CONTEXT-INFORMED SEQUENCE CLASSIFICATION: A MULTIMODAL APPROACH TO VEHICLE DIAGNOSTICS

Hugo Math
BMW Group, Munich, Germany
University of Augsburg, Augsburg, Germany
hugo.math@bmwgroup.com

Rainer Lienhart
Department of Computer Science
University of Augsburg, Augsburg, Germany
rainer.lienhart@uni-augsburg.de

ABSTRACT

Effective vehicle diagnostics are critical for safety and predictive maintenance, but often rely solely on asynchronous discrete sequences of Diagnostic Trouble Codes (DTCs), overlooking valuable environmental context. This paper introduces BiCarFormer, a multimodal bidirectional Transformer that fuses DTC sequences with tokenized sensory data (temperature, pressure, humidity) via a co-attention mechanism and multimodal embeddings. By integrating these heterogeneous modalities, BiCarFormer addresses the complexity and noise inherent in real-world automotive data. Evaluations on a large-scale fleet dataset of 22,137 error codes and 360 error patterns demonstrate that our approach significantly outperforms single-modality baselines. We also show that in this setting, Transformers can learn the fluctuation of quantized continuous values through attention.

Track: Industry & Applications

1 INTRODUCTION

Modern vehicles generate complex, irregular event sequences known as Diagnostic Trouble Codes (DTCs) Pirasteh et al. (2019). While these discrete codes provide a structured log of system states, relying on them in isolation captures only a partial picture of vehicle health. Crucial context, such as continuous environmental sensory data (e.g., temperature, pressure, voltage), is often ignored in standard diagnostic pipelines due to its high dimensionality and noise. Yet, for domain experts, this environmental context is often the deciding factor in distinguishing between critical system failures, referred to as Error Patterns (EPs)[1] Math et al. (2025b); Math & Lienhart (2025; 2026). Existing data-driven approaches typically treat vehicle diagnostics as a unimodal sequence prediction task using RNNs or Transformers Hafeez et al. (2021; 2024); Math et al. (2025a). Moreover, time series foundation models (TSFMs) such as Chronos Ansari et al. (2024) operate on one modality and on a continuous event sequence. Vehicle data is intrinsically multimodal, consisting of continuous signals (voltage, RPM) interleaved with discrete event codes (DTCs), a modality mix that purely continuous TSFMs struggle to integrate.

To bridge this gap, we introduce BiCarFormer, a multimodal bidirectional Transformer Vaswani et al. (2017) designed for industrial-scale diagnostics. BiCarFormer employs a co-attention mechanism to fuse discrete DTC embeddings with environmental features, allowing the model to learn context-aware representations of vehicle behavior. Evaluated on a real-world fleet dataset containing 22,137 unique codes and 360 error patterns, our approach demonstrates that integrating environmental context significantly improves multi-label classification performance over unimodal baselines.

2 PROBLEM FORMULATION: ASYNCHRONOUS MULTI-MODAL STREAMS

We model the vehicle state as a realization of two non-stationary stochastic processes. Let $S^{dtc} = \{(c_i, t_i, m_i)\}_{i=1}^{L}$ be the sequence of discrete diagnostic events, where c_i represents the categorical

[1] To clarify the distinction: a DTC is a raw, localized symptom generated by a control unit (e.g., 'P0562: System Voltage Low'), whereas an Error Pattern (EP) represents the actual logical failure diagnosed by engineers (e.g., 'Alternator output degraded requiring replacement').

error code, and $t_i, m_i \in \mathbb{R}^+$ represent the absolute timestamp and mileage. Simultaneously, we observe a sequence of environmental conditions $S^{env} = \{(e_j, v_j, u_j, t'_j)\}_{j=1}^{L_{env}}$, where each tuple consists of a descriptor e_j (e.g., "Battery Voltage"), a scalar value v_j, and a unit u_j. Crucially, S^{dtc} and S^{env} are unaligned ($t_i \neq t'_j$) and sampled at different rates ($L_{env} \gg L$); therefore, we truncate the sequences to maximum lengths of $L = 256$ and $L_{env} = 4096$, respectively. Our objective is to map the joint history $\{S^{dtc}, S^{env}\}$ to a set of binary Error Patterns $\mathbf{y} \in \{0,1\}^K$. The cardinality of the data is given in Table 2, and a temporal point process representation is shown in Fig. 1.

2.1 BiCarFormer Architecture

We introduce **BiCarFormer**, a dual-stream Bidirectional Transformer that learns intra-modal dependencies via co-attention (Fig. 2).

Hierarchical DTC Embedding. A single code c_i is a composition of three hierarchical features: the Electronic Control Unit (c^{ecu}), the Base-DTC (c^{base}), and the Fault-Byte (c^{byte}). To preserve the semantic hierarchy of the 22,137 unique codes without exploding the vocabulary, we employ hierarchical embedding fusion rather than flattening:

$$\mathbf{h}_i^{dtc} = [\mathbf{E}_{ecu}(c_i^{ecu}) \parallel \mathbf{E}_{base}(c_i^{base})] + \mathbf{E}_{byte}(c_i^{byte}) \tag{1}$$

where $\parallel$ denotes concatenation and $\mathbf{E}$ are learnable lookup tables. To capture the irregular spatio-temporal dynamics, we augment the DTC embedding with continuous sinusoidal embeddings $\Phi(\cdot)$ Zuo et al. (2020) for time and mileage:

$$\mathbf{u}_i = \mathbf{h}_i^{dtc} + [\Phi(t_i) \parallel \Phi(m_i)] \tag{2}$$

Tokenized Environmental Embedding. Continuous sensor data is often noisy and high-dimensional. We tokenize S^{env} by discretizing the scalar values v_j into quantiles Greenwald & Khanna (2001). This projects the continuous dynamics onto the same discrete manifold as the DTC vocabulary, enabling the Transformer to perform cross-modal attention. The environmental embedding $\mathbf{h}_j^{env}$ is formed by concatenation of the description and the quantized value bucket, then summed with the unit:

$$\mathbf{e}_j = [\mathbf{E}_{desc}(e_j) \parallel \mathbf{E}_{val}(\text{Quantize}(v_j))] + \mathbf{E}_{unit}(u_j) \tag{3}$$

This approach allows the Transformer to attend to specific value ranges (e.g., "High Temp") as distinct unified tokens and optimize the embedding space (see Fig. 2 for the proportions).

Co-Attention Mechanism. To address the asynchrony between S^{dtc} and S^{env}, we employ a co-attention mechanism Lu et al. (2019) in which the diagnostic stream queries the environmental context, and vice versa. We utilize Rotary Positional Embeddings (RoPE) Su et al. (2024) to inject relative token position. The cross-modal context vector $\mathbf{c}_i$ for the i-th DTC is computed as:

$$\mathbf{c}_i = \sum_{j=1}^{L_{env}} \text{Softmax}\left(\frac{\mathcal{R}_{\theta_u}(\mathbf{Q} \cdot \mathbf{u}_i)\mathcal{R}_{\theta_e}(\mathbf{K} \cdot \mathbf{e}_j)^T}{\sqrt{d}}\right)\mathbf{V} \cdot \mathbf{e}_j \tag{4}$$

where $\mathcal{R}$ is the rotary application with a corresponding θ. This allows specific faults to "attend" to relevant environmental precursors (e.g., a voltage drop preceding a communication fault).

3 Experiments

Dataset. We evaluate BiCarFormer on a large-scale industrial fleet dataset comprising 22,137 unique DTC types and 360 target Error Patterns (EPs) in 50,000 test samples. The dataset exhibits significant class imbalance and irregular sampling intervals, characteristic of real-world data.

Emergent Fluctuation Detection. Beyond classification accuracy, we analyze the interpretability of the learned representations. Fig. 3 visualizes the co-attention scores assigned by DTC queries to the environmental stream. Remarkably, the attention mechanism does not merely correlate with absolute values but explicitly activates at 'trigger points' of significant variance (e.g., sudden voltage drops). This suggests that BiCarFormer implicitly learns to detect value fluctuations purely from quantized tokens, effectively acting as a soft edge detector for environmental anomalies.

Table 1: **Downstream evaluation of multi-label EP classification.** All models use 25M parameters. **BiCarFormer** consistently outperforms uni-modal baselines. Notably, the significant gap in **Macro F1 (0.71 vs 0.63)** demonstrates our model's superior ability to detect **rare error patterns**, whereas baselines struggle with the class imbalance.

Model	AUROC (Micro)	F1 Score (Micro)	F1 Score (Macro)
BiCarFormer	**0.809**	**0.77**	**0.71**
DTC-TranGRU Hafeez et al. (2024)	0.602	0.36	0.28
BERT Devlin et al. (2019)	0.768	0.71	0.63

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024. URL https://arxiv.org/abs/2403.07815.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Michael Greenwald and Sanjeev Khanna. Space-efficient online computation of quantile summaries. SIGMOD Rec., 30(2):58–66, May 2001. ISSN 0163-5808. doi: 10.1145/376284.375670. URL https://doi.org/10.1145/376284.375670.

Abdul Basit Hafeez, Eduardo Alonso, and Aram Ter-Sarkisov. Towards sequential multivariate fault prediction for vehicular predictive maintenance. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1016–1021, 2021. doi: 10.1109/ICMLA52953.2021.00167.

Abdul Basit Hafeez, Eduardo Alonso, and Atif Riaz. Dtc-trangru: Improving the performance of the next-dtc prediction model with transformer and gru. Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, 2024. URL https://api.semanticscholar.org/CorpusID:269951398.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Curran Associates Inc., Red Hook, NY, USA, 2019.

Hugo Math and Rainer Lienhart. Towards practical multi-label causal discovery in high-dimensional event sequences via one-shot graph aggregation. In NeurIPS 2025 Workshop on Structured Probabilistic Inference & Generative Modeling, 2025. URL https://openreview.net/forum?id=1HZfpuDVeW.

Hugo Math and Rainer Lienhart. Trace: Scalable amortized causal discovery from single sequences via autoregressive density estimation, 2026. URL https://arxiv.org/abs/2602.01135.

Hugo Math, Rainer Lienhart, and Robin Schön. Harnessing event sensory data for error pattern prediction in vehicles: A language model approach. Proceedings of the AAAI Conference on Artificial Intelligence, 39(18):19423–19431, 4 2025a. doi: 10.1609/aaai.v39i18.34138. URL https://ojs.aaai.org/index.php/AAAI/article/view/34138.

Hugo Math, Robin Schön, and Rainer Lienhart. One-shot multi-label causal discovery in high-dimensional event sequences. In NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science, 2025b. URL https://openreview.net/forum?id=z7NT8vGWC2.

Parivash Pirasteh, Slawomir Nowaczyk, Sepideh Pashami, Magnus Löwenadler, Klas Thunberg, Henrik Ydreskog, and Peter Berck. Interactive feature extraction for diagnostic trouble codes in predictive maintenance: A case study from automotive domain. In Proceedings of the Workshop on Interactive Data Mining, WIDM'19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362962. doi: 10.1145/3304079.3310288. URL https://doi.org/10.1145/3304079.3310288.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomput., 568(C), 3 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes process. In Hal Daumé III and Aarti Singh (eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pp. 11692–11702. PMLR, 7 2020. URL https://proceedings.mlr.press/v119/zuo20a.html.

A Appendix

A.1 Data Characteristics

Table 2: Feature Space Statistics and Notation.

Feature	**Symbol**	**# Distinct Values**
Diagnostic Stream S^{dtc}		
Full DTC	c	22,137
Electronic Control Unit	c^{ecu}	132
Base-DTC	c^{base}	17,044
Fault-Byte	c^{byte}	2
Environmental Stream S^{env}		
Sensor Description	e	2,559
Quantized Value	v	3,288
Measurement Unit	u	18

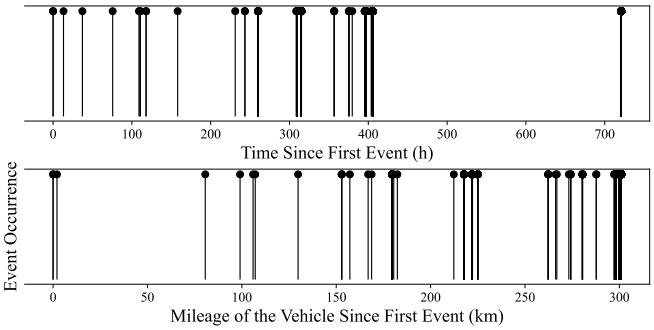

Figure 1: **Temporal and spatial point process representation of events from a vehicle**. Bold vertical lines indicate multiple events happening at the same time t_i or mileage m_i.

A.2 Figures

A.2.1 Architecture

A.2.2 Attention Scores Interpretation

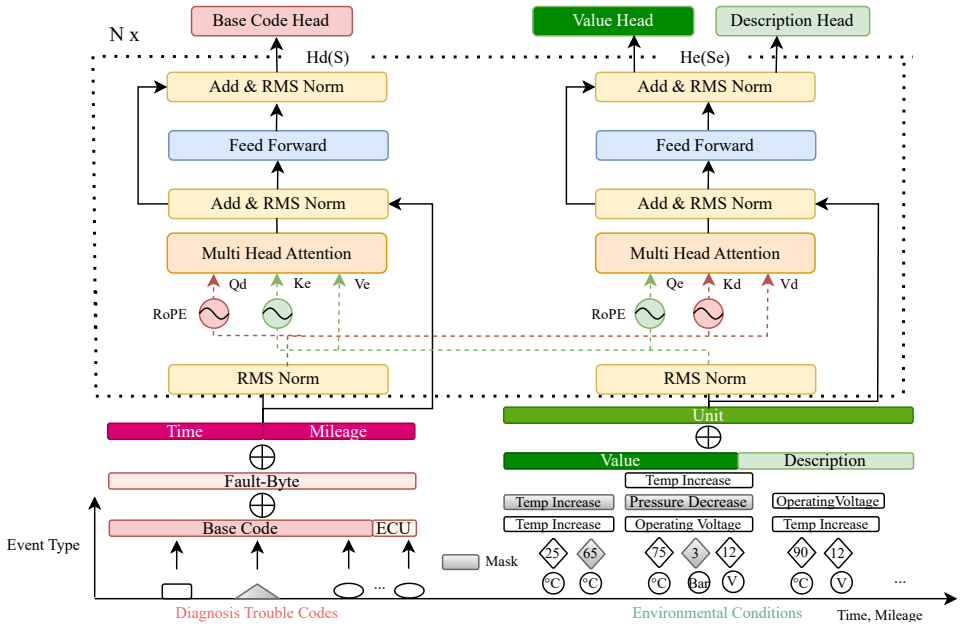

Figure 2: **BiCarFormer architecture with multimodal masking and embedding fusion**. Both parallel transformers are computing cross-attention scores conditioned on each modality $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$. Two final representations are generated for each modality: DTC ($\boldsymbol{H}_d$) and e. conditions ($\boldsymbol{H}_e$). Multiple embeddings are defined at the input level to take into account token-specific features.

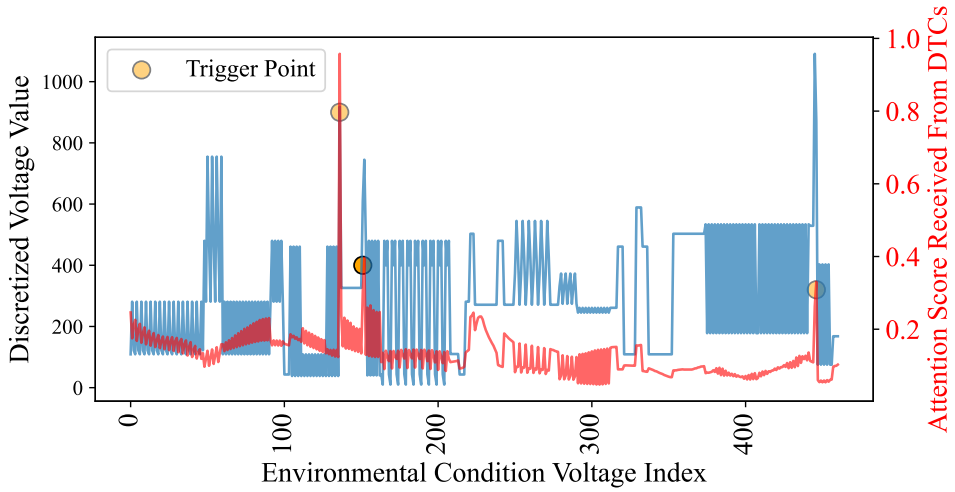

Figure 3: Co-attention scores (orange) spike precisely during sudden fluctuations in quantized voltage (blue), demonstrating the model's ability to detect environmental instability.

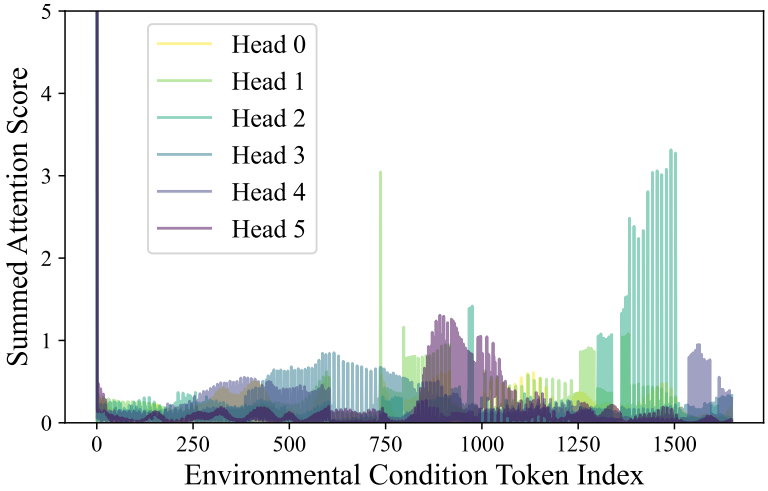

Figure 4: **Amount of attention received by each environment condition from the DTCs**. The y-axis was truncated to improve clarity as well as the number of heads printed. We take $\boldsymbol{A}_{dtc \to env}$ of the last layer.

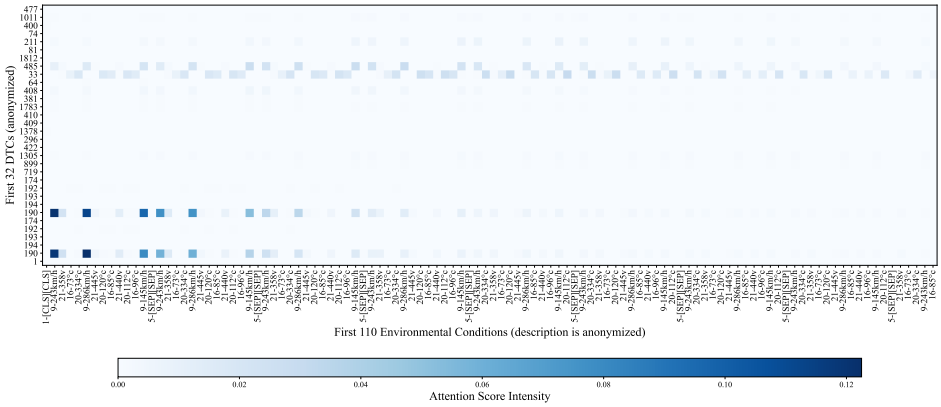

Figure 5: Cross Attention Scores for $\boldsymbol{A}_{dtc \to env}$. The DTCs are shown on the y-axis (anonymized), and the environmental conditions with their 3 elements (d, v, u) concatenated are shown on the x-axis (the description d is anonymized). The intensity of each cell reflects the attention weight, where darker shades indicate higher attention values.