10.37965/jdmd.2026.1589

A Physics-Conditioned Multimodal Token-Stream Decoder and Contract-Grounded Workflow for Cross-Condition Aircraft Structural Health Monitoring

Yu Yang[1,2], Tianbao Nie[3] and Xiang Li[3,*]

[1] National Key Laboratory of Strength and Structural Integrity, Xi'an 710065, China;

[2] Aircraft Strength Research Institute of China, Xi'an 710065, China

[3] Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System, Xi'an Jiaotong University, Xi'an 710049, China

* Corresponding author: lixiang@xjtu.edu.cn

Received 30 June 2026; Revised 21 July 2026; Accepted 11 August 2026; Published online 18 August 2026

Fatigue cracking around rivet holes is a major concern in aircraft skin panels, while practical structural health monitoring must handle heterogeneous sensors, unseen loading amplitudes, missing channels, and limited coupon data. This study develops SHM-GPT, a compact physics-conditioned multimodal token-stream decoder, together with a contract-grounded monitoring workflow. A 26-slot causal representation combines acoustic-emission, strain, Fibre Bragg Grating, and operating-condition information. Percentage-life progress is estimated with a continuous Huber regression head; early-onset classification and modality reconstruction are auxiliary training objectives, and a variance-floor penalty discourages constant predictions. Spectral constraints are applied to the condition encoder and query-key projections to limit sensitivity under load shifts. The 4.72-million-parameter model is evaluated on a 15-coupon working set drawn from an in-house 7050-aluminium centre-hole campaign. Across five scenarios, it obtains the lowest mean absolute error in four, and meets the stated spread and information diagnostics in three: boundary load extrapolation, loading-profile shift, and leave-one-specimen-out evaluation, with relative margins of 11.93%, 15.28%, and 3.01% over the strongest baseline, respectively. A small Temporal Convolutional Network is better in the two-coupon regime, and all learned methods fail the spread diagnostic when acoustic emission is removed. The deterministic contract checker matches the rule-defined verdict on 50 of 50 synthetic traces. The evidence therefore supports a shared multimodal decoder and an auditable decision-support prototype on this benchmark.

10.37965/jdmd.2026.1589

Keywords: structural health monitoring; multimodal time series; cross-condition generalisation; contract-based decision support

1. Introduction

Aircraft skin panels and wing stiffeners are pierced by tens of thousands of rivet and connection holes, and fatigue cracking originating at these stress concentrators is the dominant mechanism behind in-service structural failure of an airframe. Current airworthiness standards therefore demand quantitative remaining-life estimation throughout the aircraft lifecycle. Despite three decades of structural health monitoring research, no deployed system simultaneously satisfies the operating requirements of an aircraft condition-based maintenance pipeline: heterogeneous sensor inputs spanning five decades of sampling rate (acoustic emission at 10^6 Hz, strain and Fibre Bragg Grating at 10 Hz); loading conditions that vary across stress amplitude, frequency, and multi-stage profiles; instrumented coupon counts that rarely exceed fifteen; and downstream tasks spanning health-index construction, percentage-life regression, early-onset crack detection, and missing-modality reconstruction [1–3]. Recent reviews report that even strong deep architectures degrade by 20–40% under unseen stress amplitudes or absent modalities [3–5], and the agentic systems built around them offer no formal account of which numerical claims in a maintenance recommendation are grounded in physical measurement.

These difficulties are tightly entangled for an airframe monitoring system. Acoustic emission is hit-driven, so its effective sampling rate fluctuates by up to fortyfold between specimens, whereas strain and Fibre Bragg Grating arrive on a regular ten-Hertz grid; a unified model must reconcile these incommensurable axes before any predictive head is invoked. Service stress amplitudes span a continuous range that no fatigue rig can densely populate, so the training distribution is sparser than the deployment distribution by construction, turning cross-condition extrapolation into a load-bearing requirement. Instrumented coupons are expensive to fabricate and load to failure, with each long-life experiment lasting tens to hundreds of hours; the resulting dataset is small enough that a model with more than a few million parameters risks memorising the training coupons. Sensor channels are routinely partial — fibres break, acoustic-emission sensors detach, strain gauges drift — so a reliable system must reconstruct the missing channel from those that remain. Finally, the consumer is an inspection technician, a maintenance planner, or an airworthiness regulator, each relying on a recommendation that synthesises diagnosis, prognosis, and action; that synthesis is the layer at which hallucinated numeric claims do the most harm [6,7].

Large pre-trained time-series models have expanded the range of forecasting architectures and transfer settings [8–15]. Directly applying that paradigm to aircraft structural health monitoring nevertheless raises four obstacles. Percentage-life labels vary across coupons because absolute cycle counts are not commensurable [16]. Cross-condition evaluation needs an explicit operating-condition representation. The

small number of instrumented coupons makes broad pre-training claims inappropriate for an in-house benchmark. Finally, maintenance recommendations must keep numerical model outputs separate from natural-language rendering [6,7,17].

Three methodological threads motivate the proposed design: autoregressive sequence modelling provides a common interface for heterogeneous temporal inputs [9,13,14]; spectral constraints can limit model sensitivity to perturbations [18,19]; and Structural Operational Semantics with contract-grounded tool execution provides a basis for auditable state transitions [17,20,21]. These elements are combined in SHM-GPT, a compact multimodal decoder whose continuous progress estimate is conditioned on a Paris-law-informed vector and consumed by a four-role monitoring workflow.

The token stream in the headline configuration refers to the causal representation of multimodal observations and operating conditions, not to a discretised progress output. Percentage-life progress is predicted by a continuous Huber regression head. A legacy 1024-bin cross-entropy head is retained only as an ablation because it produced a near-constant output in the evaluated boundary-extrapolation setting. This distinction aligns the model description with the implemented objective and prevents the ablation from being presented as the primary formulation.

SHM-GPT differs from the compared task-specific models in three respects. A continuous Paris-law-informed vector conditions the multimodal sequence; one compact decoder supplies a headline progress head plus early-onset and reconstruction heads used as auxiliary training objectives; and a deterministic contract checker evaluates the resulting maintenance state before optional language rendering. The two auxiliary outputs are not independently benchmarked in the present experiments.

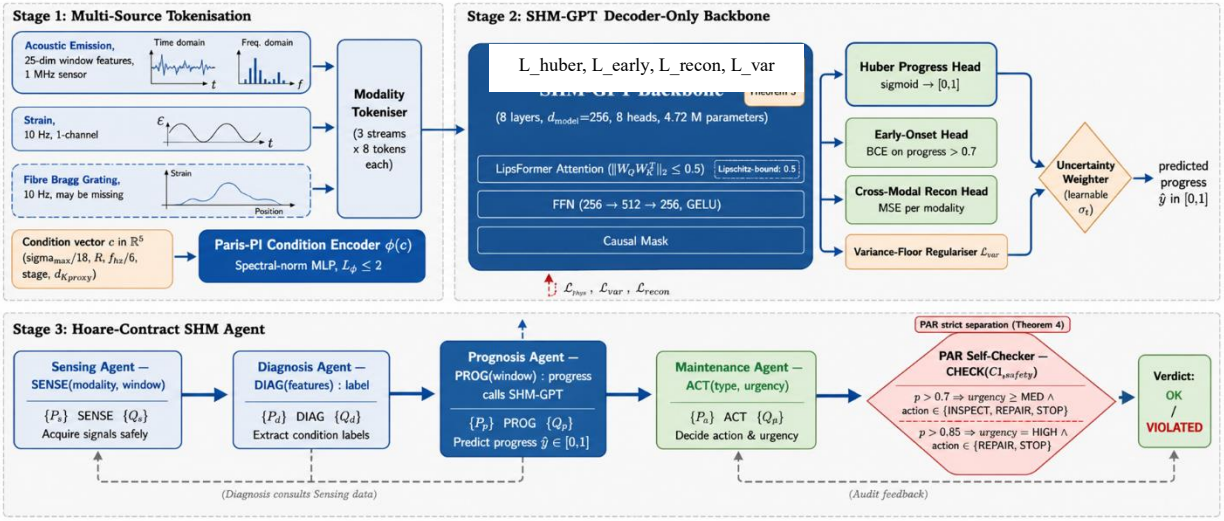

Fig. 1. Overview of the physics-conditioned multimodal decoder and four-role contract workflow.

The contributions of this work are summarised as follows.

A physics-conditioned token-stream formulation is introduced in which 25

observed conditioning tokens and one learned [PROG] query feed a shared 4.72-million-parameter decoder. The three heads read the query state at position 25; the headline progress output is continuous Huber regression, while the 1024-bin cross-entropy output is evaluated only as a legacy ablation. Replacing the latter restores prediction standard deviation from 0.033 to 0.165 in scenario A1.

A Paris-law-informed condition encoder and spectrally constrained attention projections provide an explicit sensitivity-control mechanism. On the reported benchmark, the model records lower mean absolute error than the strongest baseline by 11.93% in boundary load extrapolation and 15.28% under loading-profile shift. These empirical margins are not interpreted as a formal guarantee of extrapolation accuracy.

A contract-grounded monitoring workflow is specified through four functional roles, a five-production Backus-Naur Form grammar, five Structural Operational Semantics rules, and a deterministic Prediction-Action-Result checker. The checker returns the rule-defined verdict on 50 of 50 synthetic traces; this is an implementation test of the stated contract, not a proof of operational safety.

2. Related Work

2.1. Decoder-Only Time-Series Foundation Models

Recent time-series models differ substantially in architecture and objective. TimesFM uses decoder-only patch forecasting [8]; Chronos discretises scaled values into tokens [9]; Timer and AutoTimes use autoregressive forecasting interfaces [11,13]; and Time-MoE adds sparse experts at billion-parameter scale [14]. By contrast, Moirai uses a masked universal forecasting architecture [12], MOMENT is an encoder-family model for several time-series tasks [10], and MSGNet is a task-specific multivariate forecasting network rather than a foundation model [22]. The survey in [15] provides a broader taxonomy.

Discrete output vocabularies can be effective in large and diverse forecasting corpora, but that behaviour should not be assumed for a small structural-health dataset. In the present A1 ablation, the 1024-bin cross-entropy head produced a prediction standard deviation of 0.033 and a mean absolute error of 0.2385, whereas the continuous Huber head produced 0.165 and 0.2309. The headline configuration therefore uses Huber regression. This dataset-specific observation is not presented as a general limitation of tokenised time-series models.

2.2. Structural Health Monitoring with Deep Learning

Deep learning for structural health monitoring has been surveyed extensively [1–4]. Representative work includes guided-wave imaging [23], graph-convolution and transformer fusion for composite damage detection [2], cross-attention for guided-wave localisation [4], transformer-based indirect rail monitoring [24], explainable composite diagnostics [25], and Paris-law-informed fatigue modelling [26]. Prognostic studies include contrastive recurrent representations [27], attention-based remaining-life prediction [28], diffusion-based augmentation [29], and interval health monitoring [30]. Related multimodal industrial-diagnosis studies further illustrate the use of heterogeneous observations in intelligent operation and maintenance [37,39]. The present benchmark evaluates progress regression from a shared decoder trained with early-onset and reconstruction

auxiliaries rather than claiming broad pre-training across industrial domains.

Recent studies have continued to diversify data-driven monitoring and prognostics across structures, machinery, and aircraft-oriented systems [2,4,24–26,31,32]. Parallel developments also show the breadth of limited- and heterogeneous-data health assessment in electric-aircraft and industrial settings [38,42,43], as well as physics-integrated prediction elsewhere in electric-aircraft engineering [41]. Against this background, the present study focuses on one aircraft-coupon benchmark.

Three methodological themes are relevant. Operating conditions can be represented categorically, aligned through domain adaptation, or encoded continuously; the present encoder uses a bounded continuous vector containing a Paris-law proxy. The modality patchers implement middle fusion by producing eight tokens per sensor stream for joint decoder processing. The decision layer is a deterministic post-prediction state machine whose contract can be inspected independently of an optional renderer. No claim is made that these choices dominate all alternative fusion or domain-generalisation strategies.

2.3. Lipschitz-Bounded and Spectral-Normalised Attention

LipsFormer introduced a transformer design based on Lipschitz-oriented components, including scaled-cosine attention and normalisation choices [18], while explicit Lipschitz regularisation has also been studied as a way to reduce transformer overconfidence [19]. The present implementation does not reproduce the complete LipsFormer block: it uses standard scaled dot-product attention with power-iteration clipping of the query-key projection product. It is therefore described below as spectrally constrained or LipsFormer-inspired attention, and the resulting bound is interpreted as a sensitivity relation rather than an accuracy guarantee under unseen loads.

2.4. Cross-Domain Transfer and Domain Generalisation

Cross-domain diagnostics has considered semi-supervised pre-training [31], multi-fidelity transfer [32], predictive-maintenance ensembles [33], small-data prognostics [5], meta-learning-based domain generalisation [34,35], and federated fault diagnosis under communication constraints [40]. The present approach instead places a continuous operating-condition vector inside the causal decoder. The comparison is conceptual because no federated or broad cross-domain benchmark is evaluated here.

2.5. Contract-Grounded Agent Systems and Hallucination Control

Recent work on safe artificial intelligence and verified tool execution emphasises traceability between numerical claims and deterministic computations [7,17], with Structural Operational Semantics [20] and Hoare logic [21] providing formal specification tools. In this study, Sensing, Diagnosis, Prognosis, and Maintenance are four functional roles within one SHM-Agent state machine rather than four independently trained or autonomous agents. The runtime checker compares a trace with two explicit policy rules and rejects detected numeric substrings that are absent from the allowed trace or constant set.

3. Methodology

3.1. Problem Formulation

Let $\mathcal{S} = \{s_1, \ldots, s_{N_s}\}$ denote the set of instrumented aircraft coupons. For each

coupon s, a sliding window of length $W = 128$ strain samples (about 12.8 s at 10 Hz) yields a multi-modal observation

$$\mathbf{x}_t = (\mathbf{x}_t^{\mathrm{AE}} \in \mathbb{R}^{W\times 25}, \mathbf{x}_t^{\mathrm{STR}} \in \mathbb{R}^{W\times 1}, \mathbf{x}_t^{\mathrm{FBG}} \in \mathbb{R}^{W\times 1}, h_t^{\mathrm{FBG}} \in \{0,1\}) \quad (1)$$

where the acoustic-emission features are 25-dimensional (19 time-domain plus 6 frequency-domain), the strain and Fibre Bragg Grating channels are scalar at 10 Hz, and h_t^{FBG} is a binary availability mask. Here, the subscript t indexes the sliding window along the load history, N_s is the number of instrumented coupons, the superscripts AE, STR and FBG denote the acoustic-emission, strain and Fibre Bragg Grating modalities respectively, and R is the set of real numbers, so that $\mathrm{R}^{W\times 25}$ is the space of real W-by-25 feature matrices. The static condition vector

$$\mathbf{c} = (\tfrac{\sigma_{\max}}{18}, R, \tfrac{f_{\mathrm{Hz}}}{6}, 1[\mathrm{multi-stage}], \Delta K_{\mathrm{proxy}}) \in \mathbb{R}^5 \quad (2)$$

collects the loading amplitude, stress ratio, cycling frequency, loading profile flag, and a Paris-law stress-intensity-range proxy $\Delta K_{\mathrm{proxy}} = \sigma_{\max}(1-R)$. Here, the constants 18 and 6 normalise the maximum stress $\sigma_{\max}$ (in kN) and the cycling frequency f_{Hz} (in Hz) onto the unit interval, and $1[\mathrm{multi-stage}]$ is the indicator that equals 1 for a multi-stage loading profile and 0 otherwise. The target is the percentage-life progress $y_t \in [0,1]$,

$$y_t = \frac{N_{\mathrm{elapsed}}(t)}{N_{\mathrm{failure}}} \quad (3)$$

which renders absolute cycle counts incommensurable across coupons and forces a percentage-only formulation.

Here, $N_{\mathrm{elapsed}}(t)$ is the number of fatigue load cycles accumulated up to window t, N_{failure} is the total number of cycles the coupon sustains before failure, and their ratio y_t is the percentage-life progress.

3.2. SHM-GPT Architecture

The SHM-GPT shared decoder comprises a Paris-law-informed condition encoder, per-modality patchers, a 26-slot causal layout, and eight spectrally constrained transformer blocks with one headline progress head, two auxiliary training heads, and one variance-floor regulariser (Fig. 1). The model contains approximately 4.72 million parameters.

3.2.1. Paris-Law Condition Encoder

A two-layer perceptron with spectral-norm-constrained weights produces the condition embedding

$$\varphi(\mathbf{c}) = \mathbf{W}_2\mathrm{GELU}(\mathbf{W}_1\mathbf{c}), \quad \|\mathbf{W}_1\|_2 \le 1, \|\mathbf{W}_2\|_2 \le 1 \quad (4)$$

so that the global Lipschitz constant satisfies $L_\varphi \le 2$. Two spectrally-bounded layers, rather than one, allow non-linear interaction between $\sigma_{\max}$ and R while preserving an a-priori upper bound on the encoder gradient, which provides an a-priori cap on the encoder gain.

Here, $\mathbf{W}_1$ and $\mathbf{W}_2$ are the learnable weight matrices of the two perceptron layers, GELU is the Gaussian Error Linear Unit activation, and $\|\cdot\|_2$ denotes the spectral norm (the largest singular value), whose unit bound on each layer caps the gain of the encoder.

To avoid symbol overloading, the condition vector is denoted by c, the projection-product clipping threshold by γ, the condition-encoder Lipschitz constant by L_φ, and the full forward-map constant by L_f throughout the manuscript.

3.2.2. Modality Tokeniser and Causal Sequence

Three strided one-dimensional convolutions $\mathrm{ConvPatch}_m$ project each modality stream into eight tokens of dimension $d = 256$. When $h_t^{\mathrm{FBG}} = 0$, the Fibre Bragg Grating tokens are replaced by a learnable mask embedding. The final causal sequence of length 26 is

$$\mathrm{seq}_t = [\varphi(\mathbf{c}), \mathrm{ConvPatch}_{\mathrm{AE}}(\mathbf{x}_t^{\mathrm{AE}}), \mathrm{ConvPatch}_{\mathrm{STR}}(\mathbf{x}_t^{\mathrm{STR}}), \mathrm{mask\_or}(\mathrm{ConvPatch}_{\mathrm{FBG}}(\mathbf{x}_t^{\mathrm{FBG}}), h_t^{\mathrm{FBG}}), \mathbf{e}_{\mathrm{PROG}}] \tag{5}$$

The layout contains 25 conditioning tokens at positions 0–24 and a learned [PROG] query at position 25 (Fig. 2). The query is a fixed trainable embedding rather than the ground-truth progress value, so it does not reveal the target.

The continuous progress, early-onset, and reconstruction heads consume the hidden state of the learned query at position 25 after it has attended causally to positions 0–24. The token-24 annotation embedded in Fig. 3 is a legacy indexing error and should read token 25. The legacy 1024-class head uses the same query state only in the ablation of Section 4.4.

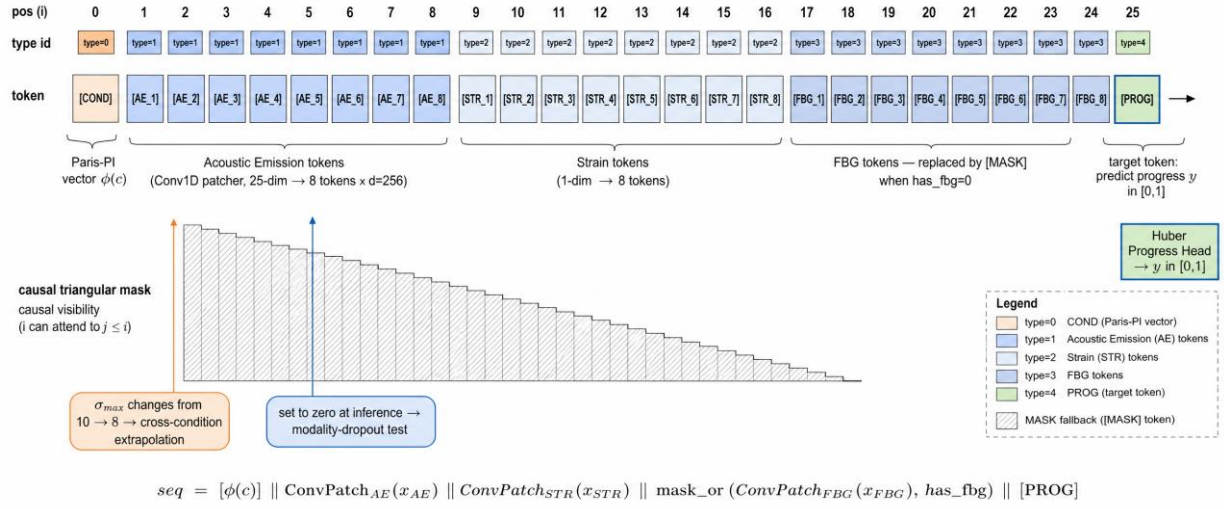

Fig. 2. Causal layout with 25 conditioning tokens and one learned [PROG] query at position 25.

3.2.3. Spectrally Constrained Attention

Each decoder block uses standard scaled dot-product attention, while two-step power-iteration clipping limits the query-key projection product before the attention operation in Eq. (6). Because the attention calculation differs from the scaled-cosine and normalisation components of the original LipsFormer, the implementation is described as LipsFormer-inspired spectral constraint rather than a reproduction of LipsFormer.

$$\mathrm{Attn}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_h}}\mathbf{x}W_Q W_K^T \mathbf{x}^T\right)\mathbf{x}W_V \tag{6}$$

The attention output is followed by a residual connection and a feed-forward block; eight spectrally constrained decoder blocks form the shared backbone (Fig. 3).

Here, **x** is the input token matrix of the attention block, W_Q, W_K and W_V are the learnable query, key and value projection matrices, d_k Scenario-wise spread: see Fig. 8 and softmax is :

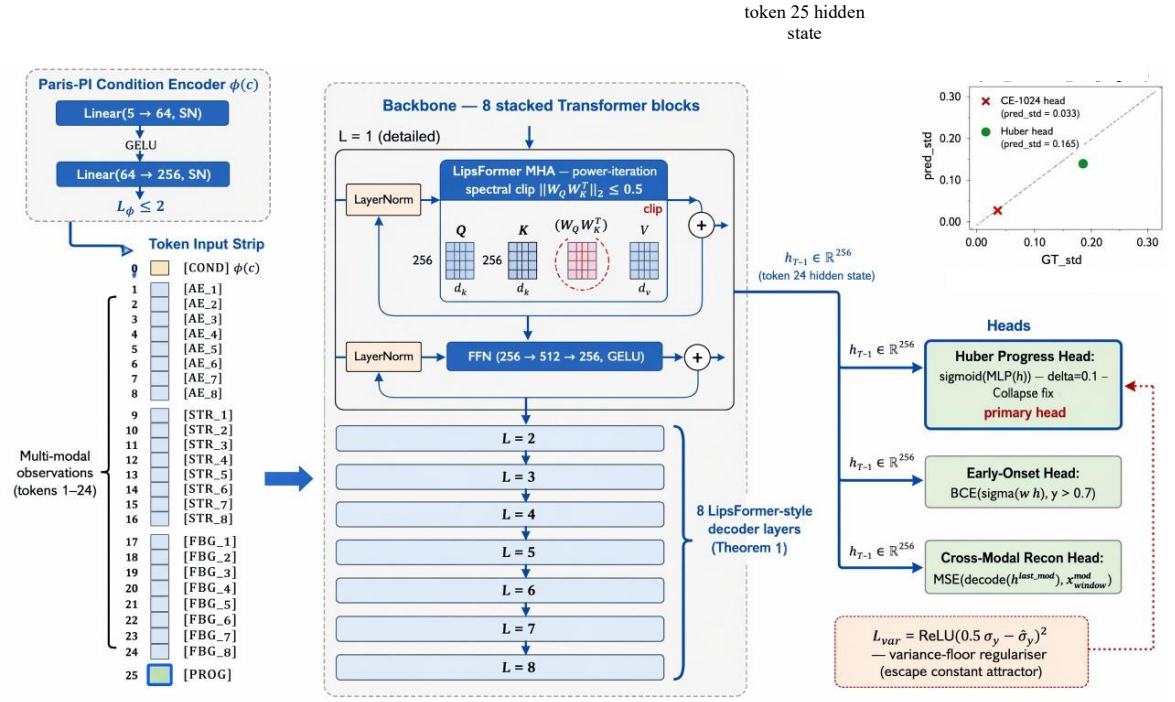

Fig. 3. SHM-GPT shared backbone with spectrally constrained attention and continuous Huber progress head.

3.2.4. Huber Progress Head and Variance-Floor Regulariser

Two progress heads were evaluated. The initial design used cross-entropy over 1024 uniform bins, but Section 4.4 shows that this head produced a near-constant prediction in scenario A1. The headline configuration instead uses a sigmoid-bounded perceptron with the Huber objective in Eq. (7).

$$\mathcal{L}_{\text{huber}}(\hat{y}, y) = \begin{cases} \frac{1}{2}(\hat{y} - y)^2 & |\hat{y} - y| \leq \delta \\ \delta(|\hat{y} - y| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}, \delta = 0.1 \qquad (7)$$

Here, $\hat{y}$ is the percentage-life progress predicted by the regression head, y is the ground-truth label of Eq. (3), and $\delta = 0.1$ is the threshold at which the loss switches from its quadratic to its linear regime.

A variance-floor regulariser

$$\mathcal{L}_{\text{var}} = \text{ReLU}\left(0.5\sigma_y - \sigma_{\hat{y}}\right)^2 \qquad (8)$$

The variance-floor regulariser in Eq. (8) is added with weight 0.1. The architecture contains one headline progress head and two auxiliary training heads for early-onset classification and per-modality reconstruction. Together with the variance-floor penalty, these yield four optimisation terms: Huber, binary cross-entropy,

reconstruction mean-squared error, and variance regularisation. Auxiliary-task performance is not independently evaluated in the present benchmark.

Here, σ_y is the standard deviation of the ground-truth labels over the batch, $\sigma_{\hat{y}}$ is the standard deviation of the corresponding predictions, and ReLU is the rectified linear unit; the term grows only once the prediction spread falls below half the label spread.

The variance-floor regulariser balances prediction diversity and fitting accuracy through a one-sided activation. When prediction spread is at least half of the ground-truth spread, the regulariser is exactly zero and the Huber and auxiliary losses determine the fit, so accurate predictions are not pushed toward artificially high variance. When prediction spread falls below the floor, the quadratic penalty supplies a gradual gradient away from the constant-prediction attractor. Its weight of 0.1 limits this correction relative to the data-fitting objective. The term therefore enforces minimum diversity only in the degenerate regime rather than rewarding variance indiscriminately.

3.3. Training Objective

The aggregate objective combines the four loss terms through the uncertainty-weighting scheme of [36],

$$L = \sum_{t\in\{huber,early,recon,var\}}\left(\frac{1}{2\sigma_t^2}\rho_t L_t + log\sigma_t\right) \quad (9)$$

where σ_t is a learnable scalar and ρ_t a pre-scaling factor calibrated on the first training batch so that all losses enter on the same order of magnitude. The optimiser is AdamW (learning rate 10^{-4}, weight decay 10^{-4}, gradient-norm clipping 1.0) with a cosine schedule over 10 epochs at batch size 24.

In Eq. (9), the summation ranges over four loss terms—Huber progress, early-onset classification, modality reconstruction, and variance regularisation. The learnable uncertainty scalars weight loss terms; they are not additional task heads.

3.4. Sensitivity-Control Design and Runtime Check

Spectral clipping and trace checking serve different engineering purposes. The former constrains selected parameter operators during training, whereas the latter checks rendered numeric substrings against a deterministic trace after inference. Neither mechanism is treated as a formal guarantee of prediction accuracy or maintenance safety.

The condition encoder and query-key projection product are spectrally constrained during optimisation. These constraints limit selected operator norms, but a global input-output Lipschitz bound would additionally require an explicit input domain, bounded token norms, a LayerNorm denominator bound, and constants for softmax, residual, and feed-forward paths. Because those assumptions are not established here, no formal extrapolation or Lipschitz theorem is claimed.

The spectral constraint is therefore interpreted as a sensitivity-control design choice. Its contribution is not isolated by the single-seed A1 ablation, and the empirical load-shift results should not be attributed to this mechanism alone.

Monotonicity is not enforced in the reported objective. The trajectory reversals visible in Fig. 9 confirm that percentage-life predictions should be interpreted window by window rather than as a guaranteed monotone crack-growth path. A monotonic

regulariser remains a future experimental option.

The runtime checker applies a lexical numeric-substring comparison to the reported trace schema. Detected numerals must match an allowed policy constant or deterministic tool field. This guard does not establish semantic equivalence of number formats or units and does not detect unsupported qualitative wording.

The 50-trace experiment in Section 4.5 is consequently a software test against a rule-defined oracle. It does not validate the policy thresholds, the predictive model, or unrestricted language generation.

3.5. Hoare-Contract Industrial Agent Protocol

The SHM-Agent is a single contract-governed state machine with four functional roles—Sensing, Diagnosis, Prognosis, and Maintenance—and five productions: SENSE, DIAG, PROG, ACT, and CHECK. Each production heads one Structural Operational Semantics rule describing a state transition (Fig. 4). The illustrative policy contract is

$$\text{progress} > 0.7 \Rightarrow \text{urgency} \geq \text{MED} \wedge \text{action} \in \{\text{INSPECT}, \text{REPAIR}, \text{STOP}\} \quad (10)$$

$$\text{progress} > 0.85 \Rightarrow \text{urgency} = \text{HIGH} \wedge \text{action} \in \{\text{REPAIR}, \text{STOP}\} \quad (11)$$

Here, the progress value is the highest percentage-life estimate available to the workflow; urgency belongs to {LOW, MED, HIGH}, and action belongs to {CONTINUE, INSPECT, REPAIR, STOP}. The constants 0.7 and 0.85 are illustrative policy thresholds used to exercise the checker. They are not derived from an airworthiness standard and must be replaced by operator-approved values before field use.

The first rule disallows LOW urgency or CONTINUE above 70% predicted progress; the second requires HIGH urgency with REPAIR or STOP above 85%. CHECK applies these deterministic Boolean rules after prediction. Because strict greater-than comparisons are used, exactly 70% does not activate the first clause and exactly 85% activates only the first. This boundary convention is part of the software policy, not a physical crack-growth threshold.

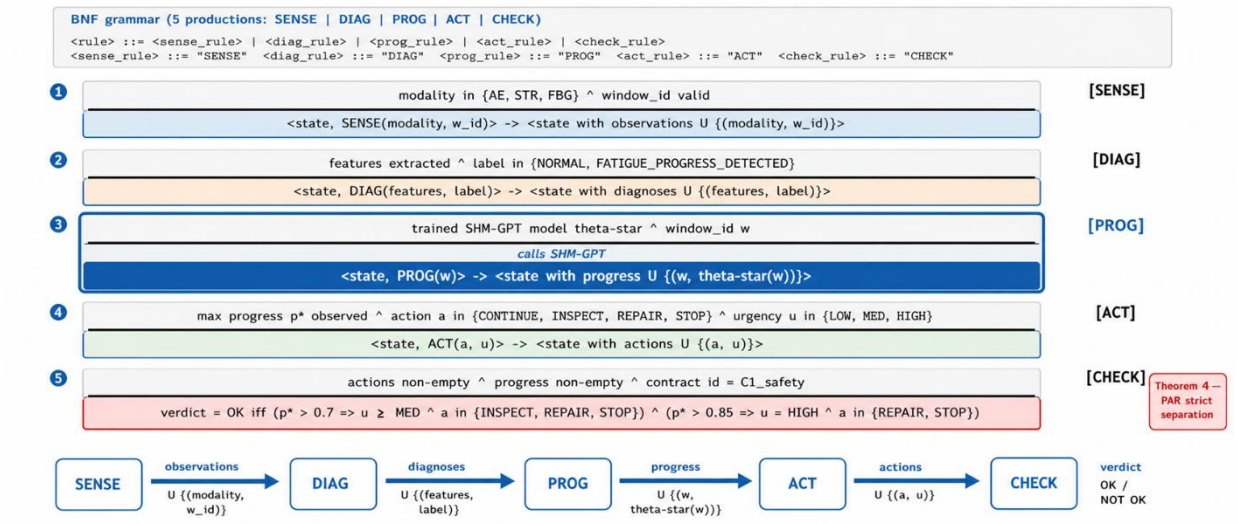

Fig. 4. Structural Operational Semantics rules for the four functional roles within one SHM-Agent workflow.

10.37965/jdmd.2026.1589

PROG invokes the trained SHM-GPT; ACT records an action and urgency; and CHECK compares the resulting state with the illustrative contract, returning OK or NOT OK. An optional renderer may paraphrase the trace, after which the implemented substring checker compares detected numerals with allowed constants and tool outputs (Fig. 5). The checker is a traceability guard within the tested schema; it does not prove that unrestricted natural language is free of hallucination.

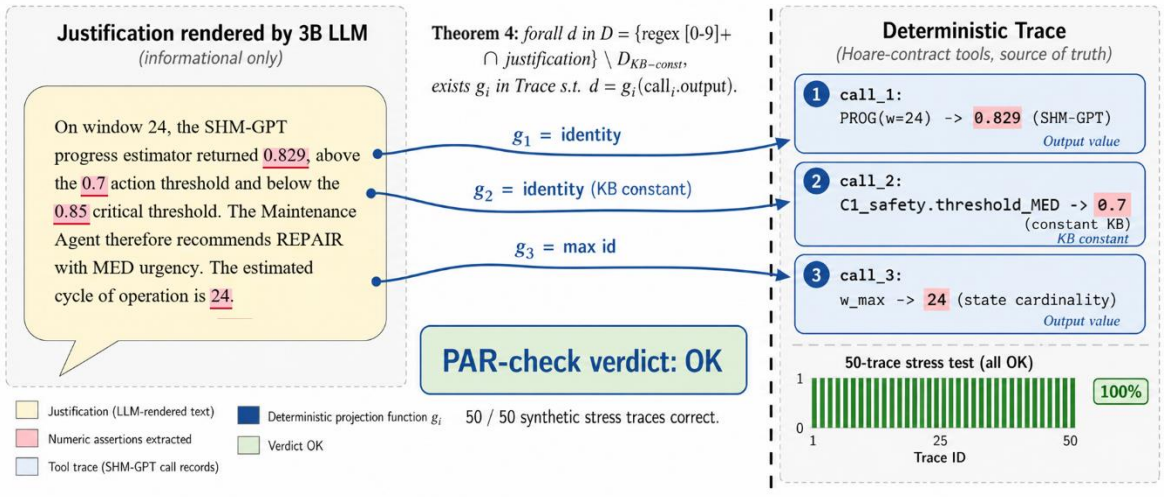

Fig. 5. Numeric-substring traceability in the illustrative renderer and deterministic tool trace.

10.37965/jdmd.2026.1589

3.6. Training Procedure

Algorithm 1 summarises training. Pre-scaling factors are calibrated once on the first batch, and the variance-floor regulariser is inactive whenever prediction spread already exceeds the stated floor.

Algorithm 1. SHM-GPT training under the Hoare-contract pipeline

Initialise θ; pre-scaling $\rho_t \leftarrow 1$
  for epoch $= 1 \ldots E$ do
    for batch $\mathcal{B}$ in training set do
      Compute $\varphi(\mathbf{c})$ with spectral-norm-clipped weights
      Build causal sequence seq from Eq. (5)
      Forward through 8 spectrally constrained blocks; clip $W_Q W_K^T$ to $\gamma = 0.5$
      Compute Huber, early, reconstruction, variance losses; combine via Eq. (9)
      if first batch then
        Calibrate $\rho_t \leftarrow 1/\max(\mathcal{L}_t, \varepsilon)$
      end if
      Backpropagate; clip $\|\nabla\theta\|_2 \leq 1$; AdamW step
    end for
    Cosine learning-rate update
  end for
  return $\theta^\star$

Algorithm 1 trains the shared decoder end-to-end. The condition embedding and 26-slot layout are assembled before propagation through eight spectrally constrained blocks. The Huber, early-onset, reconstruction, and variance-floor terms are combined through Eq. (9), followed by gradient clipping, an AdamW update, and a cosine learning-rate schedule. The contract checker is evaluated after prediction and does not contribute a training loss.

The symbols in Algorithm 1 follow Eqs. (4)–(9). The clipping constant applies to the query-key projection product, and the small positive denominator guard prevents division by a vanishing first-batch loss. The returned parameters comprise the modality patchers, condition encoder, decoder blocks, the headline progress head, and two auxiliary training heads.

4. Experimental Study

4.1. Datasets and Scenarios

The primary dataset is StructureData, an in-house campaign containing twenty-seven 7050-aluminium centre-hole coupons. The reported modelling analysis uses a pre-existing working whitelist of fifteen coupons with aligned multimodal records; thirteen contain all three sensor modalities. The remaining twelve coupons are outside this aligned working set and do not contribute to the reported training or test results. This distinction is stated explicitly to prevent the 27-coupon campaign from being interpreted as the effective modelling sample size. Percentage-life labels follow Eq. (3).

The experimental campaign was conducted on an LF5105 electro-hydraulic servo fatigue testing machine with a rated capacity of 100 kN. Each 7050-aluminium sheet coupon contains a circular centre notch that localises stress concentration and approximates crack initiation and growth near aircraft rivet or connection holes. In the long-life groups, sinusoidal cyclic loading was applied at 5 Hz and R=0.1 with paired maximum/minimum loads of 10/1 kN or 8/0.8 kN; replicate specimens were tested to support cross-specimen validation rather than single-coupon fitting.

During each run, the acoustic-emission, strain-gauge, and Fibre Bragg Grating systems recorded the full process from crack initiation to propagation. Two broadband differential acoustic-emission sensors (100-900 kHz) were mounted symmetrically at

the specimen ends, 160 mm apart, and acquired by a PAC PCI-DSP 32-channel system at 1 MHz. The Fibre Bragg Grating sensors were bonded near the notch and read by an AGS-WA40 interrogator, while transverse and longitudinal strain gauges around the notch were sampled by a JHDY-0108 dynamic strain instrument.

Raw records were converted into aligned 0.1-s observations before modelling. Acoustic-emission waveforms were divided into 100,000-point windows at 1 MHz and summarised by 25 time- and frequency-domain features. Strain and Fibre Bragg Grating records flagged during acquisition preprocessing were removed before timestamp alignment, after which each retained row contained synchronous sensor values. The precise acquisition-side rejection thresholds are not recoverable from the modelling record and are therefore a reproducibility limitation rather than an inferred methodological detail.

Fig. 6 shows the aluminium centre-hole coupon and the multimodal sensor arrangement on the fatigue testing machine.

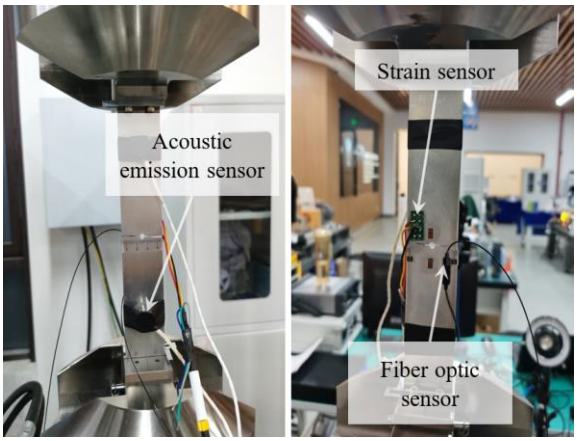

Fig. 6. Experimental 7050-aluminium centre-hole coupon and acoustic-emission, strain-gauge, and Fibre Bragg Grating sensor arrangement.

Five scenarios probe distinct operating shifts. A1 trains on five 10-kN long-life coupons and tests on three 8-kN coupons. A2 is five-fold leave-one-specimen-out evaluation within the 10-kN group and therefore measures cross-coupon, not cross-condition, generalisation. A3 trains on nine multi-stage coupons and tests on five single-stage coupons. A4 trains on two high-stress coupons and tests on five 10-kN coupons as a small-data stress test. A5 repeats A1 with acoustic emission removed at inference.

Three random seeds (42, 123, 2024) are used per scenario; all numbers are mean and population standard deviation across seeds. The four baselines are a four-layer Long Short-Term Memory, a six-layer Temporal Convolutional Network, a three-hidden-layer Multi-Layer Perceptron on window summary statistics, and a Naive constant predictor emitting the training-mean progress, all consuming the same windows and condition vectors. Implementation uses PyTorch 2.4 on a single RTX 4000 Ada 20 GB GPU; training takes about 4.5 minutes per (scenario, seed).

4.2. Fairness Gate

A scenario is labelled eligible only when three stated descriptive diagnostics pass. The surface diagnostic requires the lowest mean absolute error and at least a 1% relative margin over the next method. The spread diagnostic requires prediction standard deviation to be at least half the ground-truth standard deviation. The information diagnostic requires mean absolute error below 95% of the optimal single-constant error; the unrounded long-life threshold is 0.2356. These diagnostics screen obvious near-constant solutions but are not statistical significance tests.

The three diagnostics separate surface inferiority, low prediction spread, and

insufficient improvement over a constant. Their thresholds are heuristic and calibrated to the reported long-life labels; they are neither confidence intervals nor regulatory acceptance criteria. Results are therefore reported for all five scenarios, including the two that fail the spread diagnostic.

4.3. Main Results

Table 1 reports the mean absolute error of each method on each scenario, with the lowest value per row in bold; Fig. 7 shows the same data as a grouped bar chart with three-seed error bars.

Table 1. Mean absolute error on five scenarios over three random seeds.

Scenario	SHM-GPT	LSTM	TCN	MLP	Naive
A1 (σ=10 → σ=8)	0.2252 ± 0.0040	0.2635 ± 0.0223	0.2637 ± 0.0064	0.2557 ± 0.0063	0.2937 ± 0.0000
A2 (σ=10 LOSO five-fold)	0.2091 ± 0.0028	0.2156 ± 0.0085	0.2259 ± 0.0108	0.2254 ± 0.0022	0.2837 ± 0.0000
A3 (multi → single stage)	0.1813 ± 0.0023	0.2340 ± 0.0255	0.2683 ± 0.0055	0.2140 ± 0.0062	0.2837 ± 0.0000
A4 (σ=17/18 → σ=10, N=27)	0.2358 ± 0.0007	0.2324 ± 0.0092	0.2128 ± 0.0111	0.2350 ± 0.0053	0.2837 ± 0.0000
A5 (σ=10 → σ=8 + AE mask)	0.2162 ± 0.0020	0.2385 ± 0.0225	0.2669 ± 0.0132	0.2290 ± 0.0039	0.2937 ± 0.0000

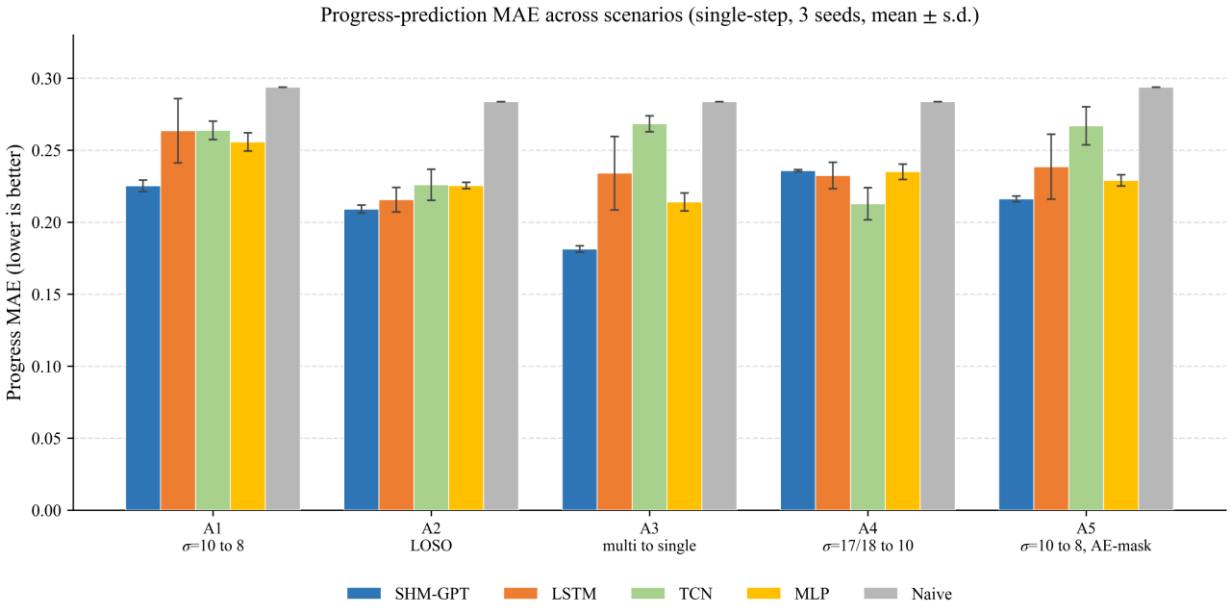

Fig. 7. Mean absolute error across scenarios with three-seed error bars.

SHM-GPT has the lowest mean absolute error in four of five scenarios (A1, A2, A3, and A5). Three scenarios also meet the spread and information diagnostics,

corresponding to three of five overall. Relative to the strongest listed baseline, the mean margins are 11.93% in A1, 15.28% in A3, and 3.01% in A2. The associated prediction-to-truth standard-deviation ratios are 0.58, 0.74, and 0.56. A2 contains five folds under each of three seeds, whereas the table reports the scenario-level mean and population standard deviation. Given only three seeds, these differences are descriptive and are not presented as statistically significant.

The scenarios support a limited interpretation. A1 tests a load-amplitude shift, A2 tests held-out coupons at a fixed load, and A3 tests a loading-profile shift. The Huber ablation directly supports the choice of progress head, but the main comparison does not isolate causal effects of the Paris-law encoder, spectral constraint, or variance term. A4 and A5 expose the small-data and missing-modality limits discussed in Section 4.6.

The baselines respond heterogeneously. The Multi-Layer Perceptron is the strongest competitor in A1 and A3 because its window summary statistics retain partial progress information under condition shift, yet cannot exceed strain root-mean-square. The Long Short-Term Memory leads among baselines in A2, where within-condition temporal regularity lets recurrent state lock onto coupon-specific trajectories. The Temporal Convolutional Network, with fewer parameters, is strongest in A4 yet weakest in A3. The Naive predictor is, as expected, weakest throughout.

Fig. 8 compares prediction spread with half the ground-truth spread and compares mean absolute error with the optimal-constant diagnostic. A1–A3 pass both diagnostics, while A4 and A5 fall below the spread floor. Decisions use the unrounded information threshold of 0.2356 even where a figure label is displayed to three decimals. In A5, SHM-GPT has lower mean absolute error than the Multi-Layer Perceptron (0.2162 versus 0.2290), but its prediction standard deviation is only 0.104; A5 is therefore reported as a stress test rather than an eligible result.

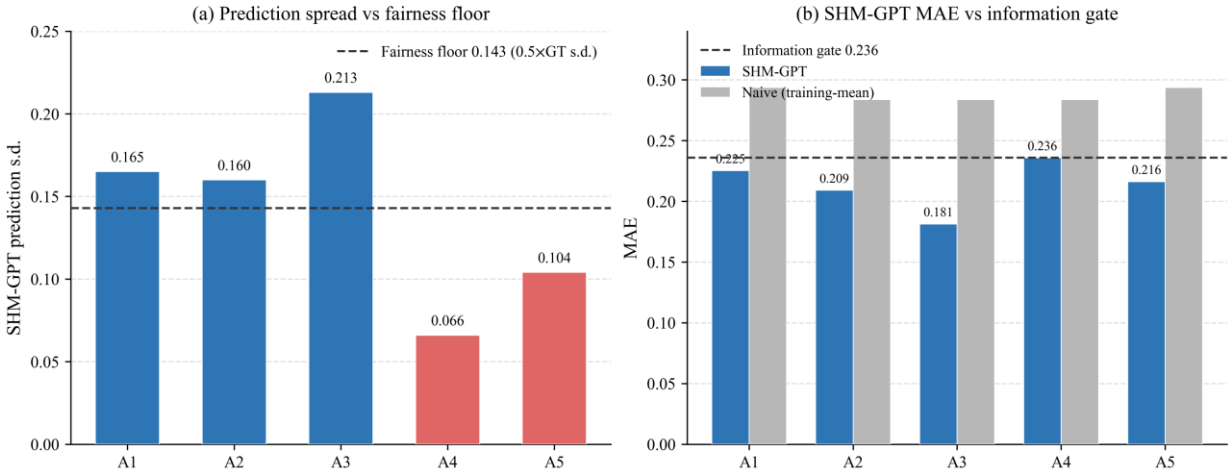

Fig. 8. Fairness diagnostic: prediction spread (a) and optimal-constant comparison (b).

Fig. 9 shows per-window trajectories for the three 8-kN test coupons. The predictions are noisy and do not reproduce the monotonic ground-truth progression throughout each coupon. SHM-GPT generally stays within the same broad range

as the recurrent and convolutional baselines and approaches high progress near the end of specimens 024 and 025, but specimen 020 retains a pronounced mid-life offset. Because the legacy cross-entropy head is not plotted in this figure, its collapse is assessed only in Table 2 and Figs. 10–11. Fig. 9 is therefore treated as a qualitative error-pattern diagnostic, not evidence of trajectory-wise tracking or monotonicity.

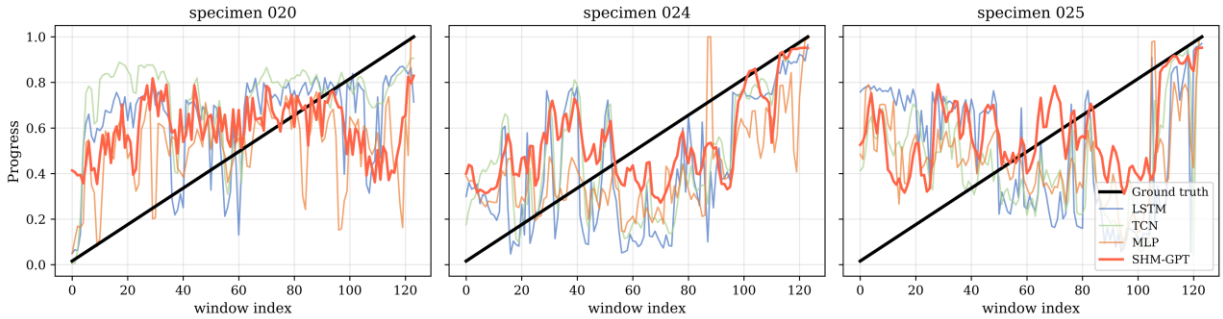

Fig. 9. Per-window prediction trajectories on three 8-kN coupons; the panel is used for qualitative error-pattern inspection.

4.4. Ablation Study

Five variants are evaluated on A1 with seed 42 (Table 2, Fig. 10). Replacing the Huber head with the 1024-bin cross-entropy head reduces prediction standard deviation from 0.165 to 0.033 and raises mean absolute error from 0.2309 to 0.2385. The remaining removals produce small, non-monotonic changes in this single run: 0.2256 without the Paris-law encoder, 0.2311 without the spectral-attention constraint, and 0.2312 without the variance-floor regulariser. Accordingly, only the head replacement provides direct evidence of collapse in this ablation; the table does not establish independent performance gains from every other component.

Table 2. Ablation study on scenario A1, seed 42.

Variant	MAE	Prediction std
Full SHM-GPT (Huber + LipsFormer + Paris-PI + variance floor)	0.2309	0.165
w/o Paris-law condition encoder	0.2256	0.168
w/o spectral-attention constraint	0.2311	0.165
w/o variance-floor regulariser	0.2312	0.166
head = cross-entropy-1024 (legacy)	0.2385	0.033

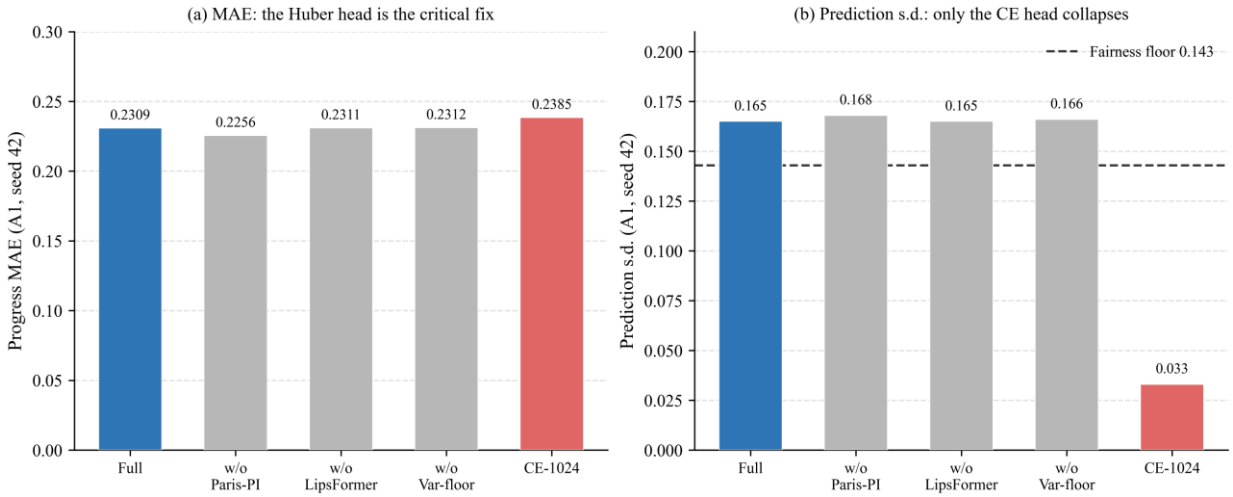

Fig. 10. Ablation of four design choices on scenario A1, seed 42: progress mean absolute error (a) and prediction standard deviation (b). Only the cross-entropy head collapses the prediction variance below the fairness floor.

The direct head comparison is already contained in Fig. 10: the Huber head yields mean absolute error 0.2309 and prediction standard deviation 0.165, whereas the 1024-bin cross-entropy head yields 0.2385 and 0.033. This supports the continuous progress head for A1 but does not establish a general ranking between regression and tokenised forecasting objectives.

The ablation supports a deliberately narrow interpretation. The Huber head is the only component whose replacement directly produces a low-spread output in this test. The slight improvement after removing the Paris-law encoder and the negligible changes from the other removals show that a single scenario and seed cannot rank those mechanisms. Additional scenarios and repeated runs would be required for component-level causal claims.

4.5. Hoare-Contract Agent Stress Test

The workflow is evaluated on an end-to-end trace from an 8-kN test coupon and on 50 synthetic traces containing both rule-conforming and randomly generated action/urgency pairs. The displayed trace reports progress 0.829 at window/state index 24 and returns OK. Across the synthetic set, the checker matches the verdict defined by the two Boolean policy rules on 50 of 50 traces. This result verifies the implementation against its test oracle; it does not validate the thresholds or certify maintenance safety.

Half of the synthetic traces use random actions and urgency levels, so both conforming and non-conforming states are exercised. For example, CONTINUE at progress 0.92 returns NOT OK, STOP at 0.31 returns OK, and REPAIR at 0.78 depends on urgency under the stated rule. Numeric-substring traceability is then checked against tool outputs and the two allowed policy constants. The test covers the implemented schema and parser, not all possible forms of numerical or semantic hallucination.

Separating deterministic fields from optional language rendering reduces one source of numerical inconsistency, but the renderer still requires independent validation for omissions, misleading wording, units,

and unsupported qualitative claims. Model replacement or quantisation would also require re-validation of prediction quality and interface compatibility. The reported checker should therefore be understood as a narrow software guard rather than a complete hallucination-control mechanism.

4.6. Limitations and Stress Tests

Two scenarios fail the spread diagnostic and are reported as stress tests with their principal limitations.

Scenario A4 is a small-data stress test in which two high-stress coupons yield about twenty-seven training windows. A Temporal Convolutional Network outperforms SHM-GPT (0.2128 versus 0.2358), and SHM-GPT prediction standard deviation is 0.066. The spectral constraint does not supply missing information or imply low error outside training coverage. Parameter-efficient adaptation and physics-guided synthetic data are possible future remedies, but they are not evaluated here.

Scenario A5 removes the acoustic-emission channel at inference. SHM-GPT has lower mean absolute error than the listed baselines, but its prediction standard deviation is 0.104 and every learned method falls below the spread floor. Training-time channel dropout, learned mask tokens, reliability-gated fusion, and full-to-partial modality distillation are plausible remedies that require separate experiments. No missing-modality robustness claim is made from the current result.

The eight-step-ahead stress test (Fig. 11) predicts progress eight windows ahead from a single 128-sample input window. SHM-GPT reaches mean absolute error 0.2585 versus 0.2612 for the Multi-Layer Perceptron, with prediction standard deviation 0.158 above the 0.134 spread floor. Its error nevertheless exceeds the information threshold of 0.232. Because the sliding-window stride is not documented in the modelling record, the eight-window horizon is not converted to seconds. Longer context and explicit crack-growth roll-out remain future work.

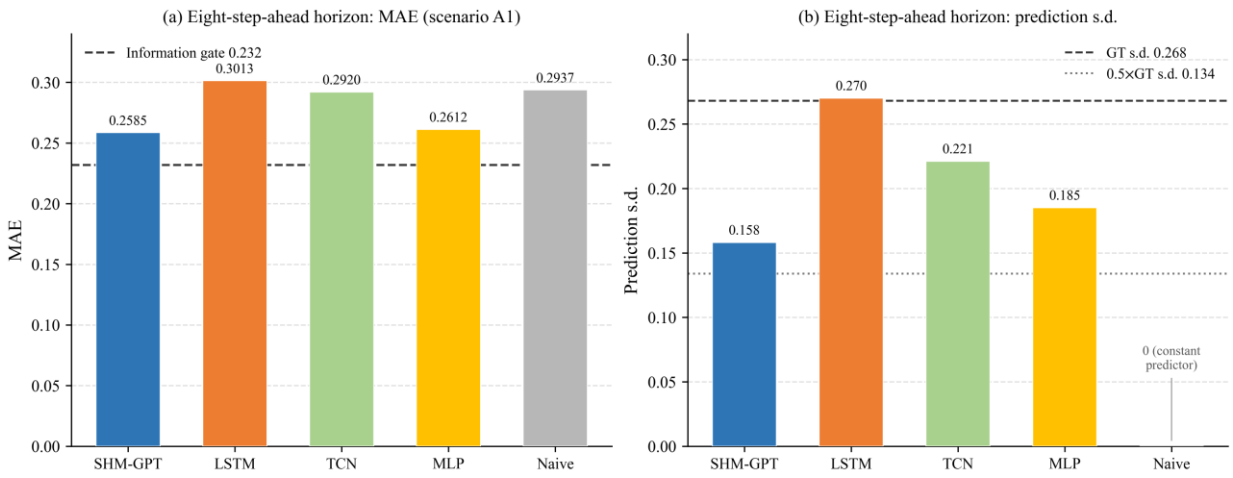

Fig. 11. Eight-step-ahead horizon stress test on scenario A1, seed 42: mean absolute error (a) and prediction standard deviation (b) across methods.

5. Conclusion

A compact physics-conditioned multimodal decoder and a contract-grounded four-role workflow have been evaluated for aircraft structural health monitoring. The continuous Huber progress head, rather than the legacy 1024-bin output, is the supported headline configuration. On the 15-coupon working set, SHM-GPT has the lowest mean absolute error in four of five scenarios and meets the descriptive spread and information diagnostics in three. The corresponding margins over the strongest listed baseline are 11.93%, 15.28%, and 3.01%. The deterministic checker also matches its rule-defined verdict on 50 of 50 synthetic traces.

The evidence remains benchmark-specific. A Temporal Convolutional Network is better in the two-coupon scenario; all learned methods show insufficient spread after acoustic-emission removal; the horizon test fails the information diagnostic; and only three random seeds are available. The current dataset is an in-house single-platform campaign rather than a broad pre-training corpus, preprocessing thresholds are incompletely documented, and the policy constants are illustrative. Multi-platform validation, stronger and more recent baselines, fully specified preprocessing, repeated component ablations, and operator-approved maintenance rules are required before deployment-oriented claims can be considered.

References

1. Y. Chen, D. Zhang, R. Yan, M. Xie, "Applications of Domain Generalization to Machine Fault Diagnosis: A Survey," IEEE/CAA Journal of Automatica Sinica 12, 1 (2025).

2. Z. Li, Y. Li, J. Lu, H. Zhu, Y. Zheng, et al., "Baseline-Free Assisted Lamb Wave-Based Damage Detection in CFRP Composites Using Graph Convolutional Networks and Transformer Models," Measurement 242, 116159 (2025).

3. Y. Wu, B. Sicard, S. A. Gadsden, "Physics-Informed Machine Learning: A Comprehensive Review on Applications in Anomaly Detection and Condition Monitoring," Expert Systems with Applications 255, C (2024).

4. H. Zhang, G. Wang, F. Teng, S. Lv, L. Zhang, et al., "Transformer Model Combining Cross-Attention and Self-Attention Guided by Damage Index for Pipeline Damage Localization Based on Helical Guided Waves," Mechanical Systems and Signal Processing 231, 112669 (2025), doi:10.1016/j.ymssp.2025.112669.

5. C. Li, S. Li, Y. Feng, K. Gryllias, F. Gu, M. Pecht, "Small Data Challenges for Intelligent Prognostics and Health Management: A Review," Artificial Intelligence Review 57, 8 (2024).

6. G. Amir, O. Maayan, T. Zelazny, G. Katz, M. Schapira, "Verifying the Generalization of Deep Learning to Out-of-Distribution Domains," Journal of Automated Reasoning 68, 3 (2024).

7. D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, et al., "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems," arXiv preprint arXiv:2405.06624 (2024).

8. A. Das, W. Kong, R. Sen, Y. Zhou, "A Decoder-Only Foundation Model for Time-Series Forecasting," in Proceedings of the 41st International Conference on Machine Learning (2024), pp. 10148–10167.

9. A. F. Ansari, L. Stella, C. Türkmen, X. Zhang, P. Mercado, H. Shen, et al., "Chronos: Learning the Language of Time Series," Transactions on Machine Learning Research 2024, 7 (2024).

10. M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, A. Dubrawski, "MOMENT: A Family of Open Time-Series Foundation Models," in Proceedings of the 41st International Conference on Machine Learning (2024), pp. 16115–16152.

11. Y. Liu, H. Zhang, C. Li, X. Huang, J. Wang, M. Long, "Timer: Generative Pre-trained Transformers Are Large Time Series Models," in Proceedings of the 41st International Conference on Machine Learning (2024), pp. 32287–32307.

12. G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, D. Sahoo, "Unified Training of Universal Time Series Forecasting Transformers," in Proceedings of the 41st International Conference on Machine Learning (2024), pp. 53140–53164.

13. Y. Liu, G. Qin, X. Huang, J. Wang, M. Long, "AutoTimes: Autoregressive Time Series Forecasters via Large Language Models," Advances in Neural Information Processing Systems 37 (2024).

14. X. Shi, S. Wang, Y. Nie, D. Li, Z. Ye, Q. Wen, M. Jin, "Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts," International Conference on Learning Representations 13, 1 (2025).

15. Y. Liang, H. Wen, Y. Nie, Y. Jiang, M. Jin, D. Song, et al., "Foundation Models for Time Series Analysis: A Tutorial and Survey," in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2024), pp. 6555–6565.

16. M. Arias Chao, C. Kulkarni, K. Goebel, O. Fink, "Aircraft Engine Run-to-Failure Dataset under Real Flight Conditions for Prognostics and Diagnostics," Data 6, 1 (2021).

17. Y. Liu, X. Peng, J. Cao, X. Wang, S. Deng, et al., "ToolGate: Contract-Grounded and Verified Tool Execution for Large Language Models," arXiv preprint arXiv:2601.04688 (2026).

18. X. Qi, J. Wang, Y. Chen, Y. Shi, L. Zhang, "LipsFormer: Introducing Lipschitz Continuity to Vision Transformers," in International Conference on Learning Representations (2023).

19. W. Ye, Y. Ma, X. Cao, K. Tang, "Mitigating Transformer Overconfidence via Lipschitz Regularization," Uncertainty in Artificial Intelligence 39, 1 (2023).

20. G. D. Plotkin, "A Structural Approach to Operational Semantics," Journal of Logic and Algebraic Programming 60–61, 1 (2004).

21. C. A. R. Hoare, "An Axiomatic Basis for Computer Programming," Communications of the ACM 12, 10 (1969).

22. W. Cai, Y. Liang, X. Liu, J. Feng, Y. Wu, "MSGNet: Learning Multi-Scale Inter-Series Correlations for Multivariate Time Series Forecasting," in Proceedings of the AAAI Conference on Artificial Intelligence (2024), pp. 11141–11149.

23. H. Jing, S. Yuan, J. Chen, Y. Meng, "A Multi-Scale Deep Residual Network-Based Guided Wave Imaging Evaluation Method for Fatigue Crack Quantification," Measurement Science and Technology 36, 015022 (2025).

24. S. Ma, K. A. Flanigan, M. Bergés, J. D. Brooks, "Transformer-Based Indirect Structural Health Monitoring of Rail Infrastructure with Attention-Driven Detection and Localization of Transient Defects," arXiv preprint arXiv:2510.07606 (2025).

25. M. M. Azad, H. S. Kim, "An Explainable Artificial Intelligence-Based Approach for Reliable Damage Detection in Polymer Composite Structures Using Deep Learning," Polymer Composites 46, 2 (2025).

26. Y. Wang, Y. Zeng, J. Tang, H. Tan, T. Sun, C. Wu, "Physics-Informed Learning of Fatigue Crack Growth in Corroded Steel: A Paris-Law Extension with Corrosion Degree and Stress Ratio Optimized by Bayesian Tuning," Theoretical and Applied Fracture Mechanics 143, 105456 (2026).

27. Q. Zhu, Z. Zhou, Y. Li, R. Yan, "Contrastive BiLSTM-Enabled Health Representation Learning for Remaining Useful Life Prediction," Reliability Engineering & System Safety 249, 110210 (2024).

28. F. Gan, H. Shao, B. Xia, "An Adaptive Model with Dual-Dimensional Attention for Remaining Useful Life Prediction of Aero-Engine," Knowledge-Based Systems 293, 1 (2024).

29. W. Wang, H. Song, S. Si, W. Lu, Z. Cai, "Data Augmentation Based on Diffusion Probabilistic Model for Remaining Useful Life Estimation of Aero-Engines," Reliability Engineering & System Safety 252, 1 (2024).

30. T. Yang, G. Li, K. Li, X. Li, Q. Han, "The LPST-Net: A New Deep Interval Health Monitoring and Prediction Framework for Bearing-Rotor Systems Under Complex Operating Conditions," Advanced Engineering Informatics 62, C (2024).

31. L. Yang, X. Zhang, F. Zhu, Z. Wang, X. Zhang, "Semi-Supervised Cross-Domain Fault Diagnosis via Contrastive Pre-Training and Annotation-Efficient Alignment Strategy," Journal of Industrial Information Integration 50, 101076 (2026).

32. D. Lee, J. G. Lee, M. Choi, C. Park, C.-W. Kim, et al., "Multi-Fidelity Sub-Label-Guided Transfer Network with Physically Interpretable Synthetic Datasets for Rotor Fault Diagnosis," Engineering Applications of Artificial Intelligence 148, 110467 (2025), doi:10.1016/j.engappai.2025.110467.

33. L. Wang, Z. Zhu, X. Zhao, "Dynamic Predictive Maintenance Strategy for System Remaining Useful Life Prediction via Deep Learning Ensemble Method," Reliability Engineering & System Safety 245, 1 (2024).

34. A. Gholamzadeh Khoee, Y. Yu, R. Feldt, "Domain Generalization Through Meta-Learning: A Survey," Artificial Intelligence Review 57, 10 (2024).

35. J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, H.-I. Suk, "Domain Generalization for Medical Image Analysis: A Review," Proceedings of the IEEE 112, 10 (2024).

36. L. Kirchdorfer, C. Elich, T. Kutsche, H. Stuckenschmidt, L. Schott, J. M. Köhler, "Analytical Uncertainty-Based Loss Weighting in Multi-Task Learning," in Pattern Recognition (DAGM GCPR 2024), Lecture Notes in Computer Science 15297 (2024).

37. S. Yu, X. Li, Y. Lei, B. Yang, N. Li, K. Feng, "Multimodal Data-enabled Large Model for Machine Fault Diagnosis Towards Intelligent Operation and Maintenance," Journal of Industrial Information Integration 50, 101061 (2026).

38. W. Zhang, X. Chang, H. Hao, Y. Zhang, X. Li, F. Zhu, "State of Charge Prediction for Electric Aircraft Lithium-Ion Batteries with Limited Data by Stochastic Quantization and Informer Network," Green Energy and Intelligent Transportation, 100437 (2026).

39. X. Chen, X. Li, Y. Lei, B. Yang, N. Li, K. Feng, "Neuromorphic Computing-Enabled Multimodal Data Fusion for Intelligent Machine Fault Diagnosis," Journal of Industrial Information Integration 51, 101108 (2026).

40. X. Li, W. Fan, S. Yang, W. Zhang, X. Li, "Flexible Federated Learning in Machinery Fault Diagnostics With Light Communication," IEEE/CAA Journal of Automatica Sinica 13, 3 (2026), pp. 680–691.

41. W. Zhang, Z. Wang, X. Li, S. Xiang, "Physics-integrated intelligent method for propeller aerodynamic property predictions of electric aircraft," Engineering Applications of Artificial Intelligence 173, 114408 (2026).

42. S. Yang, L. Ling, X. Li, J. Han, S. Tong, "Industrial battery state-of-health estimation with incomplete limited data toward second-life applications," Journal of Dynamics, Monitoring and Diagnostics 3, 4 (2024), pp. 246–257.

43. W. Zhang, H. Hao, Y. Zhang, H. Yang, X. Li, "State of charge prediction for lithium-ion batteries in electric aircraft based on self-supervised informer," Applied Soft Computing 186, 114283 (2026).