universal Forecasting
    - Multi in/output patch size projections
    - Any-variate Attention mechanism
    - Mixture distribution predictions
- LOTSA data
  - Largest collection of open-data for pre-training time series forecasting models
  - 27B obs (231B including number of variates per time series)
- Limitations & Future work
  - Heuristic approach to tackling cross-frequency learning (multi patch size mapping)
  - Limited support for high-dimensional time series
  - LOTSA data - better diversity of domains and frequency
  - Multi-modality - Text + Time Series for cold-start problems or judgemental forecasting

https://github.com/SalesforceAIResearch/uni2ts

Conclusion

- Moirai
  - Modifications to the Transformer architecture for Universal Forecasting
    - Multi in/output patch size projections
    - Any-variate Attention mechanism
    - Mixture distribution predictions
- LOTSA data
  - Largest collection of open-data for pre-training time series forecasting models
  - 27B obs (231B including number of variates per time series)
- Limitations & Future work
  - Heuristic approach to tackling cross-frequency learning (multi patch size mapping)
  - Limited support for high-dimensional time series
  - LOTSA data - better diversity of domains and frequency
  - Multi-modality - Text + Time Series for cold-start problems or judgemental forecasting

Thank You!

Questions?

https://github.com/SalesforceAIResearch/uni2ts