



Seeing Structure in Time

Interpretable, generalizable, and causality-aware
time series learning

Tengfei Ma

Department of Biomedical Informatics
Stony Brook University

October 8, 2026 @ NJIT

Time series connect observations to decisions

Measurements ordered in time $\langle x_t, t \rangle$

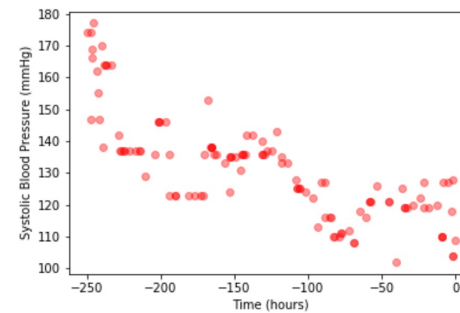
- Signals may be regular or irregular, univariate or multivariate.

Examples

- Healthcare: ECG, lab values, medical device data, hospital admission numbers
- Others: traffic, weather, power grid, particle trajectories...



ECG: continuous physiological signals



Blood pressure: sparse measurements

Tasks: Classification and diagnosis / Forecasting / Anomaly detection

Three gaps in time series models

Opaque predictions

People need to understand why a model predicts an outcome.

Interpretable shapes

Poor generalization

New hospitals, sensors, and machines provide little labeled data.

Reusable representations

Limited failure insight

Anomaly flags offer little warning or explanation.

Predictive and causal structure

Key idea: learn the right temporal abstractions (shapes, discrete latent regime, causal structure) to make model behavior easier to interpret and transfer

Roadmap

I

Shapes as tokens

IGN and VQShape

Interpretable and generalizable classification

II

Discrete latent regimes

SC-JEPA

Anticipating anomalies

III

Causal relationships

CAAD

Detecting structural failures

IV

Real-world applications

Healthcare, physics, battery

Extending pretrained models for various application

Meaningful representations for time series

Foundation models transfer learned representations across tasks and domains.

Representation What should count as a token in a continuous signal?

Structure How can the model preserve local shapes and relationships?

Learning objective Which pretraining task captures useful temporal dynamics?

This talk studies abstractions that remain useful across changing settings

PART I

Shape-based representations for generalizable and interpretable time series classification

IGN (ICLR 2025) and VQShape (NeurIPS 2024)

Shapelet features measure the best local match

A shapelet is a short subsequence that helps distinguish classes.

Distance between shapelet and time series:

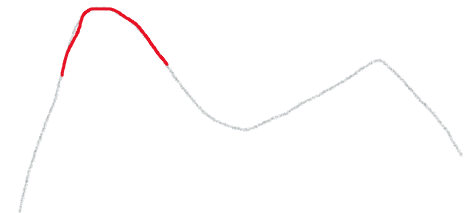
- Slide a shapelet across a series and keep the smallest matching distance.

$$d(\mathbf{x}, \mathbf{s}) = \min_t \left\| \mathbf{x}_{t \rightarrow t+L} - \mathbf{s} \right\|_2$$

Shapelet-based time series representation:

- One distance per shapelet gives the classifier its input:

$$\left[d(\mathbf{x}, \mathbf{s}_1), d(\mathbf{x}, \mathbf{s}_2), \dots, d(\mathbf{x}, \mathbf{s}_K) \right]$$

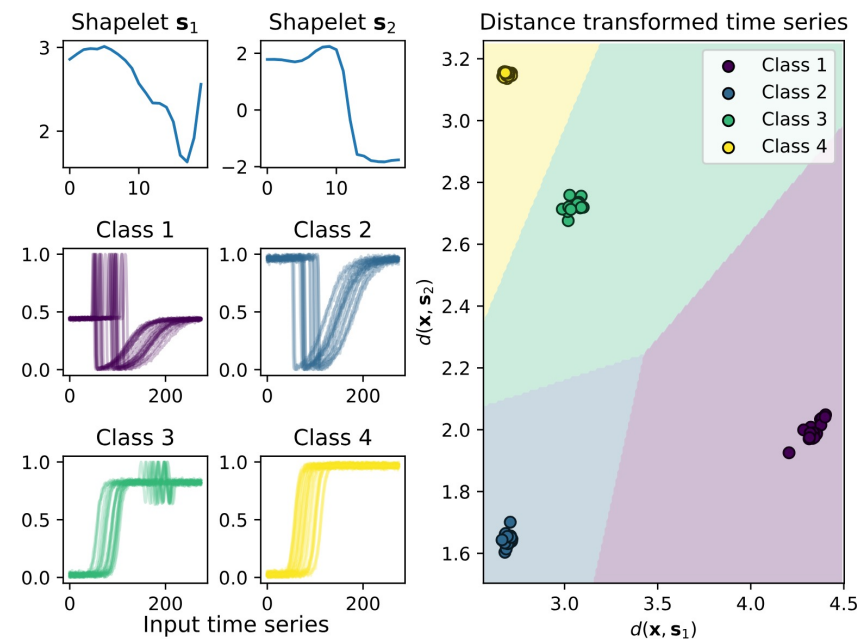


x: time series. s: shapelet. T and L: series and shapelet lengths. Smaller distance means a closer match.

Shapelets capture class-specific local patterns

A shapelet is a short subsequence that helps **distinguish classes**.

Matching a series to a set of shapelets gives an interpretable feature vector.



An example of how time series are transformed into linearly separable distances.

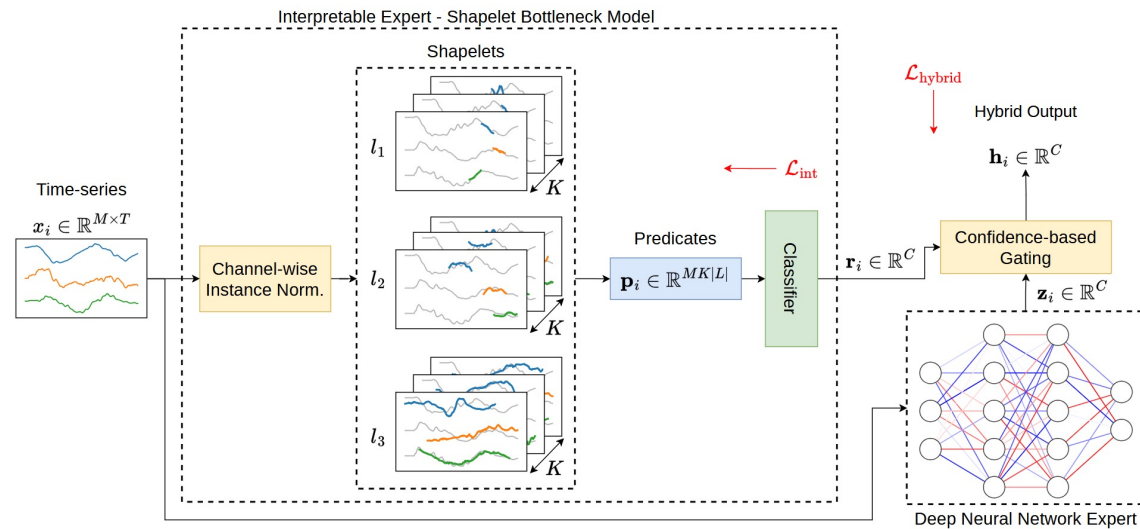
Shapelet example: tslearn documentation. The next slide defines the matching distance.

IGN combines interpretable patterns with a DNN

Shapelet limitation	IGN response
Expensive shapelet search	Learn shapelets with the model
Large distances obscure matches	Represent matches as probabilities
A shapelet bottleneck misses patterns	Mix a shapelet expert with a DNN

Interpretability Gated Networks

MOE architecture: a confidence gate combines a shapelet bottleneck model with a DNN expert.



IGN turns shapelet matches into class scores

Shapelet Bottleneck Model: the shapelet expert combines match probabilities using class-specific weights.

For sample i , class c ,

$$r_{i,c} = \sum_{\substack{k=1,\dots,K \\ m=1,\dots,M \\ l \in L}} w_{c,k}^{m,l} p_{i,k}^{m,l}$$

p : match probability

w : learned class weight

Each contribution can be traced to a temporal pattern in one channel at one scale.

i : sample. c : class. k : shapelet. m : channel. $l \in L$: subsequence length. ICLR 2025.

IGN trusts the shapelet expert when it is confident

Expert mixture

$$\mathbf{h}_i = \mathbf{r}_i \cdot \eta(x_i) + \mathbf{z}_i \cdot (1 - \eta(x_i))$$

$\mathbf{r}_i$: shapelet scores

$\mathbf{z}_i$: DNN scores

$\eta(x_i)$: shapelet weight

Confidence gate

$$\eta(x_i) = \frac{C \cdot \sum_{c=1}^C (\hat{r}_{i,c})^2 - 1}{C - 1}$$

Uniform prediction: $\eta = 0$

One-hot prediction: $\eta = 1$

$\hat{\mathbf{r}}$ is the shapelet expert's class-probability vector. C is the number of classes. ICLR 2025.

Clinical predictions with readable explanations

In-hospital mortality prediction

MIMIC-III lab-test time series

Learned shapelets act as predicates

Class	Illustrative interpretation
Survive	Stable heart rate, respiratory rate, or temperature
Mortality	Decreasing heart rate, systolic pressure, or mean blood pressure

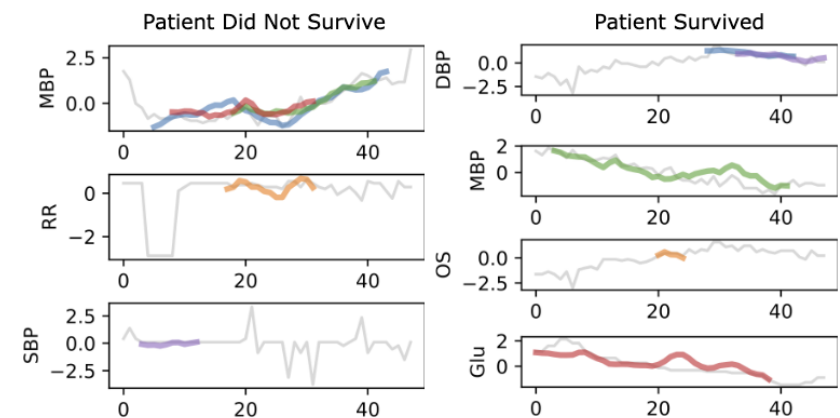
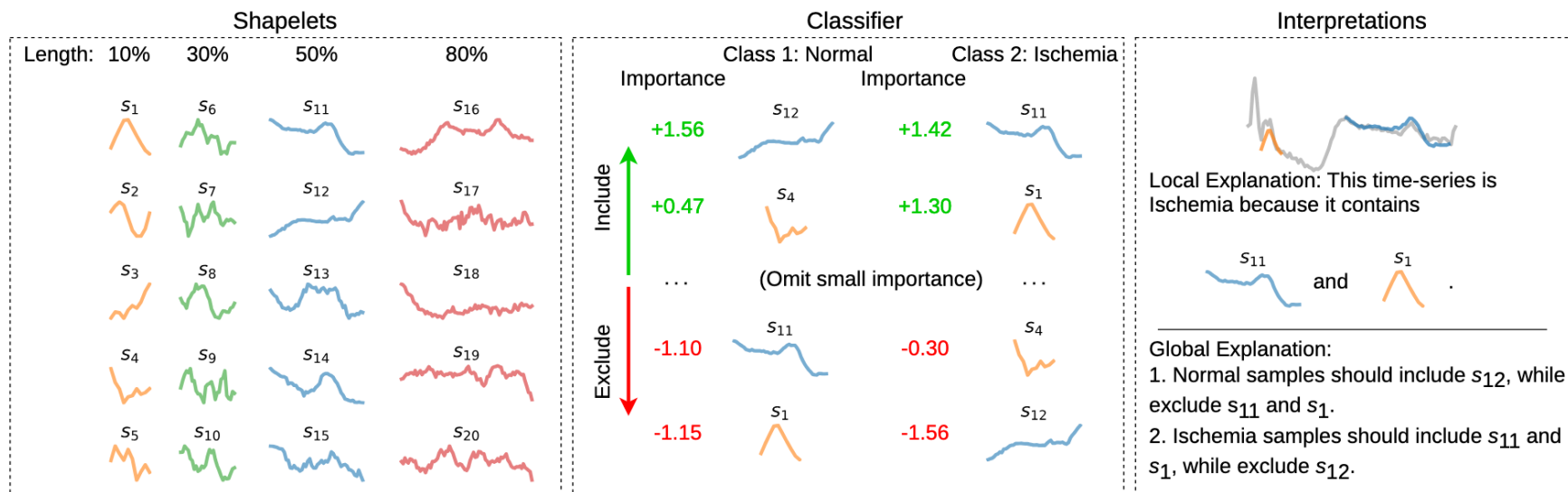


Figure 6: Local explanations of SCBM for in-hospital mortality predication. The variables in the figure are: Mean Blood Pressure (MBP), Respiratory Rate (RR), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Oxygen Saturation (OS), and Glucose (Glu).

More shapelet examples



However, limitations:

Sparse coverage

Fixed subsequence lengths

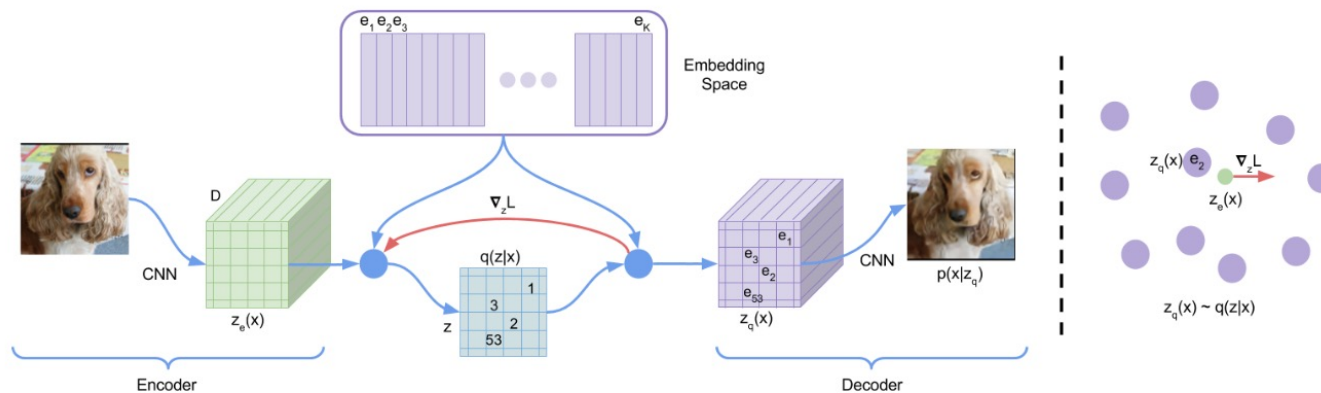
Limited transfer

VQShape extends the representation with a reusable codebook and explicit shape attributes.

Background: discrete neural representations

Vector quantization maps continuous embeddings to a learned codebook.

- **VQ-VAE: pretraining, reusable code**, enables better image/video generation/understanding, better alignment with LLMs.



Shapes as Tokens

- Codebooks in VQ-VAE
 - Still not physically meaningful
 - However, possible to replace the codebook with shapes
- Token representation
 - An attribute tuple is a shape with meta information
 - Shape can be reused, other meta info helps define its reconstruction

For a univariate TS x , a subsequence s_k can be represented by an attribute tuple $\tau_k = (z_k, \mu_k, \sigma_k, t_k, l_k)$ where

- $z_k \in \mathbb{R}^{d_{\text{code}}}$ is the code for abstracted shape of s_k ,
- $\mu_k \in \mathbb{R}^1$ is the offset of s_k ,
- $\sigma_k \in \mathbb{R}^1$ is the scale (standard deviation) of s_k and $\sigma_k > 0$,
- $t_k \in \mathbb{R}^1$ is the relative starting position of s_k in x and $0 \leq t \leq T - l_{\min}$,
- $l_k \in \mathbb{R}^1$ is the relative length of s_k w.r.t. to the length of x and $l_{\min} \leq l \leq T - t_k$.



VQShape: an interpretable time series foundation model

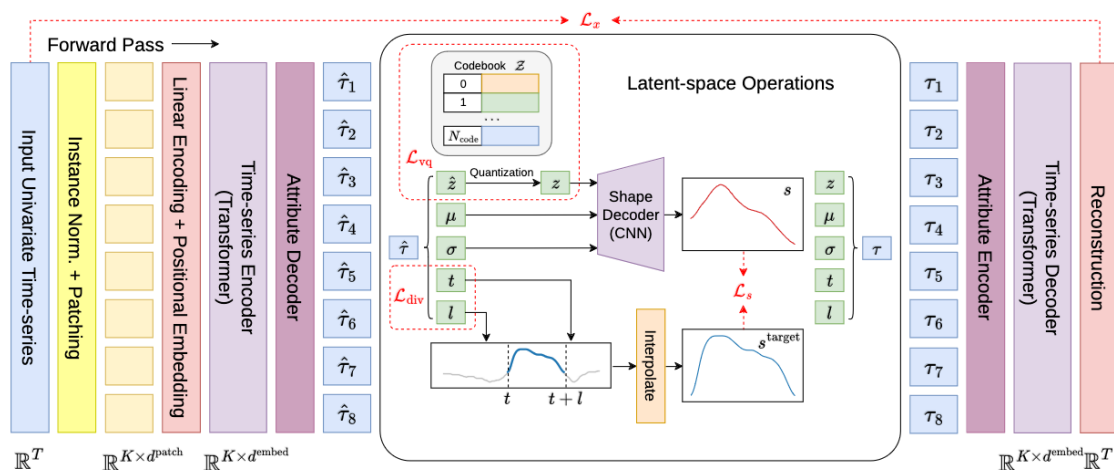


Figure 1: Overview of VQShape

Key Ideas:

- Shapes should be reused, but can be scaled in different time series
- Shape code can reconstruct a time series subsequence defined by the attribute tuple
- The combination of shapes can reconstruct original time series

VQShape decodes the shape attributes

PatchTST produces K embeddings h_k . Each embedding decodes one shape token.

$$\hat{\tau}_k = (\hat{z}_k, \mu_k, \sigma_k, t_k, l_k) = \mathcal{A}_{\text{dec}}(h_k), \quad \text{where} \quad \begin{cases} \hat{z}_k = f_z(h_k), \\ \mu_k = f_\mu(h_k), \\ \sigma_k = \text{softplus}(f_\sigma(h_k)), \\ t_k = \text{sigmoid}(f_t(h_k)) \cdot (1 - l_{\min}), \\ l_k = \text{sigmoid}(f_l(h_k)) \cdot (1 - t_k) + l_{\min}. \end{cases}$$

$$z_k = \arg \min_{z_q \in \mathcal{Z}} \|\hat{z}_k - z_q\|$$

Quantization selects the nearest codebook shape. Offset, scale, position, and duration remain continuous.

Softplus makes σ positive. Sigmoid bounds t and l so each token describes a valid subsequence.

VQShape trains the shape representation jointly

$$\mathcal{L}_{\text{pretrain}} = \lambda_x \mathcal{L}_x + \lambda_s \mathcal{L}_s + \lambda_{\text{vq}} \mathcal{L}_{\text{vq}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}}$$

$\mathcal{L}_x$ and $\mathcal{L}_s$ Reconstruct the full series and each local subsequence

$\mathcal{L}_{\text{vq}}$ Match embeddings to codes and encourage codebook coverage

$\mathcal{L}_{\text{div}}$ Spread shape tokens across positions and durations

The λ coefficients balance the objectives. Full quantization and attribute-diversity equations remain in backup.

VQShape learns confident, diverse code assignments

Quantization loss

$$\left\| \hat{z} - \text{sg}(z) \right\|_2^2 + \lambda_{\text{commit}} \left\| \text{sg}(\hat{z}) - z \right\|_2^2$$

The selected code and its continuous embedding learn to stay close.

Codebook coverage

$$\mathbb{E} \left[H(q(\hat{z}, \mathcal{Z})) \right] - H \left(\mathbb{E} [q(\hat{z}, \mathcal{Z})] \right)$$

Favor confident assignments while using a broad range of codes across the data.

sg: stop gradient. H: entropy. q: assignment distribution over codebook $\mathcal{Z}$. Both expressions are part of L_{vq} .

Decoded shapes support model inspection

Example shapes in time series

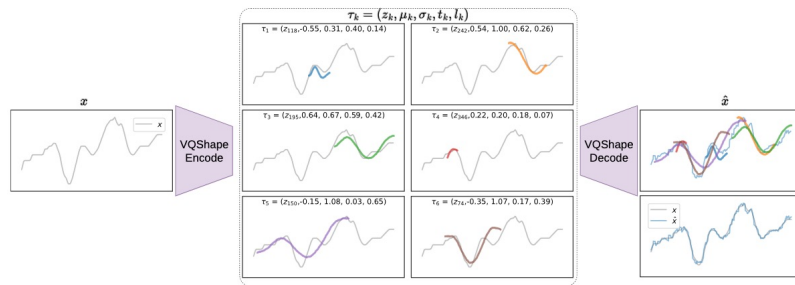
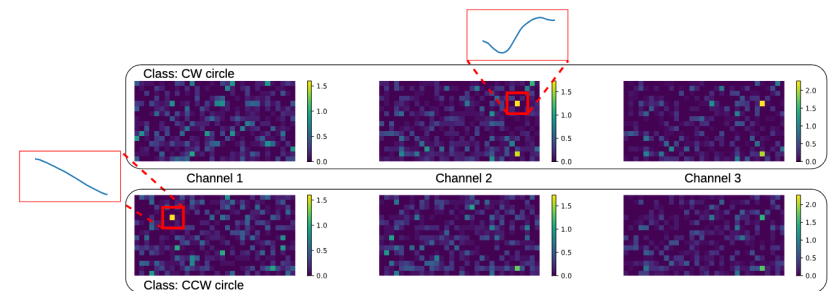


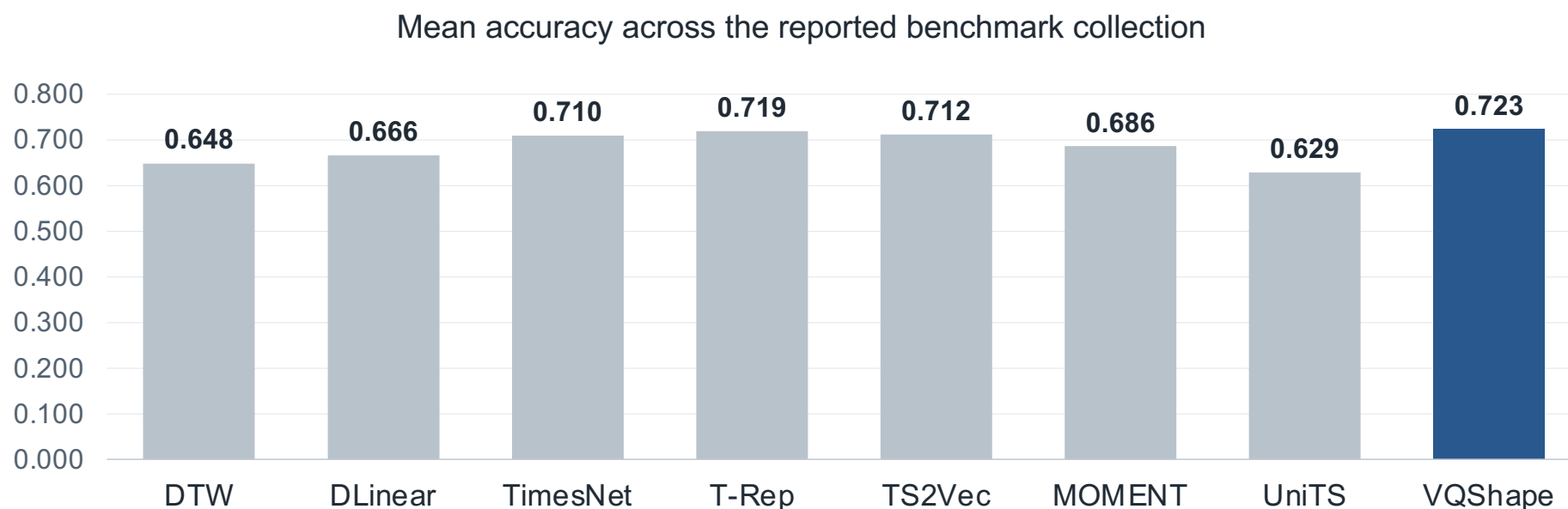
Figure 2: An example of abstracted shapes and their attributes extracted by VAShape. For better presentation, we visualize 6 of the 64 shapes.

Class-specific shape usage



Inspect both the decoded temporal patterns and where the model uses them

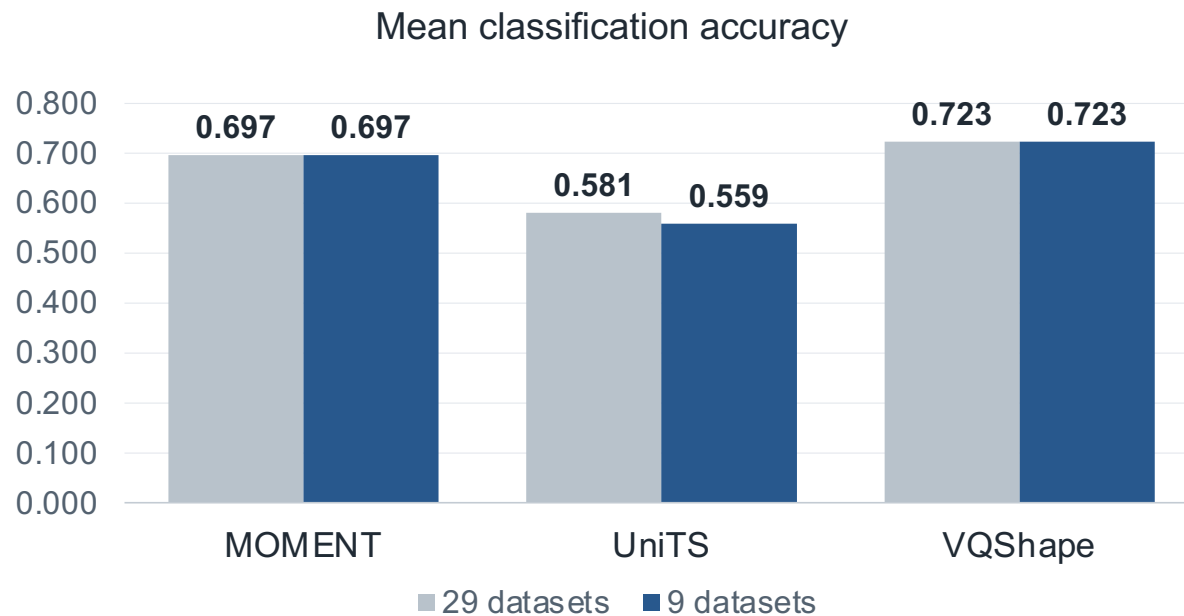
VQShape achieves strong mean classification accuracy



VQShape: 0.723 mean accuracy with interpretable, reusable shape codes

Selected methods from the original benchmark table, including statistics with N/A. Full table in backup.

Generalizability



Pretraining on only 9 datasets does not decrease the performance too much.

Transfer performance varies across healthcare tasks

Dataset	MOMENT	T-Rep	VQShape
AtrialFibrillation	0.200	0.354	0.520
Epilepsy	0.993	0.970	0.893
Heartbeat	0.722	0.725	0.663
SelfRegulationSCP1	0.840	0.819	0.904
SelfRegulationSCP2	0.478	0.591	0.596

Shape tokens transfer across tasks, with gains that depend on the dataset

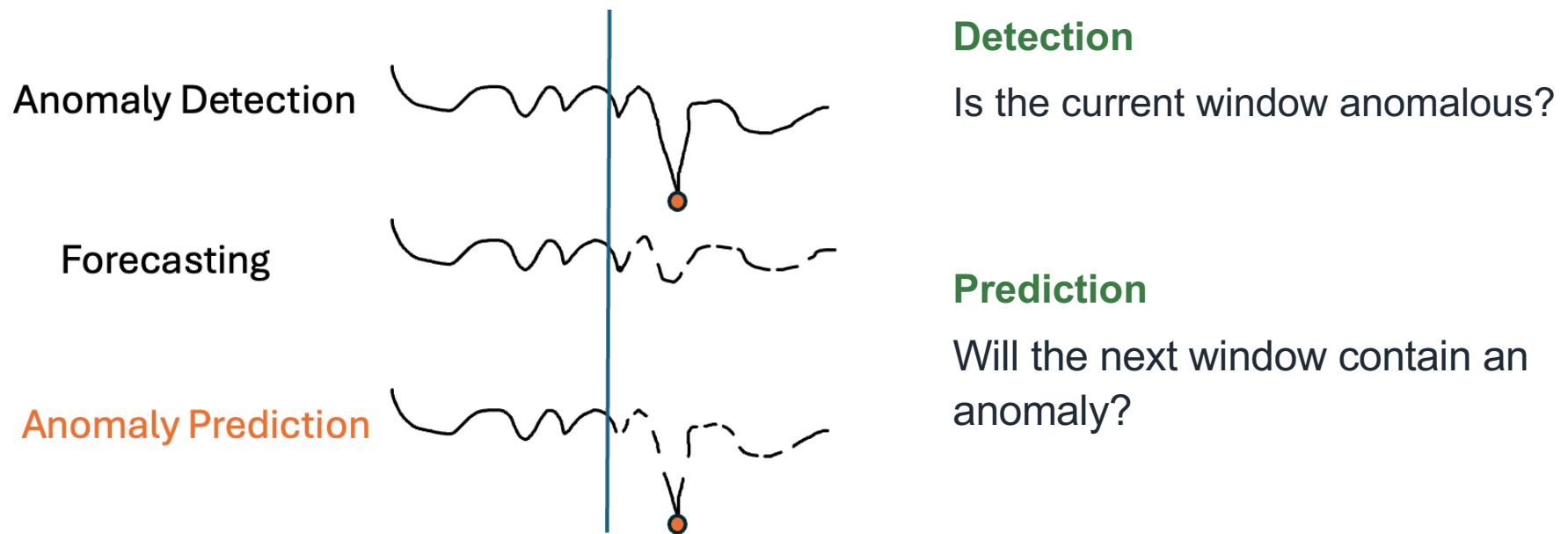
Accuracy on selected healthcare-related datasets and selected baselines. Full mixed-domain table in backup.

PART II

Anticipating anomalies with multi-resolution JEPA

SC-JEPA models future dynamics in a discrete latent code space

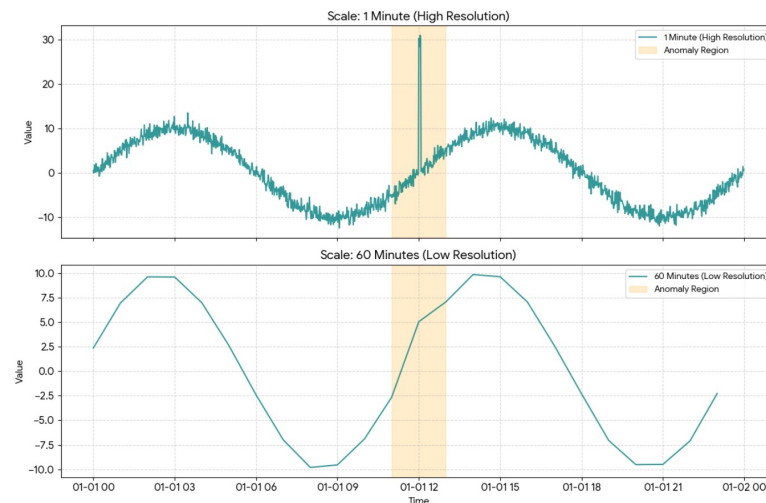
Prediction flags the next anomalous window



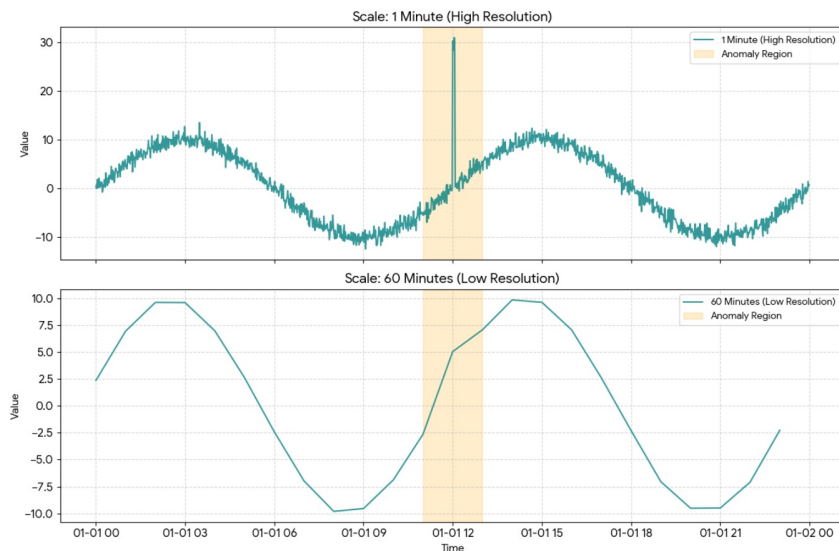
Protocol: non-overlapping windows of 100 steps. Use window t to predict the label of window $t + 1$.

Challenges

- Anomaly Prediction (precursor detection) is harder due to more subtle patterns
 - Requiring better modeling the dynamics
- Multi-scale
 - Time series data is often multi-scale. Anomalies can occur at different resolutions.



Solutions



Subtle precursors

JEPA predicts future representations

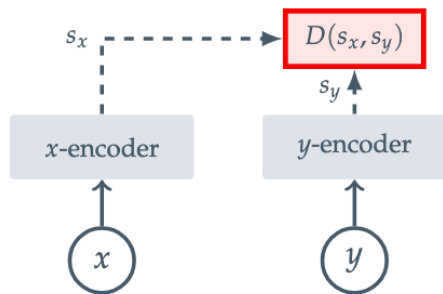
Multiple time scales

Fine and coarse views capture different changes

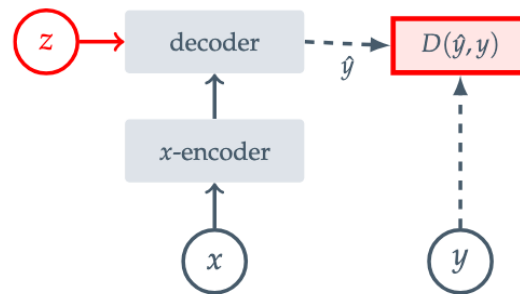
Unstable continuous latents in JEPA

A soft codebook organizes recurring regimes

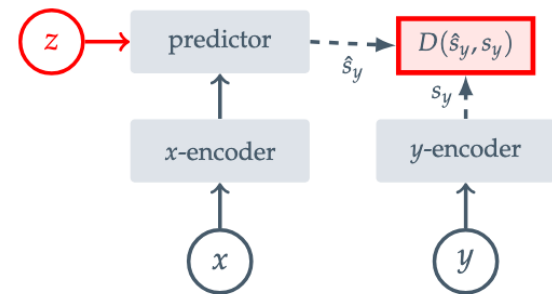
JEPA predicts representations of future observations



(a) **Joint-Embedding Architecture**



(b) **Generative Architecture**



(c) **Joint-Embedding Predictive Architecture**

Context encoder

Encode the current context

Latent predictor

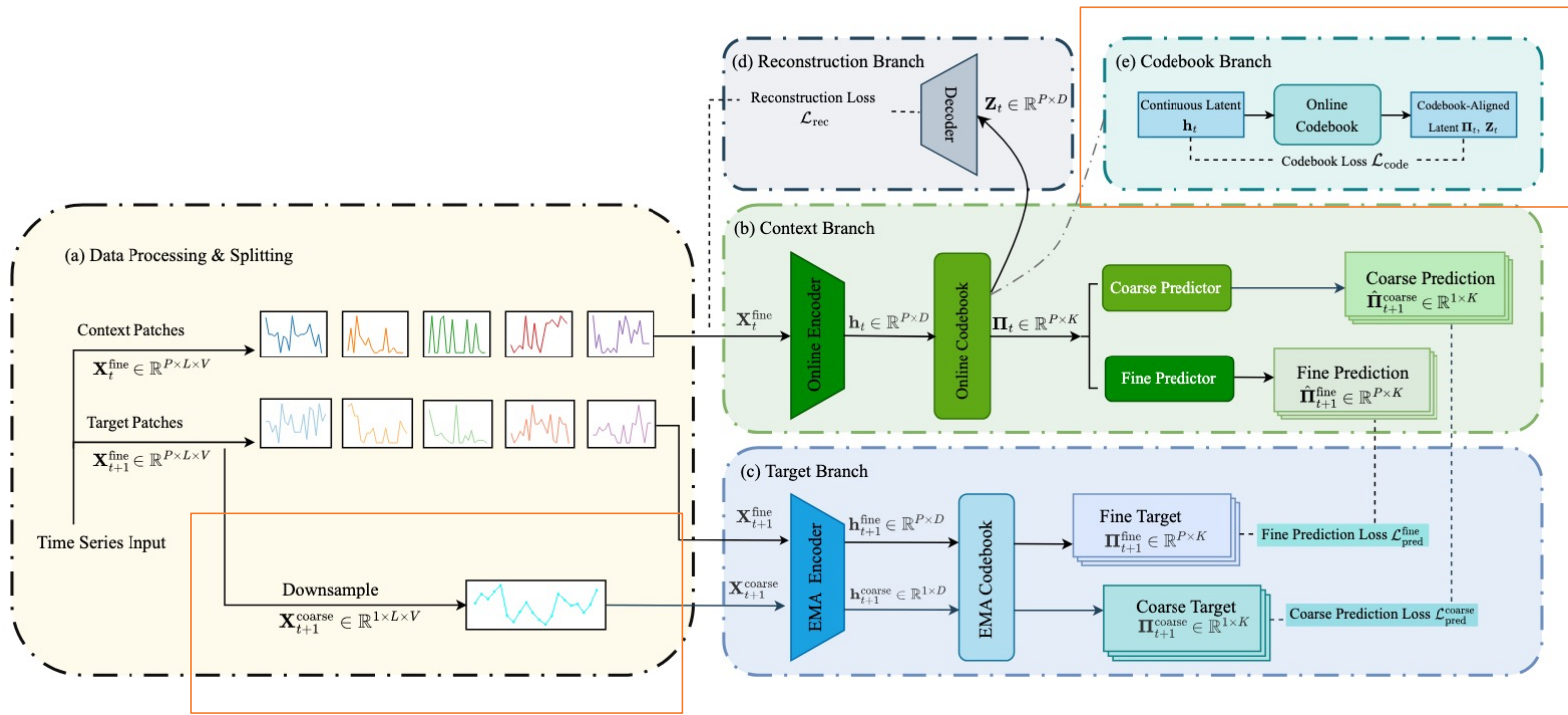
Predict the target representation

Target encoder

Provide the representation to learn

JEPA schematic adapted from the supplied Assran et al. (2023) figure.

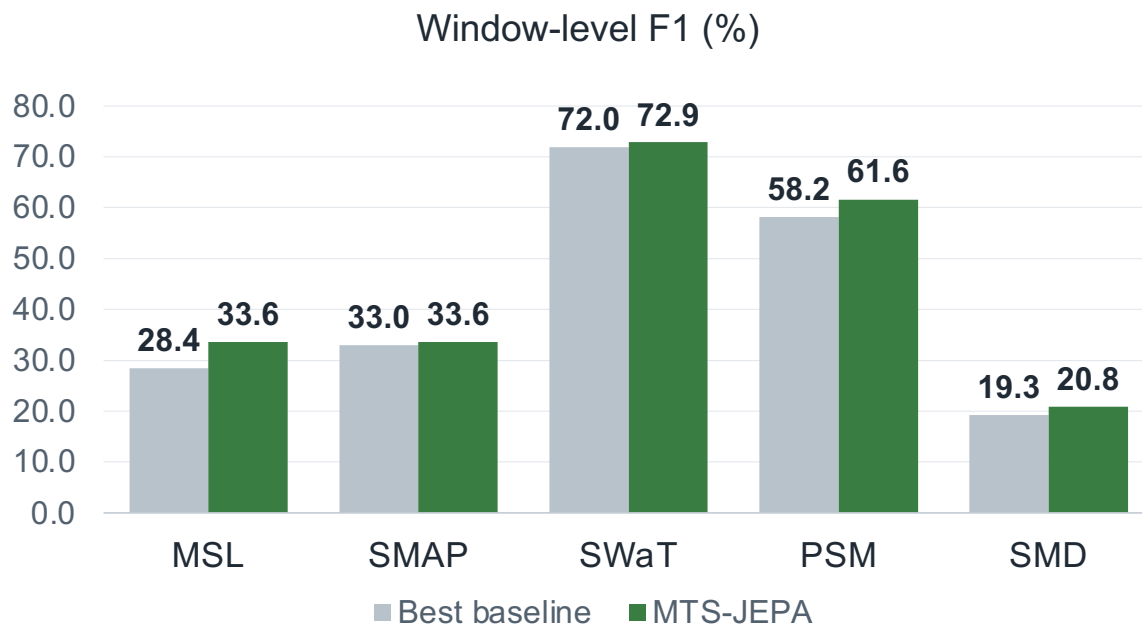
MTS-JEPA* learns fine and coarse latent dynamics



Context and target branches share codebook structure. Reconstruction and quantization stabilize prediction.

* Its formal publication in SDM2026 is renamed as SC-JEPA, i.e. we focus more on soft code than multi-scale later.

MTS-JEPA improves early-warning F1 on five benchmarks



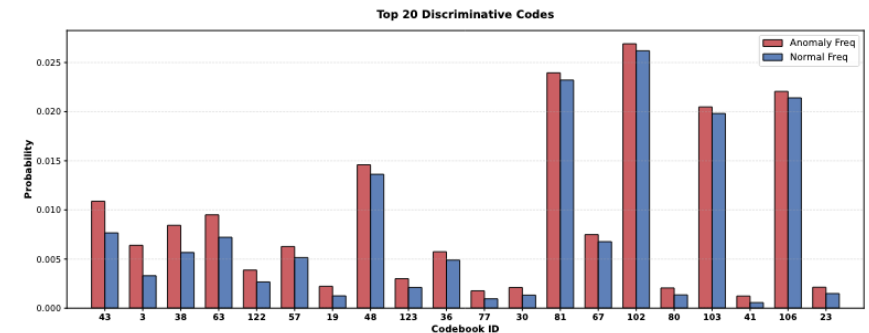
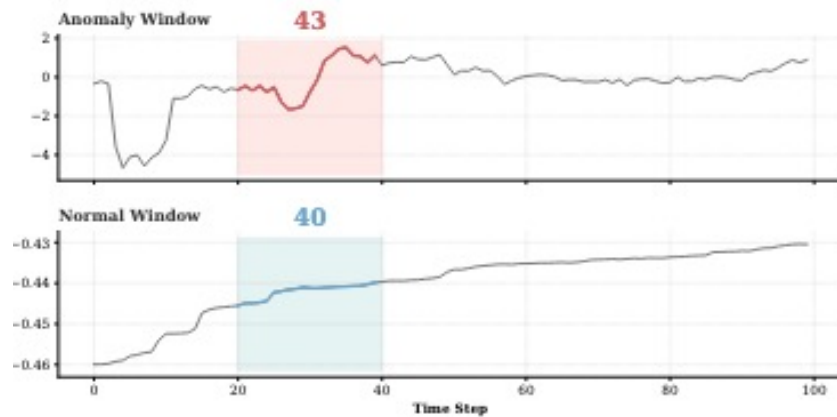
SWaT ablation: F1 (%)

Variant	F1
Full model	72.89
No downsampling	71.62
No reconstruction	14.77
No codebook	11.51

The codebook and reconstruction are central to predictive performance.

Best baseline per dataset among 14 methods. Mean of 5 seeds. The source also reports best AUC on all 5 datasets.

Latent code usage distinguishes anomalous windows

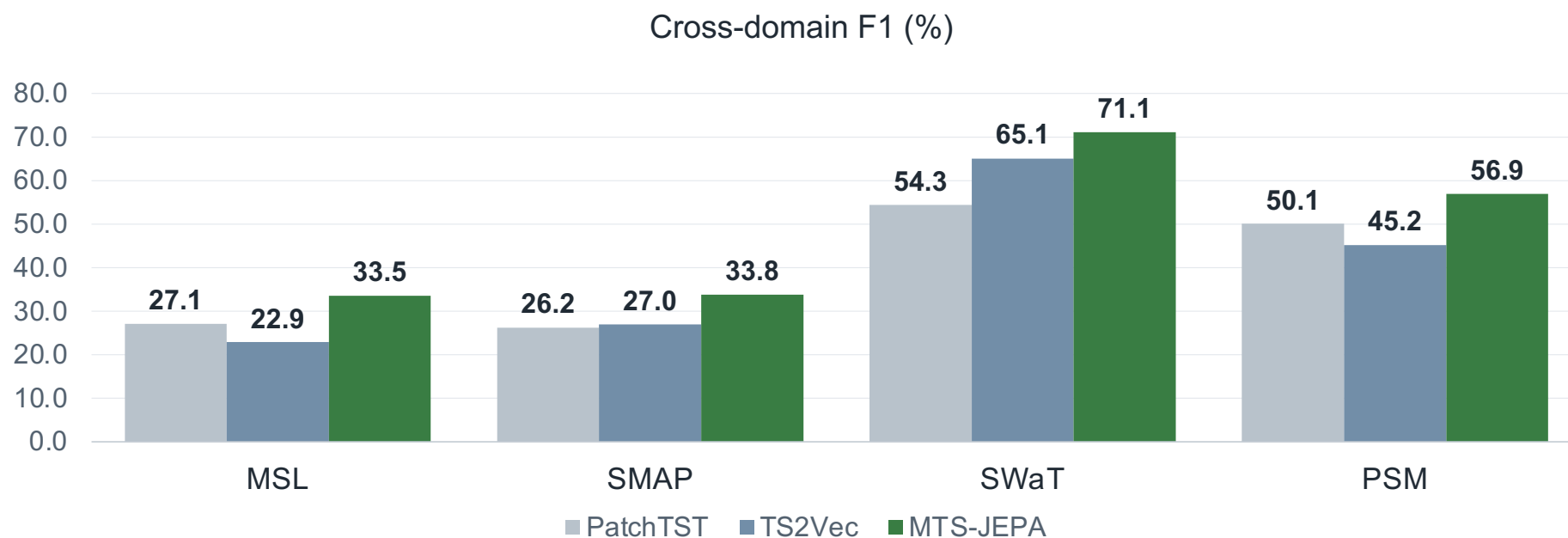


**PSM example: anomalous code 43
and normal code 40**

Code occurrence probabilities differ
between normal and anomalous windows.

The full two-pair example remains in backup. Code identities are specific to the learned codebook.

MTS-JEPA maintains stronger performance after transfer



The transferred MTS-JEPA model has the highest F1 in each displayed dataset

Mean over 5 runs. Full in-domain, cross-domain, F1, and AUC comparison in backup.

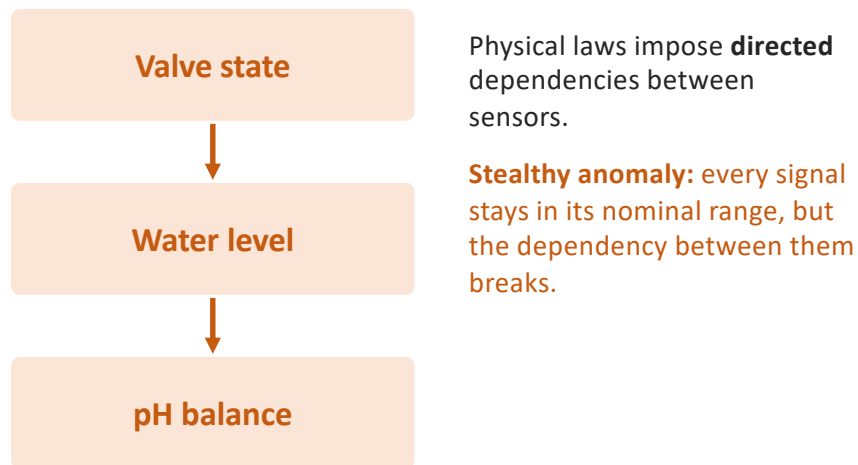
PART III

Causality-aware anomaly detection

CAAD (KDD 2026) checks which relationships in a system change

Anomalies can break dependencies without large spikes

Example: SWaT water-treatment testbed



Why existing detectors miss it

Magnitude-driven

Reconstruction / forecasting error flags spikes, not drifts in dependencies.

Correlation-centric

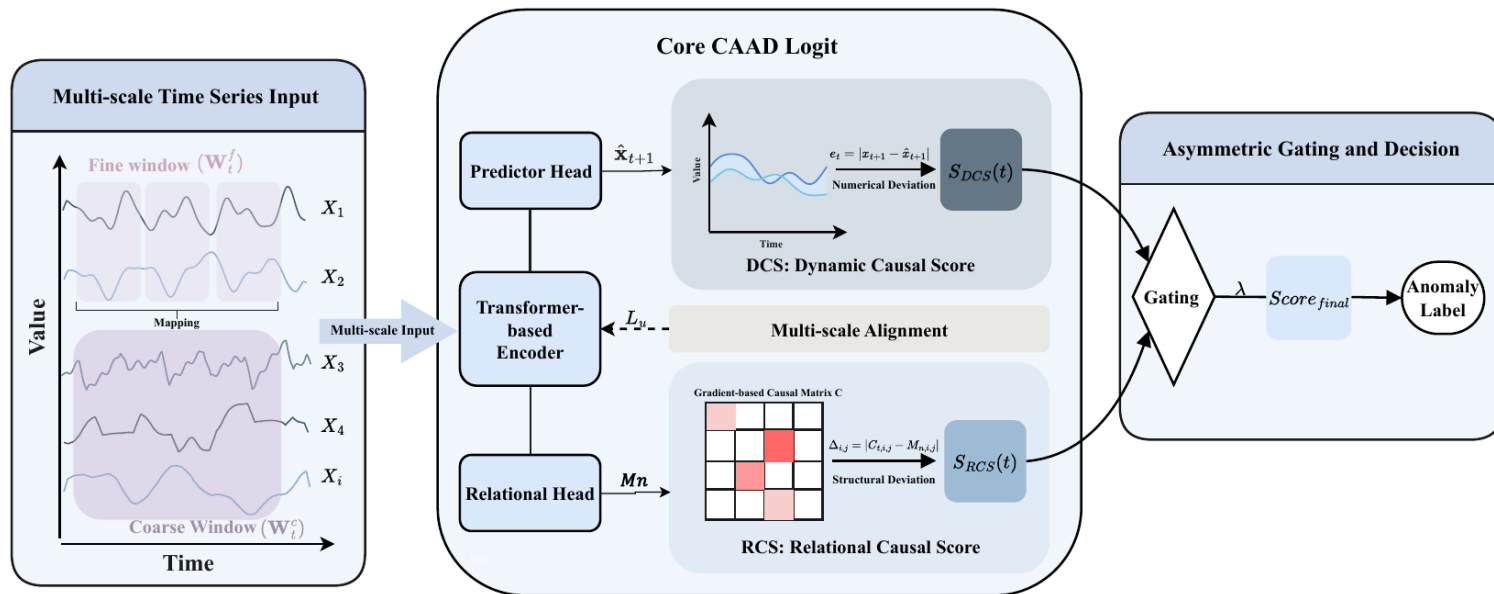
GNN / attention graphs encode co-movement; sensitive to noise.

Causality as an add-on

Static graphs, auxiliary regularizers, or post-hoc root-cause analysis.

CAAD continuously evaluates **Granger-causal consistency** for anomaly detection across windows and time scales.

CAAD combines dynamic and relational deviations



1

Multi-scale temporal alignment
learn scale-invariant causal dynamics

2

Dual-view deviation
dynamic (DCS) + relational (RCS) causal scores

3

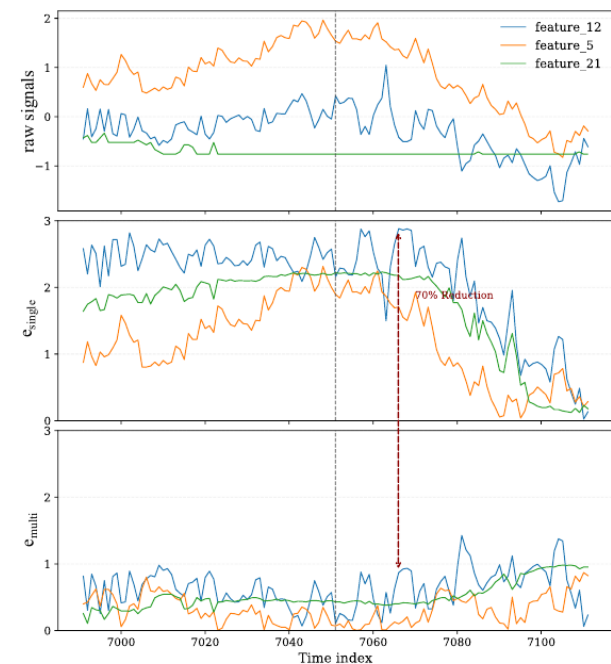
Asymmetric gating
RCS rectifies the residual-based score

Multi-scale alignment reduces spurious residuals

- **Fine windows** W^f keep local dynamics; **coarse windows** W^c (mean/max pooling) capture stable trends
- Two Transformer encoders for two scales.
- **Asymmetric (BYOL-style) alignment:** the fine “student” predicts the coarse “teacher” (stop-gradient), avoiding collapse without erasing scale differences

$$L = MSE_{fine} + MSE_{coarse} + \alpha \cdot L_u$$

- Residuals $e_t = |x_{t+1} - \hat{x}_{t+1}|$ act as **proxies for exogenous interventions**
- Multi-scale vs. single-scale: more leptokurtic, whiter residuals (Ljung–Box Q $\downarrow$ 47.9% on SWaT)



Raw signals / single-scale / multi-scale

CAAD scores temporal and structural consistency

Dynamic Causal Score (DCS)

How predictable is the evolution?

Robust-normalized residuals
Top 10% of variables

Relational Causal Score (RCS)

How much have the links changed?

Gradient-based Granger matrix
Deviation from the normal reference

$$G_{ij} = \sum_{\tau=1}^T \left| \frac{\partial \hat{x}_j}{\partial x_{i,\tau}} \right|$$

$$S_{\text{final}}(t) = S_{\text{DCS}}(t) + \lambda \cdot \text{ReLU}(S_{\text{RCS}}(t) - \tau)$$

RCS adds evidence above a threshold calibrated on normal windows.

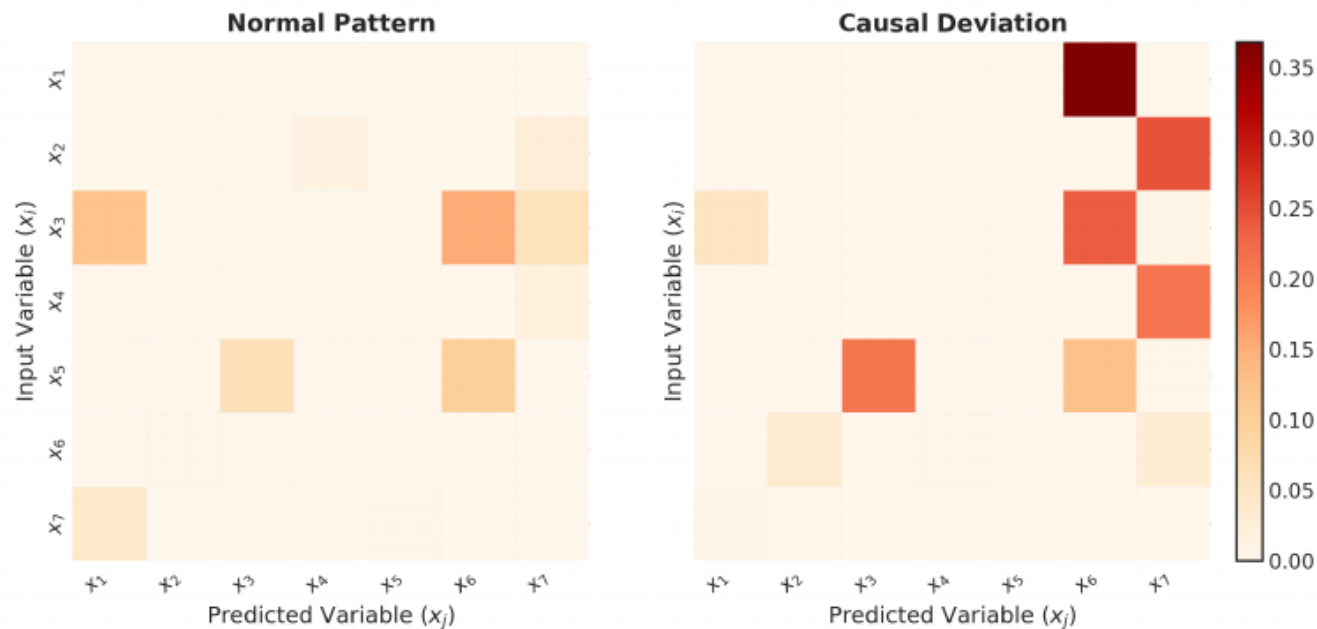
CAAD gains depend on the system's structure

Dataset	Baseline F1	CAAD F1	Baseline AUPRC	CAAD AUPRC
SWaT	0.324	0.952	0.575	0.974
PSM	0.436	0.362	0.456	0.602
CATSV2	0.262	0.330	0.265	0.321
SMD	0.191	0.170	0.202	0.090

Strongest gains on SWaT. Weakly coupled server metrics remain challenging.

Best baseline per metric among PCA, KNN, iTransformer, PatchTST, TranAD, VAE, and TimesNet. No point adjustment.

Relational deviations identify changed links

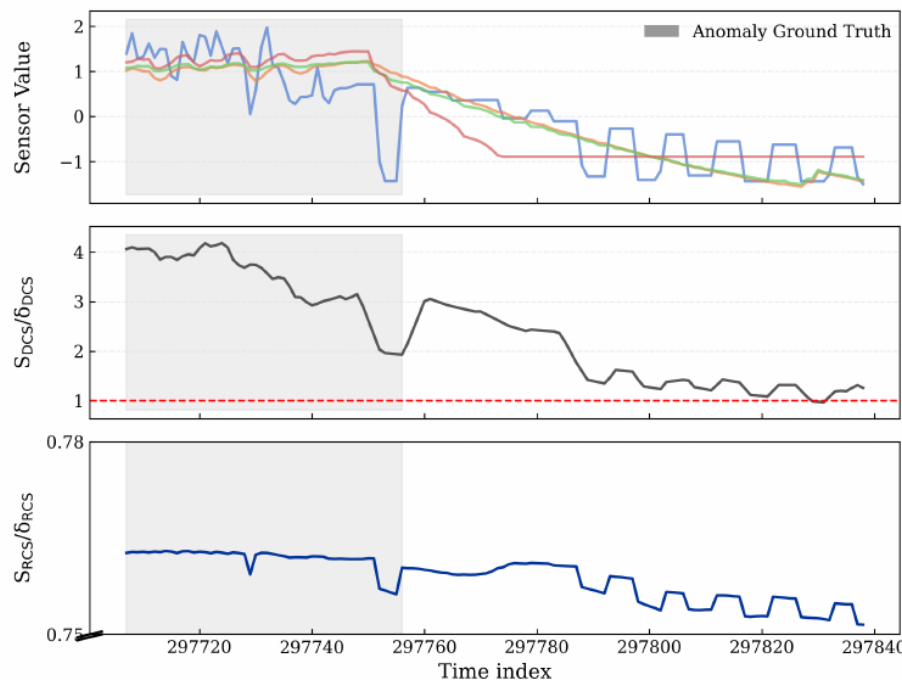


Dark cells indicate larger changes in particular input–prediction relationships.

Normal

Anomalous

Relational evidence improves the Exathlon precision result



Residuals remain elevated after the event ends.

The relational score remains flat.

0.68 precision

About 4× DCS alone

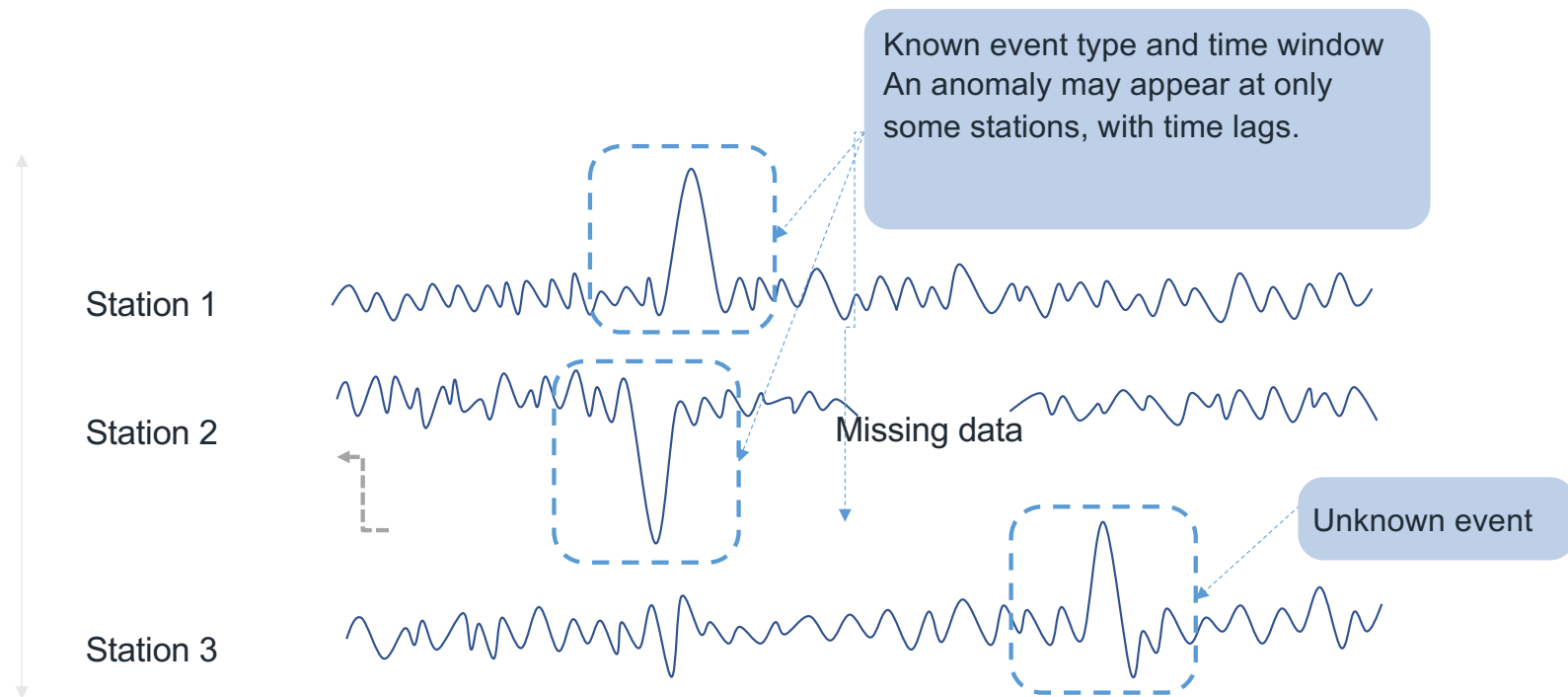
Exathlon example. Sensor signals (top), DCS (middle), RCS (bottom).

PART IV

Utilization of Pretrained Models for Real World Applications

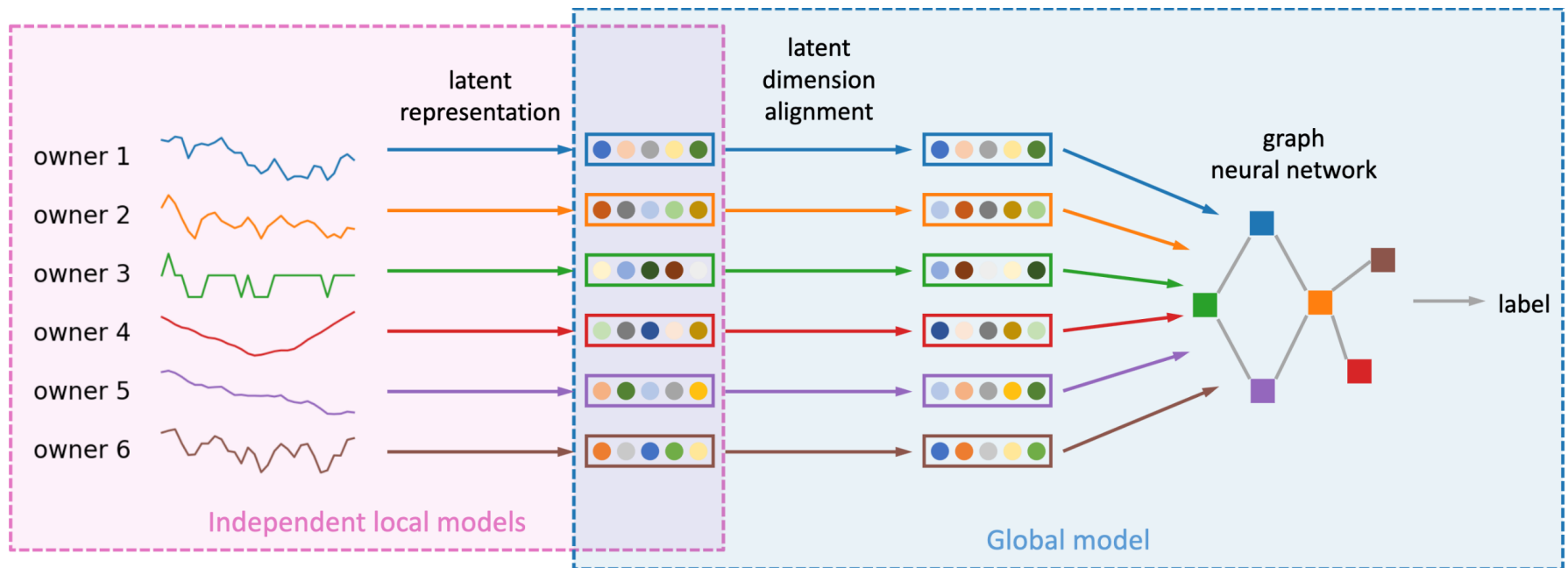
Tensor-TimesFM, battery prognostics, and detector monitoring

Power Grids: distributed observations and missing data



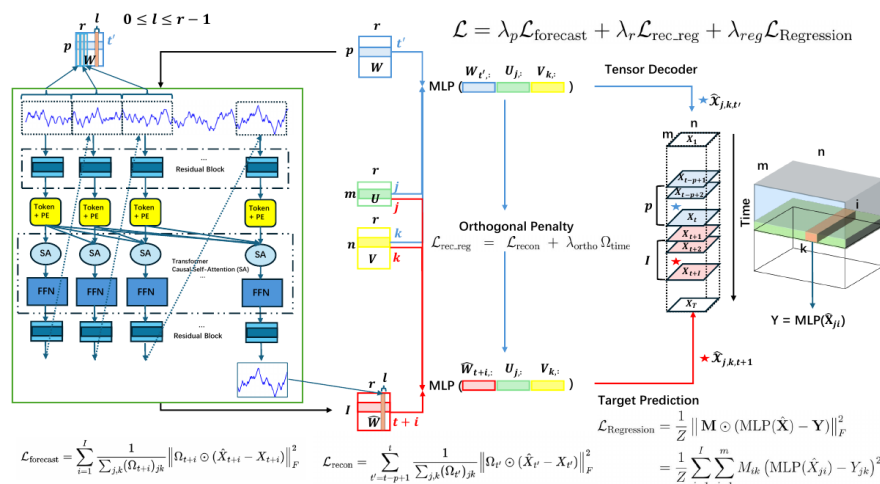
188 stations · 27 features per observation · individual feature values may be missing

Federated feature fusion (UAI23)



Tensor-TimesFM models sparse multivariate panels

Observed entries supervise a decoder with learned latent factors.



Fit only observed entries, without imputation.

Decorrelate time factors with an orthogonality penalty.

Use frozen TimesFM to forecast each latent time factor.

Latent factors improve sparse EPS forecasting

19%

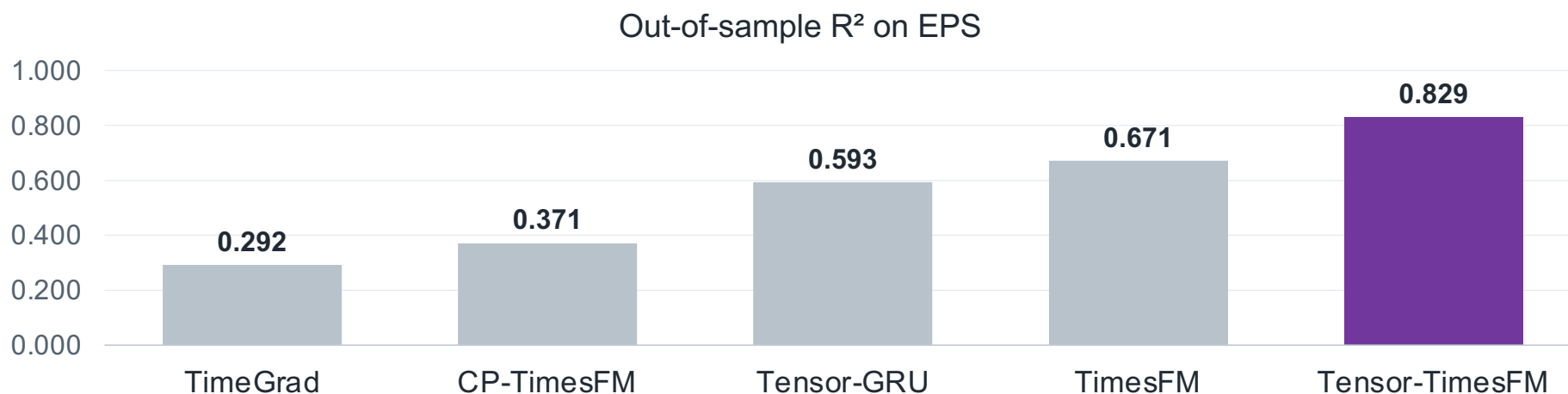
EPS cells
observed

68%

Walmart sales
zeros

41%

Rideshare cells
missing



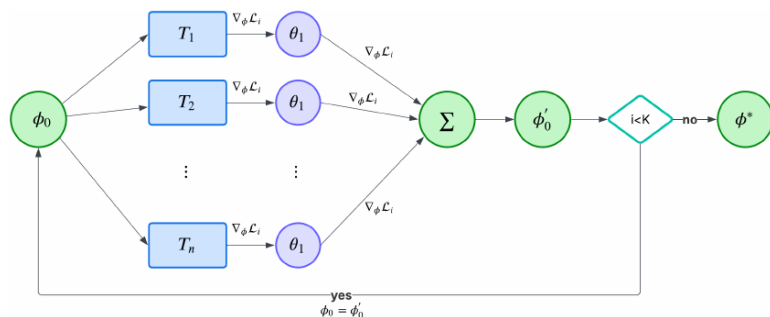
10 seeds. TimesFM alone: 0.671. Tensor-TimesFM: 0.829. The source's full architecture remains in backup.

Meta-learning improves battery life prediction

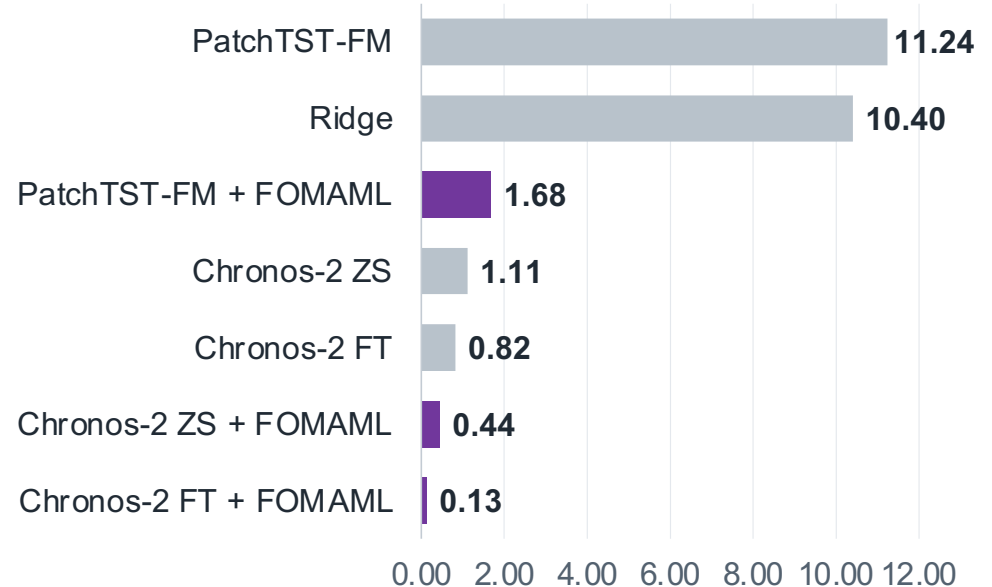
30 batteries, 9 operating conditions

Predict remaining useful life from the first 24 observed cycles.

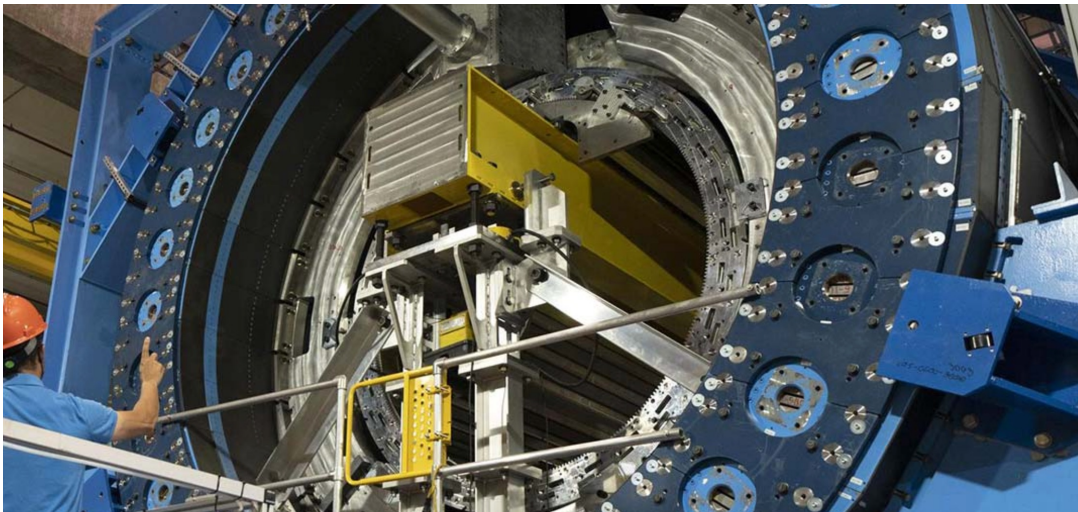
FOMAML learns an initialization that adapts to a new battery in 10 steps.



MAE in cycles (lower is better)

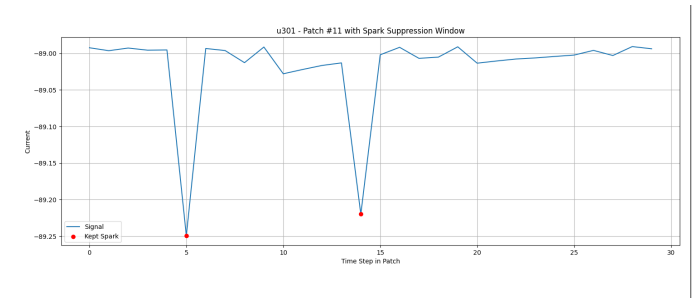


Predicting detector sparks



Multivariate current, voltage, and beam-condition streams sampled every 4 seconds.

sPHENIX at RHIC, Brookhaven.

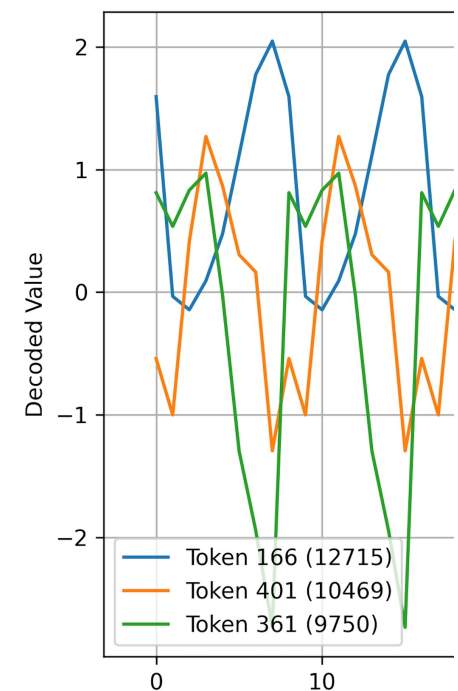


Goal: flag the precursor window before a significant current or voltage drop.

Pretrained shape tokens transfer to detector monitoring

Model	Precision	Recall	F1
Transformer	0.53	0.51	0.52
LightGBM	0.81	0.61	0.69
VQShape	0.77	0.80	0.78

A frozen VQShape representation with a linear probe reaches 0.76 F1.



Right: decoded shapes of the three most frequently activated tokens. Main VQShape model: 0.78 F1.

Seeing structure supports interpretation and transfer

Shapes as tokens

Readable predicates and reusable local patterns

Discrete latent regimes

Early warning from future latent dynamics

Causal relationships

Evidence of changed dependencies

Latent factors and pretrained priors

Adaptation to sparse panels and limited data

Looking Ahead

1

Causal root-cause diagnosis

From detecting broken relationships to localizing the source, using directed, real-time Granger matrices

2

Structure-aware foundation models

Shape tokens, discrete regimes and causal priors inside pretrained backbones; alignment with language models

3

Online & test-time adaptation

Models that keep learning under drift — non-stationary detectors, markets, patients

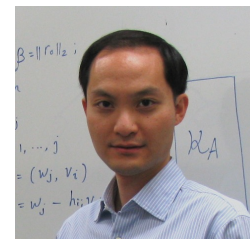
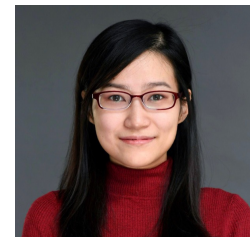
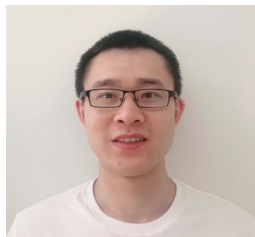
4

Deployment

Real-time monitoring for accelerator systems; bedside clinical time series

Thank you

Students and collaborators



SBU, RPI, IBM, UCSD, Yale, BNL, NJIT, NC State...

References

Yunshi Wen, Tengfei Ma, Ronny Luss, Debarun Bhattacharjya, Achille Fokoue, Anak Agung Julius. Shedding Light on Time Series Classification using Interpretability Gated Networks. ICLR, 2025.

Yunshi Wen, Tengfei Ma, Tsui-Wei Weng, Lam M. Nguyen, Anak Agung Julius. Abstracted Shapes as Tokens — A Generalizable and Interpretable Model for Time-series Classification. NeurIPS, 2024.

Yanan He, Yunshi Wen, Xin Wang, Tengfei Ma. SC-JEPA: Stabilizing Latent Predictive Learning for Time-Series Anomaly Prediction. SDM, 2026.

Xin Wang, Yunshi Wen, Yanan He, Haotian Xu, Youlan Zhao, Michel F. C. Haddad, Tengfei Ma. CAAD: Causality-Aware Multivariate Time Series Anomaly Detection via Multi-Scale Alignment and Structural Causal Consistency. KDD, 2026.

Tensor-TimesFM: High-dimensional Sparse Time Series Forecasting with Tensor Decoders and Latent Dynamics Prediction. Under review, 2026.

Xinyue Su, Bhavya Jain, Amay Bandodkar, Tengfei Ma. Foundation Time-Series Transformers for Lithium-Ion Battery Remaining Useful Life Estimation. MiLeTS Workshop @ KDD, 2026.

Tengfei Ma, Trong Nghia Hoang, Jie Chen. Federated Learning of Models Pre-Trained on Different Features with Consensus Graphs. UAI, 2023.