Reading the Device's DNA: Aligning Textual Specifications and Sensor Data for Degradation Prediction

Anonymous ACL submission

Abstract

Accurate degradation trajectory prediction (DTP) is crucial for ensuring industrial operational reliability. In real-world deployments, degradation is driven by heterogeneous stressors combining numerical sensor time-series and hardware specifications. However, existing DTP methods focus heavily on numerical signals, overlooking the rich semantic information carried by hardware specifications. This limits their cross-device generalization, as hardware specifications function like a device's "DNA", which providing stable, causally relevant semantic context that complements time-varying environmental signals. To bridge this gap, we propose *DegraMM*, a Multiway Transformer-based predictor designed to "read" the device's DNA through deep cross-modal alignment. *DegraMM* employs three specialized experts to model heterogeneous inputs and utilizes a shared multi-head attention mechanism to achieve a unified text-sensor representation geometry over modalities and time. Furthermore, selective modality gating is introduced to downweight noisy modalities under distribution shifts, while long-range temporal aggregation captures sustained degradation patterns. Evaluated on a large-scale real-world smart-meter dataset, *DegraMM* demonstrates significant improvements and superior robustness in cross-manufacturer transfer scenarios, showcasing the power of aligning textual knowledge with industrial sensor data for reliable prognostics.

1 Introduction

Degradation trajectory prediction (DTP) of industrial products has been a long topic critical for ensuring operational reliability and reducing lifecycle costs, such as industrial sensors, aviation electronics, medical instruments and meters (Huang et al., 2025). This paper uses smart meters (SMs) as a representative application for DTP. Since SMs operate under a wide range of climates and usage profiles, our proprietary dataset, which collected in harsh regions, captures this heterogeneity and enables rigorous evaluation of cross-domain generalization and robustness. Improving SM reliability directly reduces outages, maintenance costs, and customer disputes(Thunshirn et al., 2025). At the same time SMs combine multimodal data including long operational horizons and varied stressors, methods validated on SMs are likely to transfer to other long-lived industrial devices such as industrial sensors, avionics, and medical instruments.

Indeed, SMs are integral to modern power systems, supporting accurate billing, real-time monitoring, and grid stability, so their failures have immediate operational and economic consequences(Orlando et al., 2022). By the end of 2024, the number of smart meters is expected to exceed 1.2 billion globally (Li et al., 2024). Each designed for at least 16 years of service with mandatory recalibration typically every eight years (Ma et al., 2022). However, replacing SMs on a fixed schedule, without considering their actual condition, incurs substantial costs, including procurement for new meters, installation labor, and e-waste disposal. Shifting to a condition-based "replace-on-failure" strategy could save billions of RMB annually in China and even more around the world while significantly reducing plastic and metal waste (Zhao and Chen, 2023). Therefore, accurate DTP of SMs is not merely a technical task, but a key enabler of sustainable, cost-effective, and reliable grid operations.

A key difficulty for in-field DTP lies in the heterogeneous and multimodal nature of degradation drivers. In real deployments, degradation is influenced by multiple stressors that appear in different formats: external environmental factors are typically recorded as numerical time series, such as temperature and humidity, whereas device-intrinsic states can include semi-structured logs and textual descriptions such as hardware specifications,

configurations, and operational conditions. Most existing DTP approaches (Ma et al., 2023; Qiu and Tang, 2021; Duan and WenXuan, 2025; Tong et al., 2024; Qin et al., 2025) ignore the information store in hardware speciations while they function like a device's DNA—providing stable, causally relevant context that complements time-varying environmental and operational signals. Moreover, they often treat these heterogeneous inputs in a uniform manner (e.g., concatenation-based fusion) or ignore non-numerical factors altogether, resulting in shallow integration of stressors. As a result, models may over-rely on a small set of empirically dominant variables and fail to capture sustained degradation patterns and complex interactions among stressors, which limits robustness and generalization under distribution shifts.

To address these challenges, we propose *DegraMM*, a multimodal degradation trajectory predictor built on a Multiway Transformer with three specialized experts: one models external environmental stressors, one captures device-level states such as operational conditions, and the third encodes hardware specifications. A multiway attention mechanism enables joint fusion across modalities and time. Moreover, we incorporate selective modality gating to downweight noisy or weak modalities under distribution shifts, and employ long-range temporal aggregation to better capture sustained degradation patterns and complex stressor interactions.

Our key contributions are as follows:

1. We re-frame industrial degradation trajectory prediction (DTP) as a text-sensor multimodal learning challenge, highlighting the critical yet overlooked value of textual hardware specifications, the device's DNA, to overcome the shallow fusion limitations of traditional numerical-heavy methods.

2. We propose *DegraMM*, a Multiway Transformer-based framework featuring three modality-specific experts and a shared multi-head self-attention mechanism, which effectively reads the device's DNA by establishing a unified text-sensor representation geometry over modalities and time.

3. We introduce selective modality gating and long-range temporal aggregation to mitigate distribution shifts, and comprehensively validate *DegraMM* on a large-scale real-world dataset, demonstrating superior robustness and cross-manufacturer transferability in practical deployment scenarios.

2 Related Works

Data-driven DTP leverages in-field monitoring streams to learn degradation patterns directly from operational data. Early approaches include kernel methods and probabilistic fusion models, while deep learning methods (CNN/RNN/LSTM) have become dominant due to their ability to extract nonlinear temporal features (Xu et al., 2024; Guo et al., 2017; Zheng et al., 2017). However, RNN-style models often rely on relatively local temporal windows, which may be insufficient for slow, cumulative degradation processes. Transformer-based models address this limitation via self-attention, enabling long-range dependency modeling and parallel training; they have shown strong performance in RUL/DTP settings (Vaswani and Shazeer, 2017; Mo et al., 2021). In addition, advances in long-sequence time-series forecasting, such as attention mechanisms designed for long horizons and multi-scale temporal patterns, provide useful architectural building blocks for DTP, especially when degradation is driven by sustained stress exposure (Lim et al., 2021; Chen et al., 2025). However, existing methods do not effectively leverage the heterogeneous and multimodal nature of degradation stressors, which limits their performance. To address this, our paper aims to deeply align all relevant stressors—especially hardware specifications that prior work has largely overlooked—while preserving the long-range temporal modeling capability of Transformers.

3 SMs Degradation Trend Prediction Data

To systematically characterize the vulnerability of smart meters (SMs) under natural environmental stress and to provide a data for the selection and maintenance of smart meters, the State Grid Corporation of China has built a high-altitude testing base for on-site environmental verification of SMs.

As illustrated in Fig.1, the test base incorporates a highly representative cross-section of the smart-meter ecosystem: 1,369 meters from 121 manufacturers, covering the vast majority of leading vendors and prevalent meter types in the market. Following the designs in (Qin et al., 2025)

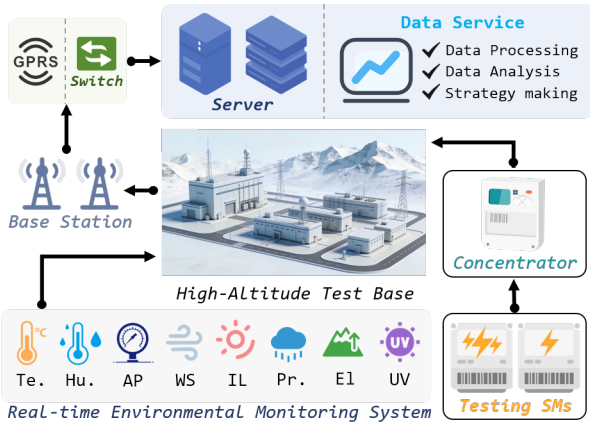

Figure 1: Electricity information acquisition platform for SMs. The abbreviations are defined as follows: Te., temperature; Hu., humidity; AP, air pressure; WS, wind speed; IL., illumination intensity; Pr., precipitation; El., elevation; UV, ultraviolet rays; GPRS, General Packet Radio Service.

and (Tang et al., 2012), an environmental monitoring subsystem collects real-time ambient data at 4-hour intervals, synchronized with SM operational telemetry. Eight types of environmental stresses were recorded, as shown in the figure caption of Fig.1. Seven discrete load levels are applied: $0.05P_s$, $0.1P_s$, $0.2P_s$, $0.5P_s$, P_s, $0.5P_{\max}$, and $P_{\max}$, where P_s denotes the power rating and $P_{\max} = 12P_s$. Five power-factor settings are configured: 1, 0.25L, 0.05L, 0.5C, and 0.8C, where 1 indicates a purely resistive load (unity power factor), L denotes inductive (lagging), and C denotes capacitive (leading). We record multiple performance-degradation indicators, including basic error, power-up/power-down error, and daily timekeeping error.

4 *DegraMM*: A Shared-Attention Multiway Transformer for Degradation Prediction

Smart-meter degradation is jointly governed by external operating conditions, internal device states, and hardware specifications, requiring a predictor that (i) unifies heterogeneous features into a coherent representation and (ii) captures their long-range temporal interactions. To this end, we propose *DegraMM*, a Multiway Transformer-based encoder with modality-specific experts. *DegraMM* encodes environmental stressors, operational conditions, and hardware specifications in dedicated expert branches, then fuses them through shared multiway self-attention to achieve deep cross-modal alignment and stronger generalization across meter batches. The expert design preserves complementary inductive biases for each modality while enabling global, joint feature fusion. In addition, selective modality gating downweights noisy or weak modalities under distribution shifts, and long-range temporal aggregation further strengthens modeling of sustained degradation patterns and complex stressor interactions.

4.1 Unified Multi-feature Fusion with Shared Multi-head Self-Attention

4.1.1 Tri-Path Feature System and Notation

As shown in Fig.2, we factor the inputs into *external*, *internal*, and *hardware* paths. For a window of length τ starting at time t,

$$\mathbf{X}_{\text{ext}}^{t:t+\tau-1} \in \mathbb{R}^{\tau \times d_{\text{ext}}}, \mathbf{X}_{\text{int}}^{t:t+\tau-1} \in \mathbb{R}^{\tau \times d_{\text{int}}}, \quad (1)$$

where the external stress features comprise the environmental and operating drivers (e.g., the eight stresses shown in Fig. 1), and the internal features comprise device responses and indicators. In addition, each device provides *hardware specifications* $\mathcal{S}_{\text{hw}}$, which are textual contents describing hardware-relevant attributes (e.g., model identifier, manufacturer notes, packaging/process hints). We treat $\mathcal{S}_{\text{hw}}$ as a third modality and include it as an additional input to the encoder to help learn hardware-specific attributes and improve alignment among all relevant stressors.

4.1.2 Modality-Specific Embeddings

We process the three modalities with separate embedding stacks that project them into a common latent space of width d_{model}. External and internal numeric embeddings

$$\mathbf{H}_{\text{ext}} = \text{LN}(\text{Linear}_{\text{ext}}(\mathbf{X}_{\text{ext}})) \in \mathbb{R}^{\tau \times d_{\text{model}}}, \quad (2)$$

$$\mathbf{H}_{\text{int}} = \text{LN}(\text{Linear}_{\text{int}}(\mathbf{X}_{\text{int}})) \in \mathbb{R}^{\tau \times d_{\text{model}}}. \quad (3)$$

Hardware text embedding We encode the textual hardware specifications $\mathcal{S}_{\text{hw}}$ using a text encoder $\text{TextEnc}(\cdot)$ and a projection into the shared width:

$$\mathbf{z}_{\text{hw}} = \text{TextEnc}(\mathcal{S}_{\text{hw}}) \in \mathbb{R}^{d_{\text{text}}}, \quad (4)$$

$$\mathbf{e}_{\text{hw}} = \text{LN}(\mathbf{z}_{\text{hw}}\mathbf{W}_{\text{hw}} + \mathbf{b}_{\text{hw}}) \in \mathbb{R}^{d_{\text{model}}}, \quad (5)$$

where $\mathbf{W}_{\text{hw}} \in \mathbb{R}^{d_{\text{text}} \times d_{\text{model}}}$. Because $\mathcal{S}_{\text{hw}}$ is typically time-invariant over the window, we broadcast $\mathbf{e}_{\text{hw}}$ across τ steps to form a per-timestep hardware stream:

$$\mathbf{H}_{\text{hw}} = \mathbf{1}_{\tau}\mathbf{e}_{\text{hw}}^{\top} \in \mathbb{R}^{\tau \times d_{\text{model}}}. \quad (6)$$

When $\mathbf{X}_{\text{hw}}$ is non-temporal (e.g., a single pooled vector), we broadcast it across the τ time steps before fusion. If hardware text is unavailable, we use a learned $[\text{UNK}_{\text{hw}}]$ embedding in place of $\mathbf{e}_{\text{hw}}$.

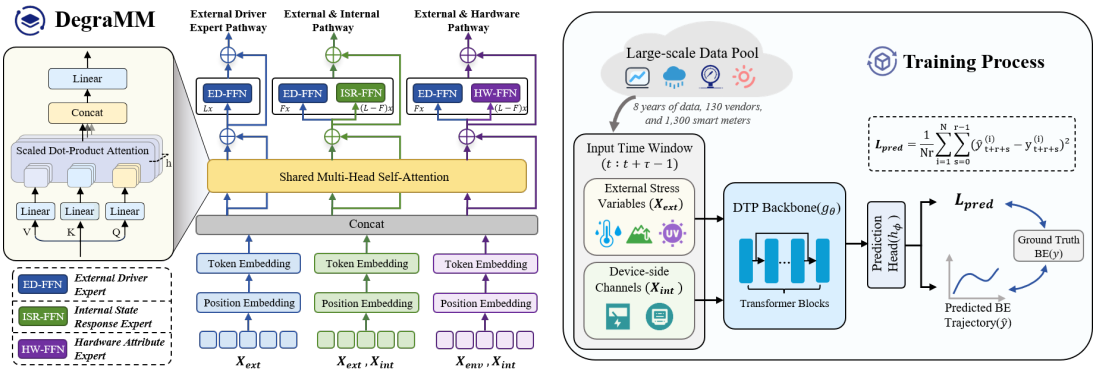

Figure 2: The Architecture of *DegraMM* and training process.

4.1.3 Temporal and Segment Encodings

To preserve temporal order and modality identity, we enrich each modality stream with (i) *learnable absolute positional encodings* and (ii) *segment (modality) embeddings*.

Positional encodings For modality $m \in \{\text{ext}, \text{int}, \text{hw}\}$, define a learnable matrix

$$\mathbf{P}_m = \left[\mathbf{p}_0^m; \ldots; \mathbf{p}_{\tau-1}^m\right] \in \mathbb{R}^{\tau \times d_{\text{model}}}, \tag{7}$$

where $\mathbf{p}_k^m \in \mathbb{R}^{d_{\text{model}}}$ is the absolute position vector for time index k.

Segment (modality) embeddings Each modality has a learnable embedding $\mathbf{s}_m \in \mathbb{R}^{d_{\text{model}}}$ that is broadcast over time via $\mathbf{1}_\tau \mathbf{s}_m^\top \in \mathbb{R}^{\tau \times d_{\text{model}}}$.

Augmented token sequences Given modality specific embeddings $\mathbf{H}_{\text{ext}}, \mathbf{H}_{\text{int}}, \mathbf{H}_{\text{hw}} \in \mathbb{R}^{\tau \times d_{\text{model}}}$, the temporally and modally augmented representations are

$$\widehat{\mathbf{H}}_{\text{ext}} = \mathbf{H}_{\text{ext}} + \mathbf{P}_{\text{ext}} + \mathbf{1}_\tau \mathbf{s}_{\text{ext}}^\top, \tag{8}$$

$$\widehat{\mathbf{H}}_{\text{int}} = \mathbf{H}_{\text{int}} + \mathbf{P}_{\text{int}} + \mathbf{1}_\tau \mathbf{s}_{\text{int}}^\top, \tag{9}$$

$$\widehat{\mathbf{H}}_{\text{hw}} = \mathbf{H}_{\text{hw}} + \mathbf{P}_{\text{hw}} + \mathbf{1}_\tau \mathbf{s}_{\text{hw}}^\top. \tag{10}$$

4.1.4 Unified Fusion for Shared Attention

We fuse the three modalities by feature-wise concatenation at each time index,

$$\widetilde{\mathbf{H}} = \text{Concat}\big(\widehat{\mathbf{H}}_{\text{ext}}, \widehat{\mathbf{H}}_{\text{int}}, \widehat{\mathbf{H}}_{\text{hw}}\big). \tag{11}$$

Here, $\widetilde{\mathbf{H}} \in \mathbb{R}^{\tau \times 3d_{\text{model}}}$, since the three input representations are concatenated along the feature dimension. This operation aligns modalities per timestep and prepares them for shared multi-head self-attention. It places external and internal signals from the same time index in a unified space while conditioning every step on the hardware context, enabling the model to learn hardware-aware cross-modal interactions.

4.1.5 Shared Multi-Head Self-Attention (MSA) Fusion

Each Transformer block applies *shared* MSA to $\widetilde{\mathbf{H}}$ followed by modality-specific feed-forward experts. For head $i = 1, \ldots, h$ with per-head width $d_k = 3d_{\text{model}}/h$,

$$\mathbf{Q}_i = \widetilde{\mathbf{H}}\mathbf{W}_i^Q, \mathbf{K}_i = \widetilde{\mathbf{H}}\mathbf{W}_i^K, \mathbf{V}_i = \widetilde{\mathbf{H}}\mathbf{W}_i^V, \tag{12}$$

$$\text{head}_i = \text{softmax}\left(\mathbf{Q}_i\mathbf{K}_i^\top / \sqrt{d_k}\right)\mathbf{V}_i, \tag{13}$$

$$\text{MSA}_{\text{shared}}(\widetilde{\mathbf{H}}) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right)\mathbf{W}^O, \tag{14}$$

with $\mathbf{W}_i^{Q,K,V} \in \mathbb{R}^{3d_{\text{model}} \times d_k}$ and $\mathbf{W}^O \in \mathbb{R}^{3d_{\text{model}} \times 3d_{\text{model}}}$. The attention output is added residually and normalized:

$$\mathbf{U} = \text{LN}\Big(\widetilde{\mathbf{H}} + \text{MSA}_{\text{shared}}(\widetilde{\mathbf{H}})\Big). \tag{15}$$

Equations (12)–(14) use a *single* set of attention weights $\{\mathbf{W}_i^{Q,K,V}, \mathbf{W}^O\}_{i=1}^h$ for the concatenated external+internal+hardware tokens, enforcing a unified attention geometry over modalities and time. This design (i) reduces parameters relative to per-modality attention, (ii) encourages cross-modal alignment at each timestep while conditioning on hardware specifications, and (iii) improves transfer by learning modality-agnostic interaction patterns that remain stable across manufacturers and hardware variants.

4.2 External Driver, Internal State Response, and Hardware Attribute Experts

To enable deep alignment among different modalities of stressors and improve cross-manufacturer generalization, *DegraMM* complements the *shared* multi-head self-attention (MSA) with three modality-specific experts applied after each MSA block: an *External Driver Expert* (ED-FFN) for environmental and operating drivers, an *Internal*

State Response Expert (ISR-FFN) for device and indicator responses, and a *Hardware Attribute Expert* (HW-FFN) that ingests textual hardware specifications (e.g., device model, controller/firmware identifiers, packaging, manufacturer notes, BOM-like descriptors) to learn hardware-specific attributes.

Three training/inference pathways Rather than always activating all experts, we define **three pathways** that selectively route expert computation across the L_x Transformer layers to encourage staged alignment.

Pathway I (External-only) focuses on environmental factor modeling by activating only ED-FFN across all L_x layers, ensuring environmental effects are fully represented and consistently aligned through depth.

Pathway II (External → Internal) uses a two-stage schedule over layers: the first F_x layers apply ED-FFN to let environmental effects intersect and align deeply, forming a strong foundation, while the remaining $L_x - F_x$ layers activate only ISR-FFN to operate on internal state information conditioned on the already-aligned environmental representations.

Pathway III (External → Hardware) mirrors Pathway II for hardware attributes: the first F_x layers activate ED-FFN to build deeply aligned environmental representations, and the remaining $L_x - F_x$ layers activate only HW-FFN, enabling the model to learn hardware-specific attributes grounded on the environmental alignment.

This staged expert scheduling preserves modality-specific inductive biases while the shared attention mechanism continuously aligns the fused streams into a unified temporal representation (Fig. 2).

4.2.1 Parameterization

Let $\mathbf{H}^{(\ell)} \in \mathbb{R}^{\tau \times 3d_{\text{model}}}$ be the input to Transformer layer $\ell \in \{0, \dots, L_x - 1\}$ and $\mathbf{U}^{(\ell)} = \mathrm{LN}\big(\mathbf{H}^{(\ell)} + \mathrm{MSA}^{(\ell)}_{\text{shared}}(\mathbf{H}^{(\ell)})\big)$ its post-attention state. We maintain fixed binary masks $\mathbf{M}_{\text{ext}}, \mathbf{M}_{\text{int}}, \mathbf{M}_{\text{hw}} \in \{0,1\}^{1 \times 3d_{\text{model}}}$ that select the external, internal, and hardware channel subspaces of the fused feature. The experts are three masked feed-forward modules. For $s \in \{\text{ext}, \text{int}, \text{hw}\}$, we define

$$\mathcal{F}_s(\mathbf{U}) = \mathbf{M}_s \odot \phi_s(\mathbf{U} \odot \mathbf{M}_s), \tag{16}$$

where

$$\phi_s(\mathbf{X}) = W_s^{(2)}\, \sigma\Big(W_s^{(1)}\mathbf{X}\Big). \tag{17}$$

The three experts are then instantiated as

$$\text{ED-FFN}(\mathbf{U}) = \mathcal{F}_{\text{ext}}(\mathbf{U}), \tag{18}$$

$$\text{ISR-FFN}(\mathbf{U}) = \mathcal{F}_{\text{int}}(\mathbf{U}), \tag{19}$$

$$\text{HW-FFN}(\mathbf{U}) = \mathcal{F}_{\text{hw}}(\mathbf{U}). \tag{20}$$

Layer-wise pathway gates We introduce *pathway-dependent* binary gates $\alpha^{(\ell)}_{\text{ext}}, \alpha^{(\ell)}_{\text{int}}, \alpha^{(\ell)}_{\text{hw}} \in \{0,1\}$ that determine which expert is active at layer ℓ. For convenience, define the indicator $\mathbb{I}[\cdot] \in \{0,1\}$. Then the three pathways correspond to:

$$\textbf{Pathway I:}\quad \alpha^{(\ell)}_{\text{ext}} = 1,\quad \alpha^{(\ell)}_{\text{int}} = 0,\ \alpha^{(\ell)}_{\text{hw}} = 0. \tag{21}$$

$$\textbf{Pathway II:}\quad \alpha^{(\ell)}_{\text{ext}} = \mathbb{I}[\ell < F_x],\quad \alpha^{(\ell)}_{\text{int}} = \mathbb{I}[\ell \geq F_x],\ \alpha^{(\ell)}_{\text{hw}} = 0. \tag{22}$$

$$\textbf{Pathway III:}\quad \alpha^{(\ell)}_{\text{ext}} = \mathbb{I}[\ell < F_x],\quad \alpha^{(\ell)}_{\text{int}} = 0,\ \alpha^{(\ell)}_{\text{hw}} = \mathbb{I}[\ell \geq F_x]. \tag{23}$$

where $\ell \in \{0, \dots, L_x - 1\}$, and F_x controls the depth devoted to environmental alignment before switching to internal-state or hardware-specific processing.

Expert-augmented block output With the pathway gates, the block output becomes

$$\mathbf{H}^{(\ell+1)} = \mathrm{LN}\Big(\mathbf{U}^{(\ell)} + \alpha^{(\ell)}_{\text{ext}}\,\text{ED-FFN}\big(\mathbf{U}^{(\ell)}\big) + \alpha^{(\ell)}_{\text{int}}\,\text{ISR-FFN}\big(\mathbf{U}^{(\ell)}\big) + \alpha^{(\ell)}_{\text{hw}}\,\text{HW-FFN}\big(\mathbf{U}^{(\ell)}\big)\Big). \tag{24}$$

When hardware text is unavailable, we set $\alpha^{(\ell)}_{\text{hw}} = 0$ and fall back to the other pathways.

4.2.2 Training Task

Let g_θ denote the *DegraMM* backbone consisting of the pathway-specific input construction (Eqs. (8)–(11)), the shared MSA blocks (Eq. (14)), and the pathway-aware expert scheduling (Eq. (24)). Let h_ϕ be a linear prediction head applied to the final hidden sequence, and define the pretraining model as $f_{\theta,\phi} = h_\phi \circ g_\theta$. We train *DegraMM* with multi-step next-window forecasting under a chosen pathway $p \in \{\text{ext}, \text{int}, \text{hw}\}$:

$$\hat{\mathbf{y}}_{t+\tau:t+2\tau-1} = f_{\theta,\phi}\Big(\mathbf{H}^{(0)}_p\Big), \tag{25}$$

where $\mathbf{H}^{(0)}_p \in \{\mathbf{H}^{(0)}_{\text{ext}}, \mathbf{H}^{(0)}_{\text{int}}, \mathbf{H}^{(0)}_{\text{hw}}\}$ denotes the pathway-specific input representation selected by

p, and $\mathbf{H}_{\text{ext}}^{(0)}, \mathbf{H}_{\text{int}}^{(0)}, \mathbf{H}_{\text{hw}}^{(0)}$ are constructed from $\widehat{\mathbf{H}}_{\text{ext}}, \widehat{\mathbf{H}}_{\text{int}}, \widehat{\mathbf{H}}_{\text{hw}}$ as in Sec. 4.1.2.[1] The prediction head maps the final backbone output $\mathbf{H}_p^{(L_x)} \in \mathbb{R}^{\tau \times 3d_{\text{model}}}$ to the horizon trajectory,

$$\hat{\mathbf{y}}_{t+\tau:t+2\tau-1} = h_\phi\left(\mathbf{H}_p^{(L_x)}\right), \tag{26}$$

where h_ϕ is applied independently at each timestep.

Given N training windows, we minimize the mean squared error (MSE) over windows and horizon:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N\tau} \sum_{i=1}^{N} \sum_{s=0}^{\tau-1} \left(\hat{y}_{t+\tau+s}^{(i)} - y_{t+\tau+s}^{(i)}\right)^2. \tag{27}$$

5 Experiments

5.1 Experimental Setup

5.1.1 Hardware & software

All experiments are implemented in PyTorch (v2.1) with CUDA (v12.1) and Python 3.10, and run on a single NVIDIA GeForce RTX 5090 (≥32 GB RAM). Mixed-precision training (float16) is enabled.

5.1.2 Data preprocessing & Evaluation metrics

Following prior work (Ma et al., 2022; Qiu and Tang, 2021; Tong et al., 2024), we adopt BE as the degradation metric for SMs, and we conduct all measurements under the testing conditions and prerequisites specified in the JJG596-2012 standard (Tang et al., 2012). Sliding windows of length $\tau = 128$ are formed with stride s (default $s = 32$). Each window uses inputs $(\mathbf{X}_{\text{ext}}^{t:t+\tau-1}, \mathbf{X}_{\text{int}}^{t:t+\tau-1})$ and predicts the next-window BE sequence $\mathbf{y}_{t+\tau:t+2\tau-1}$ (horizon $H = \tau$). Missing values are forward-filled for gaps ≤ 1 step and linearly interpolated for gaps ≤ 4 steps; windows with missing ratio $> \theta_{\text{miss}} = 0.2$ are discarded.

Following previous works (Ma et al., 2022; Qiu and Tang, 2021; Tong et al., 2024), we report root mean square error (RMSE) and mean absolute error (MAE)for performance of DTP. Both MAE and RMSE are lower-is-better metrics. For cross-manufacturer evaluation, we additionally report per-manufacturer scores and their macro mean ± standard deviation.

5.1.3 Baselines

We compare against representative probabilistic, kernel, and deep neural models:

KSVR (kernel support vector regression (Ma et al., 2023)) uses independent RBF kernels per feature with a composite fusion kernel;

WFB (weighted-fusion Bayesian (Qiu and Tang, 2021)) uses a Weibull likelihood with a two-level hierarchy;

BPNN (back-propagation neural network (Jian and Wei, 2021)) uses inputs {time, temperature, humidity} with architecture 3-106-1;

LSTM (long short-term memory (Zhao and Chen, 2023)) uses input/output sizes 22/1;

MVCNN (multiview CNN (Tong et al., 2024)) uses convolutional blocks followed by fully connected layers of sizes 256 and 128.

5.1.4 Implementations

We evaluate *DegraMM* at the same scale as the prevalent "base" Transformer models, with approximately 276M parameters and 12 layers. We use AdamW, cosine decay with 10% warmup, weight decay 5×10^{-2}, gradient clip 1.0, attention/FFN dropout 0.1, drop-path 0.1, GELU activations, and FFN expansion ratio $r_{\text{ff}} = 4$.

5.2 Cross-Manufacturer Generalization Performance

In this section, we benchmark *DegraMM* against representative probabilistic and kernel-based methods, as well as state-of-the-art deep neural baselines including MVCNN, on smart-meter degradation prediction across 30 manufacturers. This cross-manufacturer setting places stringent demands on model generalizability, since devices differ in both intrinsic design and operating conditions.

Overall, shown in Fig. 4, *DegraMM* delivers the highest accuracy and the most consistent performance across manufacturers. Specifically, *DegraMM* achieves **MAE** 0.0190 ± 0.0037 and **RMSE** 0.0222 ± 0.0047 (mean ± std), providing substantial error reductions over all baselines. Compared with the strongest deep competitors, *DegraMM* reduces MAE by **43.9%** relative to MVCNN (0.0338 ± 0.0066) and by **48.9%** relative to BPNN (0.0371 ± 0.0080). The improvements are even more pronounced over classical approaches: compared with KSVR, *DegraMM* achieves a **77.4%** MAE reduction ($0.0839 \pm 0.0205 \rightarrow 0.0190 \pm 0.0037$) and an **83.7%** RMSE reduction ($0.1363 \pm 0.0448 \rightarrow 0.0222 \pm 0.0047$). The per-model distributions in Fig. 4 also confirm *DegraMM*'s superior accuracy and robustness. Compared with the large dispersion of **KSVR/WFB**

[1] Concretely, $\mathbf{H}_{\text{ext}}^{(0)} = \text{Concat}(\widehat{\mathbf{H}}_{\text{ext}}, \mathbf{0}, \mathbf{0})$, $\mathbf{H}_{\text{int}}^{(0)} = \text{Concat}(\widehat{\mathbf{H}}_{\text{ext}}, \widehat{\mathbf{H}}_{\text{int}}, \mathbf{0})$, $\mathbf{H}_{\text{hw}}^{(0)} = \text{Concat}(\widehat{\mathbf{H}}_{\text{ext}}, \mathbf{0}, \widehat{\mathbf{H}}_{\text{hw}})$.

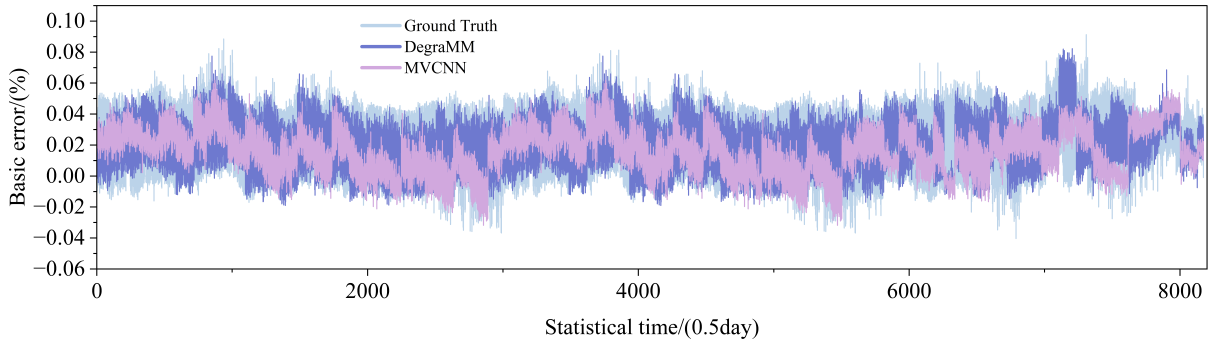

Figure 3: Timeline of prediction results across different models.

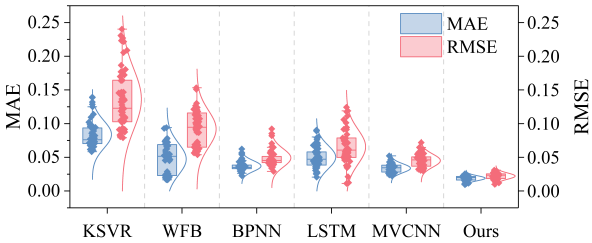

Figure 4: MAE and RMSE values under different methods.

and the noticeable cross-manufacturer variance of deep baselines, *DegraMM* achieves consistently low errors with a much tighter spread.

To further highlight the qualitative differences, Fig. 3 presents a representative meter on which MVCNN achieves its best performance. Even in this favorable setting, *DegraMM* follows the ground-truth degradation trajectory more closely, capturing the overall trend with smaller deviations. Combined with the quantitative results, this case study shows that *DegraMM* improves not only average prediction accuracy but also robustness and stability across manufacturers. These gains are largely attributable to *DegraMM* 's departure from *shallow fusion*: by explicitly incorporating hardware specifications alongside environmental and operational signals, *DegraMM* leverages stable device-level context to better explain cross-manufacturer variability. Built on a Multiway Transformer, it integrates modality-aware experts with shared attention to align and fuse heterogeneous stressors, while selective modality gating and long-range temporal aggregation enhance robustness under distribution shifts and improve transferability to previously unseen operating scenarios.

5.3 Component-Wise Analysis of *DegraMM*

In this section, we evaluate the effects of the three modality-aware experts (External Driver Expert(ED-FFN) / Internal State Response Expert(ISR-FFN) / Hardware Attribute Expert(HW-FFN)) and the shared MSA design. Table 1 reports MAE/RMSE (mean±std over manufacturers) for four variants: (i) ED-FFN only + MSA, (ii) ED-FFN + ISR-FFN + MSA, (iii) ED-FFN + ISR-FFN + HW-FFN + MSA (our full model), and (iv) three experts but *w/o* shared MSA. Both MAE and RMSE are lower-is-better metrics.

Using only the ED-FFN yields the worst performance, indicating that external conditions alone are insufficient for robust degradation prediction. Adding the ISR-FFN consistently reduces the error, and further introducing the HW-FFN leads to the best overall results, reducing MAE by 80.6% compared with using only the ED-FFN expert and highlighting the complementarity of the three heterogeneous factors. Moreover, removing the shared MSA (w/o MSA) causes a clear performance drop even with all experts enabled, increasing MAE by 61.9% and demonstrating that shared attention is crucial for effectively integrating heterogeneous information. Overall, the full *DegraMM* (three experts + shared MSA) achieves the lowest MAE/RMSE and the most stable performance across manufacturers, where robust fusion and alignment are critical for generalization.

6 Conclusion

We show that many existing degradation trajectory prediction (DTP) methods rely on *shallow fusion* of heterogeneous stressors, which limits their ability to capture long-range interactions and sustained degradation effects. To address this limitation, we propose *DegraMM*, a multimodal predictor built on a Multiway Transformer with three specialized experts for modality-aware modeling and effective fusion of heterogeneous stressors. *DegraMM* further incorporates selective modality gating to downweight noisy or weak modalities under distribution shifts, and employs long-range temporal aggregation to better represent persistent degradation patterns and complex stressor interactions.

ED-FFN	ISR-FFN	HW-FFN	MSA	MAE	RMSE
✓	–	–	✓	0.0976 ± 0.0154	0.1247 ± 0.0457
✓	✓	–	✓	0.0657 ± 0.0052	0.0692 ± 0.0064
✓	✓	✓	✓	0.0189 ± 0.0036	0.0221 ± 0.0046
✓	✓	✓	–	0.0497 ± 0.0089	0.0525 ± 0.0074

Table 1: Ablation results (mean $\pm$ standard deviation over manufacturers).

We evaluate *DegraMM* on a large-scale real-world smart-meter dataset collected in harsh operating environments, where it delivers consistent DTP improvements across diverse settings, including transfer to previously unseen operating scenarios, demonstrating strong effectiveness and robustness for practical deployment.

Limitations

Although *DegraMM* achieves strong performance on smart-meter degradation trajectory prediction, several limitations remain. First, our experiments are conducted on a real-world smart-meter dataset from a high-altitude testing base. While the data cover diverse manufacturers and harsh environmental conditions, additional validation in other regions, grid infrastructures, and industrial devices is still needed to further verify generalizability.

Second, the hardware-specification modality relies on the availability and quality of textual device information. In practice, such descriptions may be incomplete, inconsistent across manufacturers, or contain proprietary terminology, which could reduce the effectiveness of hardware-aware alignment.

Third, we mainly use basic error as the degradation indicator. Although it is an important measure of smart-meter health, other failure modes such as communication instability, timekeeping drift, and power-up/power-down anomalies are not fully modeled.

Finally, our experiments mainly evaluate prediction accuracy and cross-manufacturer generalization under offline settings. Real-world maintenance systems often require continual learning, adaptation to newly installed devices, and robustness to sensor drift or delayed measurements. Future work should investigate online updating, uncertainty estimation, and decision-oriented maintenance policies based on predicted degradation trajectories.

References

Runfei Chen, Shuyang Jiang, and Wei Huang. 2025. SeMob: Semantic synthesis for dynamic urban mobility prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15335–15355.

Junfeng Duan and Yao WenXuan. 2025. Predictive health management of smart meters: Daily measurement error forecasting. *IEEE Trans. Smart Grid*, 16(3):2429–2438.

Liang Guo, Naipeng Li, and Feng Jia. 2017. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing*, 240:98–109.

Ruizheng Huang, Zhicheng Zhang, and Yong Wang. 2025. Cross-MoE: An efficient temporal prediction framework integrating textual modality. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29927–29938.

Ma Jian and Qiu Wei. 2021. Research on the degradation trend of smart energy metering equipment based on iga-bp neural network. *Chinese Journal of Scientific Instrument*, 42(11):123–131.

Yang Li, Dong Qin, and H Vincent Poor. 2024. Bringing edge intelligence to smart meters via federated split learning. *Nature Communications*, 15(1):9044.

Bryan Lim, Sercan O. Arik, and Nicolas Loeff. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*.

Jun Ma, Wei Qiu, and Qiu Tang. 2022. Measurement error assessment for smart electricity meters under extreme natural environmental stresses. *IEEE Trans. Ind. Informat.*, 18(6):3905–3914.

Jun Ma, Zhaosheng Teng, and Qiu Tang. 2023. A novel multisource feature fusion framework for measurement error prediction of smart electricity meters. *IEEE Sensors Journal*, 23(17):19571–19581.

Yu Mo, Qianhui Wu, Xiu Li, and 1 others. 2021. Remaining useful life estimation via transformer encoder enhanced by a gated convolutional unit. *Journal of Intelligent Manufacturing*, 32:1997–2006.

Matteo Orlando, Abouzar Estebsari, and Enrico Pons. 2022. A smart meter infrastructure for smart grid iot applications. *IEEE Internet of Things Journal*, 9(14):12529–12541.

Yuhong Qin, Wei Qiu, and Yun Luo. 2025. Degradation analysis: Modeling and evaluating for electric energy meters under multistress. *IEEE Trans. Ind. Informat.*, 21(9):6615–6625.

Wei Qiu and Qiu Tang. 2021. Probability analysis for failure assessment of electric energy metering equipment under multiple extreme stresses. *IEEE Trans. Ind. Informat.*, 17(6):3762–3771.

Hong Tang, Xiaoxi Yang, and Leibing Shi. 2012. Electrical meters for measuring alternating-current electrical energy.

Philipp Thunshirn, Ines Mlinaric, and Jana Berg. 2025. A qualitative analysis of consumer motivations and barriers towards active smart meter utilization. *Energy Policy*, 203:114623.

Xuan Tong, Jun Ma, and Li Ma. 2024. A novel prediction method for smart meter error using multiview convolutional neural network. *IEEE Sensors Journal*, 24(24):42009–42017.

Ashish Vaswani and Noam Shazeer. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008.

Wenduan Xu, Stephen Clark, and Douglas Brown. 2024. Quantum recurrent architectures for text classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18020–18027.

Zhongyong Zhao and Yu Chen. 2023. Evaluation of operating state for smart electricity meters based on transformer–encoder–bilstm. *IEEE Trans. Ind. Informat.*, 19(3):2409–2420.

Shuang Zheng, Kiril Ristovski, Ahmed Farahat, and Chetan Gupta. 2017. Long short-term memory network for remaining useful life estimation. In *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 88–95. IEEE.