KNOGML: KNOWLEDGE-GROUNDED MULTIMODAL LEARNING FOR FAULT DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

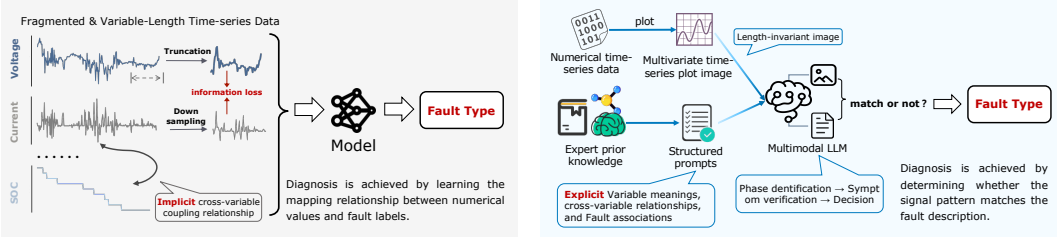
Reliable fault diagnosis is essential to industrial safety and often relies on fragmented, variable-length time series. Many existing methods achieve diagnosis by learning a mapping from numerical features to fault labels, but face two key limitations: 1) they typically rely on lossy operations such as truncation or downsampling to accommodate fixed-length model inputs, limiting the use of complete recorded segments; 2) the domain-specific physical meaning of signals and their relationships is not explicitly utilized. To this end, we propose KnoGML, a knowledge-grounded multimodal learning framework that combines multivariate plot images with structured diagnostic priors for fault diagnosis. Specifically, domain-specific fault knowledge is organized into structured prompts describing variable meanings, cross-variable relationships, and fault associations within a three-stage diagnostic protocol comprising phase identification, symptom verification, and final decision. Meanwhile, multivariate time-series data are converted into fixed-resolution grid images, providing a unified visual representation for time-series segments of varying lengths. Building on this, a multimodal large language model jointly processes the structured prompts and visual representation, transforming fault diagnosis into a matching problem between fault descriptions and signal patterns. Finally, a classification head produces a deterministic fault label, providing a fixed output interface for industrial integration. Across two battery chemistries, KnoGML outperforms all evaluated baselines in both Macro-F1 and Weighted-F1, achieving Macro-F1 scores of 90.24% and 87.61%. Ablations demonstrate the diagnostic value of explicit fault associations beyond signal descriptions alone. Code is available at https://anonymous.4open.science/r/KnoGML-77AB/.

1 INTRODUCTION

Reliable fault diagnosis is essential for ensuring industrial system integrity, production continuity, and safe and stable operation of industrial systems. With the development of the Industrial Internet of Things (IIoT), an increasing number of sensors are deployed to monitor the operating status of industrial systems, and the multivariate time-series data thereby recorded have become a primary source for analyzing industrial system faults. Such data have two characteristics closely related to diagnosis. First, actual industrial tasks and their occurrence times are stochastic; the collected data are often not complete, standardized operating cycles but rather segments with fragmented coverage and variable duration. Second, each signal carries explicit physical meaning: different trends of the same variable may correspond to different fault mechanisms, and coordinated changes among multiple variables must also be interpreted together with their physical meanings and operating phase to determine the fault type. For example, in electric vehicle batteries, a voltage drop may arise from normal discharge or faults such as internal short circuits; therefore, it should not be judged in isolation but assessed together with relationships such as current direction, state-of-charge range, and temperature variation (Hong et al., 2024; Zhang et al., 2023). Therefore, how to utilize the observational information in segments and analyze it together with the physical meanings of signals is an important problem for achieving reliable fault diagnosis.

Benefiting from the representation learning and temporal modeling capabilities of deep learning, data-driven methods have become an important research direction in fault diagnosis. Such methods directly learn features from time series and map them to fault categories, reducing dependence

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Numerical feature–label learning (b) Knowledge-grounded multimodal learning

Figure 1: Fault diagnosis from fragmented, variable-length time series. (a) A numerical pipeline adapts segments to fixed-length inputs and learns feature–label mappings. (b) KnoGML jointly processes multivariate plot images and structured diagnostic prompts to match fault descriptions with signal patterns and predict a fault label.

on manual feature design (Jia et al., 2016). For example, convolutional neural networks can extract fault-related patterns from raw vibration signals (Zhang et al., 2017), while spatiotemporal Transformers perform battery fault diagnosis by modeling dependencies among multivariate measurements (Zhao et al., 2023). However, they still have two fundamental limitations, as shown in Figure 1(a). **(1) Incomplete information usage.** Most models are constrained by a fixed input window preset during training. For variable-length segments, common practices are truncation, downsampling, or padding. Truncation discards observations outside the window, while downsampling can suppress brief changes. Padding with an appropriate mask preserves the recorded values, but does not itself specify how to compare segments with different durations. The limitation therefore concerns how a pipeline handles variable-length observations, rather than a universal restriction of numerical models. **(2) Unclear physical relationships.** Class-label supervision can teach models statistical relationships among signals, but does not explicitly state the physical conditions under which these relationships support a fault. Learning such associations therefore depends on the coverage of the training data. When fault samples are scarce or operating conditions change, such fitting becomes particularly difficult (Lei et al., 2023), causing normal changes to be mistaken for fault patterns (Zhang et al., 2023).

The above limitations indicate that fault diagnosis requires not only suitable signal representations but also interpretable and actionable diagnostic knowledge. Pretrained multimodal large language models jointly process images and text and have demonstrated the ability to analyze plots (Bai et al., 2025), offering a feasible path to address these needs. Multivariate plot images can present recorded segments of different lengths in a unified visual format, while text can directly describe signal meanings and their associations with faults. Prior studies have also shown that model-based diagnosis and knowledge-driven learning demonstrate the value of physical relationships in fault judgment (Isermann, 2005; Lei et al., 2023).

Motivated by this, we propose **KnoGML**, a **knowledge-grounded multimodal learning** framework for fault diagnosis, as shown in Figure 1(b). KnoGML converts multivariate time series into fixed-resolution grid images covering the recorded segment, accommodating segments of different lengths in a unified visual format; meanwhile, it organizes fault knowledge into structured prompts following a three-stage protocol of *phase identification*, *symptom verification*, and *final decision*. This protocol guides the model to first identify the operating phase, then inspect relevant signal patterns, and finally associate them with candidate faults. A pretrained multimodal large language model jointly processes the prompts and images, learning to match fault descriptions with signal patterns during classification training; low-rank adaptation (LoRA) updates the model in a parameter-efficient manner, and a classification head outputs a deterministic fault label.

- We reformulate fault diagnosis as a knowledge-guided matching problem between fault descriptions and signal patterns. Structured diagnostic priors and multivariate plot images bring domain knowledge and variable-length observations into a shared multimodal classification framework.
- We develop a three-stage diagnostic prompt that organizes variable meanings, cross-variable relationships, and fault associations around operating phases. This provides task-specific guidance to a pretrained multimodal model, adapted through LoRA with a discriminative classification head.

- KnoGML outperforms all evaluated baselines in both Macro-F1 and Weighted-F1 across two battery chemistries. Ablations demonstrate the value of explicit fault associations and visual input, while limited-data results show the highest Weighted-F1 across all evaluated training fractions.

2 Related Work

Fault diagnosis and domain knowledge. Model-based diagnosis relates measurements to physical behavior through residuals, parameter estimates, and consistency checks (Isermann, 2005). Data-driven methods learn representations directly from observations, including convolutional features for vibration signals (Jia et al., 2016; Zhang et al., 2017) and temporal dependencies or input–response relationships for battery measurements (Zhao et al., 2023; Zhang et al., 2023). These approaches offer complementary strengths. Physical relationships make engineering knowledge explicit, while learned representations capture patterns that are difficult to specify analytically. Prior-knowledge-embedded learning further illustrates how domain knowledge can support diagnosis with few fault samples and changing operating conditions (Lei et al., 2023). Numerical models can learn signal dependencies, but label supervision alone does not provide their physical interpretation. KnoGML supplies this information through structured diagnostic prompts, without requiring a separate physical model for each fault. Its three-stage protocol organizes the use of this knowledge around operating phases, signal patterns, and candidate faults.

Time-series representations and multimodal learning. Time-series models use diverse representations, including temporal two-dimensional variation in TimesNet (Wu et al., 2023) and attention-based features for variable-length diagnostic inputs (Kim et al., 2019). Fixed-length formatting is therefore a property of particular pipelines rather than a universal requirement. Language-based approaches include Time-LLM, which reprograms time-series inputs for a pretrained language model (Jin et al., 2024), and ChatTS, which supports textual interaction with time-series representations (Xie et al., 2025). Visual approaches range from time-series image encodings (Wang & Oates, 2015) to anomaly analysis using multimodal models (Zhou & Yu, 2025). Although pretrained models can analyze plots (Bai et al., 2025), recognizing a curve pattern does not establish its diagnostic meaning. Zhou and Yu find that images improve time-series anomaly detection, while explicit reasoning prompts do not improve performance (Zhou & Yu, 2025). Knowledge-guided methods such as FD-LLM and KG-SR-LLM also connect diagnostic information with language (Lin et al., 2025; Xiao et al., 2025). KnoGML combines multivariate plot images with phase-dependent fault descriptions and adapts the model for categorical diagnosis. The image provides a common format for variable-length records, while the prompt specifies how signal patterns relate to faults; finite image resolution still limits the visible detail.

3 Method

KnoGML combines multivariate plot images with structured diagnostic priors to classify fragmented, variable-length time series. The framework takes domain knowledge as text, so its input format and learning objective do not depend on a particular industrial system. We instantiate it for electric vehicle batteries, whose voltage, current, state of charge (SOC), and temperature provide coupled observations across charging and discharging. The battery setting supplies the signal definitions, fault descriptions, and label vocabulary used below; the full prompt is given in Appendix A.2.

3.1 Problem formulation

Let $X = \{(t_j, \mathbf{x}_j)\}_{j=1}^{T}$ be a segment with timestamps t_j and V monitored variables $\mathbf{x}_j \in \mathbb{R}^V$. The number of observations T varies across segments. Given fault knowledge $\mathcal{K}$ and a label space $\mathcal{Y}$, we construct a plot image $I = g(X)$ and a structured diagnostic prompt $P = h(\mathcal{K}, c)$, where c contains the context specified by the prompt template. The model learns

$$p(y \mid X, \mathcal{K}, c) = \mathrm{softmax}(W_c F_{\theta,\phi}(I, P) + b_c)\,, \tag{1}$$

where F is a pretrained multimodal backbone with frozen parameters θ and trainable LoRA parameters ϕ. The classification head (W_c, b_c) maps its final-layer representation to $\mathcal{Y}$, with four output units for the battery task. Figure 2 summarizes this knowledge-guided matching between fault descriptions and signal patterns.

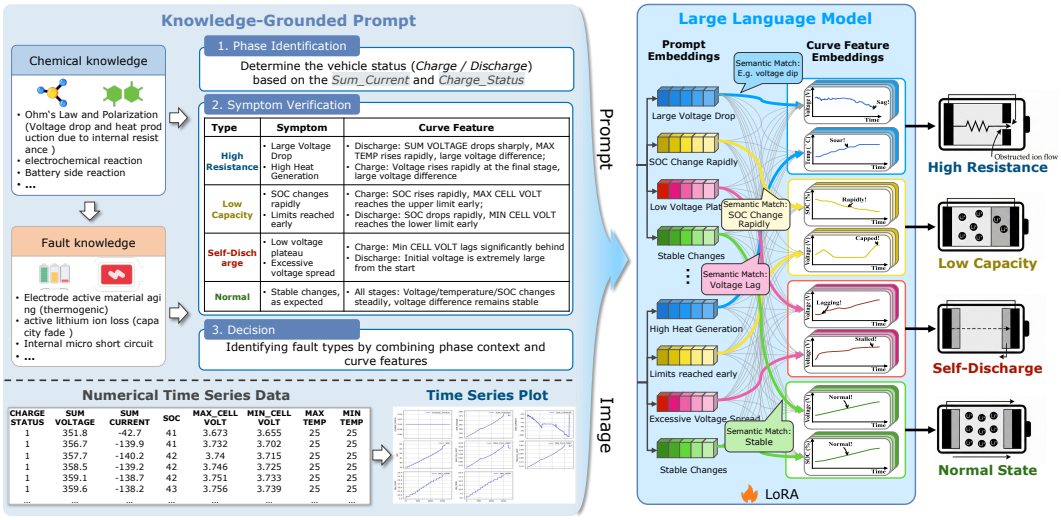


Figure 2: KnoGML instantiated for battery fault diagnosis. A structured diagnostic prompt and a multivariate plot image jointly enter a pretrained multimodal model. LoRA adapts the model for knowledge-guided matching, and a classification head returns one fault label. Connections between descriptions and curves illustrate the matching task, not measured attention.

3.2 Structured Diagnostic Priors

The fault knowledge draws on domain expert experience and basic physical principles. The former provides practical associations between operating phases, signal changes, and candidate faults; the latter explains why the variables should be considered together. For batteries, these include the dependence of terminal voltage on current and battery state, the relation between charge transfer and SOC, and the connection between electrical losses and heating. This knowledge provides qualitative diagnostic guidance without specifying complete physical models or universal thresholds.

To construct the prompt, we first state the meaning of each measured variable and the conventions needed to read it. We then organize the fault descriptions by operating phase, identifying the relevant signals, expected changes, and cross-variable relationships. Finally, we combine these descriptions with the diagnostic task and allowed labels in a shared template. The resulting text supplies *structured diagnostic priors* that supplement class-label supervision. The content can be revised for another application while retaining the same organization and model interface.

The template follows a three-stage diagnostic protocol. *Phase identification* establishes the operating phase from the signals and input conventions. *Symptom verification* guides the model to inspect relevant slopes, amplitudes, timing, and joint changes within that phase. *Final decision* relates the observed patterns to the candidate faults and specifies one allowed label. For example, the battery prompt directs attention to voltage response during changes in current, together with cell-voltage spread and temperature, when assessing high resistance. This connects a fault description to several observable signals rather than treating a voltage change in isolation.

The complete prompt contains context information, variable definitions, the diagnostic task, phase-specific fault descriptions, and the output schema. Its fault descriptions are shared across samples; contextual fields specify battery type, vehicle identifier, and operating status. The stages organize the input instructions; class-label training does not establish that the model executes them as three independently verified operations. No intermediate phase labels, symptom annotations, or generated reasoning traces are required. Appendix A.2 gives the complete battery template, including its qualitative comparisons and current-sign convention.

3.3 Multivariate Plot Images

Each segment is rendered as a fixed-resolution grid image. Subplots show individual signals against a common temporal reference, with separate vertical axes in their respective units. For batteries,

the eight panels contain charge status, pack voltage, pack current, SOC, and maximum and minimum cell voltages and temperatures. The fixed panel order lets the prompt refer to each variable consistently, while the common time reference supports comparisons across signals. For example, a voltage change and a current change can be examined over the same part of a segment. The image thus shows the observations, while the prompt describes their possible diagnostic meaning.

The default image is $I \in \mathbb{R}^{448\times448\times3}$ regardless of T. Rendering the recorded interval therefore avoids selecting a fixed-length prefix solely to satisfy the model interface. This design establishes full-interval coverage, not superior information retention over every numerical representation. This is a unified visual representation, not a lossless encoding. Finite resolution can obscure small changes or closely spaced observations; its effect is examined in Appendix B.

3.4 Knowledge-guided matching and adaptation

The pretrained model jointly processes the image and prompt. Its attention layers combine curve features with textual descriptions of phase-dependent patterns and fault associations. Classification training adapts the resulting representation to distinguish candidate faults. We use knowledge-guided matching to describe this diagnostic formulation, not an explicitly supervised alignment between individual rules and image regions. Knowledge-guided matching is learned within the multimodal backbone through the class-label objective; no separate similarity module or contrastive objective is introduced.

LoRA (Hu et al., 2022) adapts the query and value projections of the language attention layers. For a frozen projection $W_0 \in \mathbb{R}^{d\times k}$,

$$Wx = W_0x + \frac{\alpha}{r}BAx, \qquad A \in \mathbb{R}^{r\times k}, \quad B \in \mathbb{R}^{d\times r}, \tag{2}$$

where r is the adaptation rank and α controls the update scale. The vision encoder and original language-model weights stay frozen; only the LoRA parameters and classification head are updated. Training minimizes class-weighted cross-entropy over LoRA and classification-head parameters,

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log p(y_i \mid I_i, P_i), \tag{3}$$

where w_{y_i} denotes the weight of class y_i. Supervision is applied to fault labels, without losses on intermediate stages or textual explanations.

3.5 Constrained diagnostic output

At inference, $\hat{y} = \operatorname{argmax}_{y\in\mathcal{Y}} p(y \mid I, P)$ returns exactly one allowed category. For batteries, IDs 0–3 correspond to high resistance, low capacity, normal, and self-discharge. The classification head enforces this finite output space without free-form generation or text parsing, giving a deterministic decision for fixed model scores.

The framework provides control over both the task specification and the output interface. Structured prompts make the diagnostic guidance explicit and editable; the classification head guarantees a fixed categorical format for industrial integration. Prompt guidance remains distinct from an enforced physical rule, and the returned label is not itself a verified explanation.

4 Experiments

4.1 Experimental setup

Dataset. CH-BatteryGen is a semi-synthetic benchmark generated from real operating-data distributions with controlled fault injection (Liu et al., 2026). We evaluate lithium iron phosphate (LFP) and nickel–cobalt–manganese (NCM) batteries separately.

Data preparation. Each charging or discharging segment is one sample. The eight channels, sampled at 0.1 Hz, share a time axis and retain separate physical-unit axes in a 448×448 grid image (Section 3.3). This image and the structured diagnostic prompt form the model input. Segment

length counts data rows in one CSV, with consecutive rows separated by 10 seconds; Table 1 gives the range within each category.

Within each battery chemistry and fault category, vehicles are assigned in sequence to training, validation, and test sets at a ratio of 4:2:4. All segments sharing an identifier stay in one subset. The partition remains fixed across training seeds. See Table 1 for counts and Appendix A for data details.

Table 1: CH-BatteryGen vehicle counts and per-segment length ranges.

Type	Fault Category	Total Vehicles	Number of Vehicles			Segment Length (records)
			Train	Val	Test	
LFP	Normal	400	160	80	160	13k – 37k
	High Resistance	30	12	6	12	16k – 29k
	Low Capacity	30	12	6	12	13k – 32k
	Self-discharge	40	16	8	16	31k – 53k
NCM	Normal	400	160	80	160	4.2k – 12k
	High Resistance	30	12	6	12	4.2k – 11k
	Low Capacity	30	12	6	12	11k – 28k
	Self-discharge	40	16	8	16	4.0k – 10k

We report per-class F1, Macro-F1, and Weighted-F1 at the segment level. Macro-F1 weights classes equally, whereas Weighted-F1 accounts for their frequencies, providing complementary views of the imbalanced data. We summarize performance by the mean and standard deviation over three training seeds.

Baselines. We compare four groups of baselines. (1) **Traditional approaches** include LOF (Breunig et al., 2000), Isolation Forest (Liu et al., 2008), SVM (Cortes & Vapnik, 1995), KNN (Cover & Hart, 1967), and PCA (Abdi & Williams, 2010). (2) **Deep time-series models** include LSTM (Hochreiter & Schmidhuber, 1997), Anomaly Transformer (Xu et al., 2022), and TimesNet (Wu et al., 2023). (3) **Visual and multimodal baselines** include ViT (Dosovitskiy et al., 2021), CLIP (Radford et al., 2021), and ChatTS (Xie et al., 2025). (4) **General-purpose large models** are represented by GPT-4o with fault-informed prompts (OpenAI, 2024).

Implementation. The default backbone is Qwen2.5-VL-7B-Instruct (Bai et al., 2025). We use 448×448 images and train for five epochs with AdamW (Loshchilov & Hutter, 2019), learning rate 5×10^{-5}, and batch size one. The vision encoder and original language-model weights are frozen. Only query/value LoRA parameters ($r = 8$, $\alpha = 16$, dropout 0.05) and the four-output classification head are trained. Training seeds are 42, 123, and 2024. We use PyTorch 2.1.2, Python 3.11, CUDA 12.1, and an RTX 4090D (24 GB). Appendix A.3 details the baseline implementations; Appendix B reports further evaluations.

4.2 Main Results

Table 2 shows that KnoGML achieves the highest Weighted-F1/Macro-F1 on both chemistries, reaching 90.79%/90.24% on LFP and 90.10%/87.61% on NCM. Its Macro-F1 exceeds the strongest traditional baseline, KNN, by 7.44 and 1.97 percentage points, demonstrating improved overall diagnosis across both chemistries.

The per-class results show that the gains extend beyond the frequent normal class. On LFP self-discharge, TimesNet and ViT obtain 32.75% and 28.88%, compared with 90.75% for KnoGML, although some baselines exceed KnoGML on individual NCM faults.

ChatTS and GPT-4o report Macro-F1 between 61.82% and 67.99%, while CLIP remains near 57%. The gap highlights the value of task-specific diagnostic guidance and adaptation beyond general-purpose multimodal processing; the following ablations examine their contributions.

4.3 Ablation Studies

Structured diagnostic priors. We fix the backbone and training setup and compare four prompt variants (Table 3). A generic instruction removes the three-stage diagnostic guidance; a data-description prompt summarizes signal changes without fault associations; an ambiguous-knowledge prompt uses vague diagnostic descriptions; and an incorrect-knowledge prompt supplies wrong fault

Table 2: F1 scores on LFP and NCM (mean ± standard deviation, %). W-F1 and M-F1 denote Weighted-F1 and Macro-F1; **bold** and underline mark the best and second-best results.

LFP (F1 %)						
Method	High Res	Low Cap	Self-Dis	Normal	W-F1	M-F1
LOF	$55.25_{\pm 0.4}$	$50.19_{\pm 0.5}$	$48.80_{\pm 0.6}$	$88.58_{\pm 0.2}$	$81.09_{\pm 0.3}$	$60.71_{\pm 0.4}$
IForest	$19.82_{\pm 1.1}$	$25.38_{\pm 1.4}$	$15.25_{\pm 1.2}$	$69.88_{\pm 0.8}$	$59.84_{\pm 0.9}$	$32.58_{\pm 1.2}$
SVM	$45.55_{\pm 0.2}$	$48.99_{\pm 0.3}$	$37.88_{\pm 0.4}$	$84.22_{\pm 0.1}$	$76.08_{\pm 0.2}$	$54.16_{\pm 0.3}$
KNN	$84.77_{\pm 0.2}$	$\underline{86.82}_{\pm 0.3}$	$82.81_{\pm 0.3}$	$76.78_{\pm 0.5}$	$78.34_{\pm 0.4}$	$\underline{82.80}_{\pm 0.3}$
PCA	$\underline{87.92}_{\pm 0.1}$	$84.18_{\pm 0.2}$	$\underline{86.61}_{\pm 0.1}$	$25.93_{\pm 0.4}$	$38.00_{\pm 0.3}$	$71.16_{\pm 0.2}$
LSTM	$70.22_{\pm 0.8}$	$75.44_{\pm 0.9}$	$69.77_{\pm 1.1}$	$\underline{89.66}_{\pm 0.4}$	$\underline{86.05}_{\pm 0.6}$	$76.27_{\pm 0.7}$
Anomaly Trans.	$39.17_{\pm 1.5}$	$41.09_{\pm 1.2}$	$45.31_{\pm 1.8}$	$88.95_{\pm 0.6}$	$79.60_{\pm 0.9}$	$53.63_{\pm 1.3}$
TimesNet	$39.27_{\pm 1.2}$	$42.49_{\pm 1.4}$	$32.75_{\pm 1.5}$	$87.48_{\pm 0.8}$	$77.51_{\pm 0.9}$	$50.50_{\pm 1.1}$
ViT	$22.21_{\pm 1.8}$	$25.15_{\pm 2.0}$	$28.88_{\pm 2.2}$	$82.66_{\pm 1.1}$	$71.28_{\pm 1.3}$	$39.73_{\pm 1.6}$
CLIP	$48.33_{\pm 0.9}$	$50.83_{\pm 1.1}$	$42.50_{\pm 1.3}$	$86.66_{\pm 0.5}$	$78.68_{\pm 0.7}$	$57.08_{\pm 0.9}$
ChatTS	$61.22_{\pm 1.3}$	$63.11_{\pm 1.5}$	$55.55_{\pm 1.6}$	$83.15_{\pm 0.9}$	$78.42_{\pm 1.0}$	$65.76_{\pm 1.2}$
GPT-4o	$52.22_{\pm 1.6}$	$64.88_{\pm 1.4}$	$48.11_{\pm 1.9}$	$84.44_{\pm 0.8}$	$78.43_{\pm 1.1}$	$62.41_{\pm 1.4}$
KnoGML	$\mathbf{89.58}_{\pm 0.3}$	$\mathbf{89.67}_{\pm 0.4}$	$\mathbf{90.75}_{\pm 0.3}$	$\mathbf{90.97}_{\pm 0.2}$	$\mathbf{90.79}_{\pm 0.2}$	$\mathbf{90.24}_{\pm 0.3}$

NCM (F1 %)						
Method	High Res	Low Cap	Self-Dis	Normal	W-F1	M-F1
LOF	$54.57_{\pm 0.5}$	$49.28_{\pm 0.6}$	$47.95_{\pm 0.4}$	$87.53_{\pm 0.3}$	$80.09_{\pm 0.3}$	$59.83_{\pm 0.4}$
IForest	$20.22_{\pm 1.3}$	$25.14_{\pm 1.5}$	$14.98_{\pm 1.0}$	$70.22_{\pm 0.9}$	$60.10_{\pm 0.8}$	$32.64_{\pm 1.1}$
SVM	$46.33_{\pm 0.3}$	$47.27_{\pm 0.2}$	$38.55_{\pm 0.5}$	$84.99_{\pm 0.1}$	$76.69_{\pm 0.2}$	$54.29_{\pm 0.3}$
KNN	$\underline{85.31}_{\pm 0.3}$	$\underline{86.25}_{\pm 0.2}$	$83.14_{\pm 0.4}$	$87.84_{\pm 0.2}$	$\underline{87.22}_{\pm 0.2}$	$\underline{85.64}_{\pm 0.2}$
PCA	$\mathbf{90.42}_{\pm 0.1}$	$79.58_{\pm 0.2}$	$\mathbf{87.95}_{\pm 0.2}$	$26.15_{\pm 0.5}$	$38.16_{\pm 0.4}$	$71.03_{\pm 0.2}$
LSTM	$69.95_{\pm 1.0}$	$74.89_{\pm 0.8}$	$70.12_{\pm 1.2}$	$90.22_{\pm 0.5}$	$86.48_{\pm 0.6}$	$76.30_{\pm 0.8}$
Anomaly Trans.	$37.88_{\pm 1.4}$	$40.72_{\pm 1.6}$	$20.11_{\pm 2.1}$	$89.27_{\pm 0.5}$	$77.74_{\pm 0.8}$	$47.00_{\pm 1.5}$
TimesNet	$42.08_{\pm 1.6}$	$52.77_{\pm 1.3}$	$43.98_{\pm 1.8}$	$84.22_{\pm 0.9}$	$76.59_{\pm 1.1}$	$55.76_{\pm 1.2}$
ViT	$23.11_{\pm 2.1}$	$24.77_{\pm 1.9}$	$29.22_{\pm 2.4}$	$83.21_{\pm 1.0}$	$71.78_{\pm 1.2}$	$40.08_{\pm 1.7}$
CLIP	$47.55_{\pm 1.2}$	$49.66_{\pm 0.8}$	$41.55_{\pm 1.1}$	$87.22_{\pm 0.6}$	$78.93_{\pm 0.8}$	$56.50_{\pm 0.9}$
ChatTS	$61.99_{\pm 1.4}$	$62.88_{\pm 1.2}$	$55.99_{\pm 1.8}$	$\mathbf{91.11}_{\pm 0.6}$	$84.86_{\pm 0.9}$	$67.99_{\pm 1.1}$
GPT-4o	$53.33_{\pm 1.7}$	$50.47_{\pm 2.1}$	$58.75_{\pm 1.6}$	$84.72_{\pm 0.7}$	$78.70_{\pm 1.0}$	$61.82_{\pm 1.5}$
KnoGML	$84.58_{\pm 0.4}$	$\mathbf{87.08}_{\pm 0.3}$	$\underline{87.81}_{\pm 0.4}$	$\underline{90.97}_{\pm 0.2}$	$\mathbf{90.10}_{\pm 0.2}$	$\mathbf{87.61}_{\pm 0.3}$

Table 3: Prompt ablations on LFP and NCM (F1 in percent, mean ± standard deviation). The backbone and training setup are fixed. Bold indicates the best mean.

LFP (F1 %)						
Method	High Res	Low Cap	Self-Dis	Normal	W-F1	M-F1
w/o Structured Knowledge	$54.02_{\pm 1.8}$	$62.93_{\pm 1.5}$	$60.45_{\pm 2.1}$	$80.12_{\pm 1.2}$	$75.95_{\pm 1.4}$	$64.38_{\pm 1.6}$
Prompt w/ Data Description	$81.01_{\pm 0.8}$	$83.97_{\pm 0.9}$	$\mathbf{91.03}_{\pm 0.7}$	$85.17_{\pm 0.5}$	$85.32_{\pm 0.6}$	$85.30_{\pm 0.7}$
Prompt w/ Ambiguous Knowledge	$72.01_{\pm 1.2}$	$75.02_{\pm 1.3}$	$83.01_{\pm 1.1}$	$88.02_{\pm 0.7}$	$85.88_{\pm 0.8}$	$79.52_{\pm 1.0}$
Incorrect Knowledge	$40.01_{\pm 2.4}$	$44.02_{\pm 2.6}$	$58.01_{\pm 2.1}$	$62.01_{\pm 1.8}$	$59.29_{\pm 2.0}$	$51.01_{\pm 2.2}$
KnoGML	$\mathbf{89.58}_{\pm 0.3}$	$\mathbf{89.67}_{\pm 0.4}$	$90.75_{\pm 0.3}$	$\mathbf{90.97}_{\pm 0.2}$	$\mathbf{90.79}_{\pm 0.2}$	$\mathbf{90.24}_{\pm 0.3}$

NCM (F1 %)						
Method	High Res	Low Cap	Self-Dis	Normal	W-F1	M-F1
w/o Structured Knowledge	$49.98_{\pm 2.2}$	$53.01_{\pm 1.9}$	$57.22_{\pm 2.0}$	$85.01_{\pm 1.1}$	$78.77_{\pm 1.5}$	$61.31_{\pm 1.8}$
Prompt w/ Data Description	$72.02_{\pm 1.1}$	$75.03_{\pm 1.0}$	$81.01_{\pm 0.9}$	$89.97_{\pm 0.4}$	$87.28_{\pm 0.5}$	$79.51_{\pm 0.8}$
Prompt w/ Ambiguous Knowledge	$63.02_{\pm 1.4}$	$66.01_{\pm 1.5}$	$72.01_{\pm 1.2}$	$84.02_{\pm 0.9}$	$80.72_{\pm 1.0}$	$71.27_{\pm 1.2}$
Incorrect Knowledge	$31.01_{\pm 2.8}$	$36.02_{\pm 2.5}$	$41.01_{\pm 2.7}$	$50.01_{\pm 2.2}$	$47.31_{\pm 2.4}$	$39.51_{\pm 2.6}$
KnoGML	$\mathbf{84.58}_{\pm 0.4}$	$\mathbf{87.08}_{\pm 0.3}$	$\mathbf{87.81}_{\pm 0.4}$	$\mathbf{90.97}_{\pm 0.2}$	$\mathbf{90.10}_{\pm 0.2}$	$\mathbf{87.61}_{\pm 0.3}$

associations. These comparisons assess the complete prompt designs, including their diagnostic content and organization.

With generic instructions, Macro-F1 falls to 64.38% on LFP and 61.31% on NCM; NCM high-resistance F1 drops to 49.98%. Signal descriptions recover some performance, but still trail KnoGML by 4.94 and 8.10 Macro-F1 points. The advantage shows that explicit fault associations add diagnostic value beyond descriptions of curve changes. The benefit is aggregate: the data-description variant is slightly better on LFP self-discharge.

Ambiguous knowledge reduces NCM Macro-F1 to 71.27%, and incorrect knowledge produces the largest decline, to 39.51%. Precise fault descriptions outperform vague guidance, while incorrect associations undermine classification. Knowledge quality therefore matters beyond simply adding a text prompt, although these comparisons do not independently verify each physical relation.

Visual input and adaptation. Adding multivariate plot images improves Macro-F1 over the no-image variant by 8.63/7.77 points on LFP/NCM (Table 4). Visual observations thus contribute beyond the contextual fields and shared knowledge in the prompt. Time-domain curves also outperform FFT-based frequency-domain plots by 6.10 points on NCM. This comparison supports the chosen temporal representation over the tested spectral plotting alternative.

Table 4: Visual-input and adaptation ablations.

Setting	LFP		NCM	
	W-F1	M-F1	W-F1	M-F1
w/o vision	$83.56_{\pm 0.7}$	$81.61_{\pm 0.8}$	$81.03_{\pm 0.8}$	$79.84_{\pm 0.9}$
w/ FFT	$83.35_{\pm 0.6}$	$81.29_{\pm 0.7}$	$83.92_{\pm 0.5}$	$81.51_{\pm 0.6}$
w/o LoRA adaptation	$76.85_{\pm 1.5}$	$56.89_{\pm 1.8}$	$71.90_{\pm 1.6}$	$59.83_{\pm 1.9}$
w/o head tuning	$83.08_{\pm 0.9}$	$80.02_{\pm 1.1}$	$84.71_{\pm 0.8}$	$75.26_{\pm 1.2}$
KnoGML	$\mathbf{90.79}_{\pm 0.2}$	$\mathbf{90.24}_{\pm 0.3}$	$\mathbf{90.10}_{\pm 0.2}$	$\mathbf{87.61}_{\pm 0.3}$

Disabling LoRA adaptation reduces NCM Macro-F1 to 59.83%; freezing the classification head reduces it to 75.26%. These declines show that the pretrained representation alone is insufficient in this setting. Adapting both the representation and decision layer complements the diagnostic prompts and visual input.

4.4 Learning with limited training data

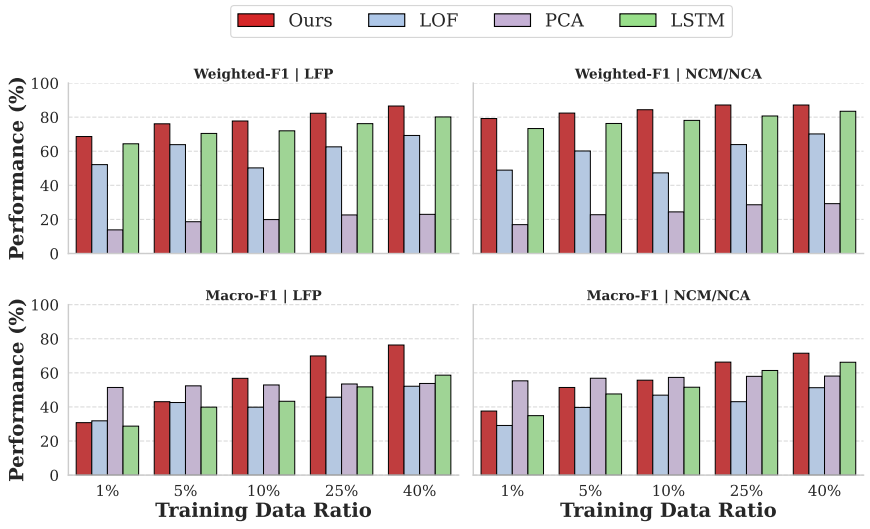


Figure 3: Weighted-F1 and Macro-F1 using 1%–40% of training vehicles on LFP and NCM, averaged over three random selections. KnoGML is labeled "Ours".

To assess diagnosis with scarce labels, we randomly select training vehicles and use their associated segments, averaging three repetitions (Figure 3). KnoGML retains the highest Weighted-F1 at all plotted ratios and the highest Macro-F1 on both chemistries at 25% and 40%. These results support diagnosis with fewer labeled vehicles, with class-balanced gains at 25% and 40%. At smaller

fractions, the advantage varies across metrics. In particular, 1% corresponds to approximately two of 200 training vehicles and cannot cover all four classes, limiting conclusions about learning every fault.

4.5 Computational efficiency

Table 5 shows the computational cost of the diagnostic gains. KnoGML raises Macro-F1 from 56.89% to 90.24% and Weighted-F1 from 76.85% to 90.79% on LFP, with higher memory use and operation counts. Throughput decreases from 8.50 to 3.20 samples per second. LoRA limits the trainable parameter set, although inference still uses the frozen backbone. The resulting accuracy–cost trade-off calls for considering the latency and memory requirements of the intended application. Appendix A.3 summarizes adaptation and evaluation settings.

Table 5: Computational cost and classification performance on LFP.

Metric	Backbone	KnoGML
Memory (GB)	2.30	4.70
Throughput (sample/s)	8.50	3.20
GFLOPs	15.20	35.80
Weighted-F1	76.85	90.79
Macro-F1	56.89	90.24

4.6 Cross-modal interaction

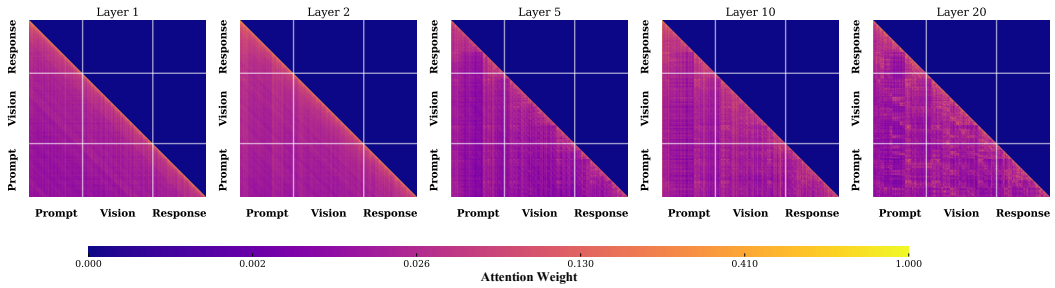


Figure 4: Attention maps for Layers 1, 2, 5, 10, and 20, with groups labeled Prompt, Vision, and Response. Colors indicate attention weights; the maps visualize interactions among the prompt, image, and classifier-input representations across layers.

Figure 4 visualizes attention within and across text, image, and classifier-input representations. Response denotes the representation preceding the classification head. The maps complement the prompt and input ablations by displaying cross-modal interactions at different layers. They provide a qualitative view of joint processing, rather than establishing which physical relationship determines a prediction.

5 Conclusion

We presented KnoGML, which formulates fault diagnosis as knowledge-guided matching between fault descriptions and signal patterns. Multivariate plot images represent variable-length segments, while structured diagnostic priors guide their interpretation. KnoGML achieves the highest Macro-F1 and Weighted-F1 among all evaluated baselines across two battery chemistries. Ablations establish the value of explicit fault associations beyond signal descriptions, supporting this approach as a design reference for knowledge-grounded industrial diagnosis.

Despite these promising results, several limitations remain. Fixed-resolution plots may obscure brief changes, and class-label supervision does not verify individual knowledge matches. Evaluation on semi-synthetic data with separated identifiers cannot establish generalization to independent field trajectories. Future work should improve temporal detail and validate knowledge use on field data.

AI USE STATEMENT

The research ideas, method, experimental design, and experimental work presented in this paper are the authors' original contributions. Generative AI tools assisted with language editing and translation, manuscript organization, consistency checks, and feedback on methodological descriptions and experimental interpretation.

REFERENCES

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010. doi: 10.1002/wics.101.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. doi: 10.1007/BF00994018.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi: 10.1109/TIT.1967.1053964.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Musashi Hinck, Matthew L. Olson, David Cobbley, Shao-Yen Tseng, and Vasudev Lal. LLaVA-Gemma: Accelerating multimodal foundation models with a compact language model, 2024. URL https://arxiv.org/abs/2404.01331.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Jichao Hong, Fengwei Liang, Jian Yang, and Shanxiao Du. An exhaustive review of battery faults and diagnostic techniques for real-world electric vehicle safety. *Journal of Energy Storage*, 99: 113234, 2024. doi: 10.1016/j.est.2024.113234.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Rolf Isermann. Model-based fault-detection and diagnosis—status and applications. *Annual Reviews in Control*, 29(1):71–85, 2005. doi: 10.1016/j.arcontrol.2004.12.002.

Feng Jia, Yaguo Lei, Jing Lin, Xin Zhou, and Na Lu. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mechanical Systems and Signal Processing*, 72-73:303–315, 2016. doi: 10.1016/j.ymssp.2015.10.025. URL https://doi.org/10.1016/j.ymssp.2015.10.025.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01728.

Eunji Kim, Sungzoon Cho, Byeongeon Lee, and Myoungsu Cho. Fault detection and diagnosis using self-attentive convolutional neural networks for variable-length sensor data in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 32(3):302–309, 2019. doi: 10.1109/tsm.2019.2917521. URL https://doi.org/10.1109/tsm.2019.2917521.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-VL technical report, 2025. URL https://arxiv.org/abs/2504.07491.

Zihao Lei, Ping Zhang, Yuejian Chen, Ke Feng, Guangrui Wen, Zheng Liu, Ruqiang Yan, Xuefeng Chen, and Chunsheng Yang. Prior knowledge-embedded meta-transfer learning for few-shot fault diagnosis under variable operating conditions. *Mechanical Systems and Signal Processing*, 200: 110491, 2023. doi: 10.1016/j.ymssp.2023.110491. URL https://doi.org/10.1016/j.ymssp.2023.110491.

Lin Lin, Sihao Zhang, Song Fu, and Yikun Liu. FD-LLM: Large language model for fault diagnosis of complex equipment. *Advanced Engineering Informatics*, 65:103208, 2025. ISSN 1474-0346. doi: 10.1016/j.aei.2025.103208. URL https://www.sciencedirect.com/science/article/pii/S1474034625001016.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008. doi: 10.1109/ICDM.2008.17.

Qingdi Liu, Yan Fu, Lishuo Liu, Yanke Lin, Jin Xin, Jianfeng Zhang, Cheng Liu, Lujia Pan, Dongxu Guo, Yuejiu Zheng, and Qiang Li. Battery fault: A comprehensive dataset and benchmark for battery fault diagnosis. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0d781fa5f639bf2caf728a68e9678362-Abstract-Conference.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Meta. Llama 3.2-Vision model card. Model documentation, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct.

OpenAI. Hello GPT-4o. OpenAI Blog, 2024. URL https://openai.com/index/hello-gpt-4o/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3939–3945, 2015. URL https://www.ijcai.org/Proceedings/15/Papers/553.pdf.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02186.

Chengyong Xiao, Xiaowei Liu, Aziguli Wulamu, and Dezheng Zhang. KG-SR-LLM: Knowledge-guided semantic representation and large language model framework for cross-domain bearing fault diagnosis. *Sensors*, 25(18):5758, 2025. ISSN 1424-8220. doi: 10.3390/s25185758. URL https://www.mdpi.com/1424-8220/25/18/5758.

Xiaomi LLM-Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, et al. MiMo-VL technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. ChatTS: Aligning time series with LLMs via synthetic data for enhanced understanding and reasoning. *Proceedings of the VLDB Endowment*, 18(8):2385–2398, 2025. doi: 10.14778/3742728.3742735. URL https://arxiv.org/abs/2412.03104.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly Transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.02642.

Jingzhao Zhang, Yanan Wang, Benben Jiang, Haowei He, Shaobo Huang, Chen Wang, Yang Zhang, Xuebing Han, Dongxu Guo, Guannan He, and Minggao Ouyang. Realistic fault detection of Li-ion battery via dynamical deep learning. *Nature Communications*, 14(1):5940, 2023. doi: 10.1038/s41467-023-41226-5.

Wei Zhang, Gaoliang Peng, Chuanhao Li, Yuanhang Chen, and Zhujun Zhang. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors*, 17(2):425, 2017. doi: 10.3390/s17020425. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5336047/.

Jingyuan Zhao, Xuning Feng, Junbin Wang, Yubo Lian, Minggao Ouyang, and Andrew F. Burke. Battery fault diagnosis and failure prognosis for electric vehicles using spatio-temporal transformer networks. *Applied Energy*, 352:121949, 2023. doi: 10.1016/j.apenergy.2023.121949.

Zihao Zhou and Rose Yu. Can LLMs understand time series anomalies? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/05774fb74e863308c4b460c9f49f6918-Abstract-Conference.html.

A Dataset and implementation details

The dataset, jointly constructed by the China Automotive Engineering Research Institute (CAERI) and Huawei, covers both LFP and NCM chemical systems across charging and discharging scenarios, with four fault labels (normal, self-discharge, high-resistance, and low-capacity). The dataset includes variable-length charging and discharging segments. Its data card reports 10,000 LFP segments and 9,478 NCM segments, including fewer charging than discharging segments for NCM. We utilize eight critical channels—including charge status, total voltage, total current, SOC, and extreme cell voltages and temperatures—to construct the multimodal visual evidence. The data spans a broad environmental temperature range from -10°C to 45°C, with a sampling frequency of 0.1 Hz. We split by published vehicle identifier, keeping all associated segments in one subset. This prevents identifier overlap; it does not establish independence of undisclosed generating trajectories. Within each chemistry and fault category, vehicles are assigned in sequence using a 4:2:4 training/validation/test ratio. All segments from a vehicle remain together. Evaluations are conducted at the segment level, classifying individual charging or discharging fragments. Furthermore, to mitigate the impact of class imbalance across fault categories, a class-weighted cross-entropy loss is employed during training. Detailed statistics of the dataset are reported in Table 1.

A.1 Input variables

The grid contains the eight channels below. Time supplies the common horizontal reference. Separate vertical axes retain the units of each variable; visual heights in different subplots are not directly comparable as physical magnitudes.

Table 6: Battery channels used to construct plot images.

Channel	Meaning	Unit
`CHARGE_STATUS`	Operating-status code	—
`SUM_VOLTAGE`	Pack terminal voltage	V
`SUM_CURRENT`	Pack current	A
`SOC`	Reported state of charge	%
`MAX_CELL_VOLT`	Maximum cell voltage	V
`MIN_CELL_VOLT`	Minimum cell voltage	V
`MAX_TEMP`	Maximum probe temperature	°C
`MIN_TEMP`	Minimum probe temperature	°C

The sampling interval is 10 seconds. The source variable specifications give precision of 1 second for relative timestamps, 1 for charge-status codes, 0.1 V for pack voltage, 0.1 A for current, 1% for SOC, 0.001 V for cell voltages, and 1°C for probe temperatures.

A.2 Prompt template

Table 7 specifies the diagnostic instructions, including their qualitative comparisons and contextual fields. The phase instructions combine operating status, current direction under the input convention, and SOC evolution. Terms such as "abnormal", "earlier", and "reasonable" are qualitative instructions rather than numerical thresholds. The template does not establish a unique physical cause for each curve pattern. It specifies the intended diagnostic procedure, while the classification head returns the final category without a generated explanation.

Table 7: Knowledge-grounded diagnostic instructions. Context fields vary by input; fault descriptions and the output schema are shared.

Module	**Content & Logic Definitions**
I. Context Information & Visual Inputs	
Context Injection	Battery Type: `<battery_type>`, Vehicle ID: `<car_id>`, Current Status: `<charge_desc>`

Table 7 continued

Module	Content & Logic Definitions
Visual Variables	The input figure displays trends for the following 8 variables over a time series: 1. `CHARGE_STATUS`: Operating-status code; interpret it jointly with current and SOC rather than treating every non-charging code as discharge. 2. `SUM_VOLTAGE`: Total voltage of the battery pack. 3. `SUM_CURRENT`: Current value during the charging or discharging process. 4. `SOC`: State of Charge (estimated remaining capacity). 5. `MAX_CELL_VOLT`: The highest individual cell voltage value in the battery pack. 6. `MIN_CELL_VOLT`: The lowest individual cell voltage value in the battery pack. 7. `MAX_TEMP`: The highest temperature point in the battery pack. 8. `MIN_TEMP`: The lowest temperature point in the battery pack.
II. Task Definition	
Based on the image, judge if the battery is faulty and classify it into one of the following categories: *high_resistance, low_capacity, normal, self_discharge*.	
III. Structured Diagnostic Protocol	
Step 1: Phase Identification	**Determine the operating phase from `CHARGE_STATUS`, `SUM_CURRENT`, and `SOC`:** • **Discharge Stage:** Identify discharge from jointly consistent status, current direction under the dataset convention, and decreasing SOC. A status code alone is insufficient. • **Charge Stage:** Identify charge from jointly consistent status, current direction under the dataset convention, and increasing SOC. If these disagree, do not force a phase assignment.
Step 2: Symptom Verification	**Inspect signal patterns (slopes, amplitudes, cross-variable relationships, and operating phases):** **A. High Resistance Suspect (high_resistance)** *Discharge Phase:* • `SUM_VOLTAGE` shows a significant and abnormal sudden drop at the beginning of discharge or when load increases. • `SOC` drops abnormally fast (compared to the actual discharge current). • The voltage spread (`MAX_CELL_VOLT - MIN_CELL_VOLT`) increases significantly. • `MAX_TEMP` rises unusually quickly or reaches an excessively high temperature during discharge. *Charge Phase:* • Charging speed is abnormally slow (`SOC` rising slope is low, inconsistent with charging current). • `SUM_VOLTAGE` climbs rapidly to the cutoff voltage at the end of charging. • The voltage spread (`MAX_CELL_VOLT - MIN_CELL_VOLT`) increases significantly during charging (especially in the later stages). • `MAX_TEMP` rises unusually quickly or becomes too high (obvious heating). **B. Low Capacity Suspect (low_capacity)** *Charge Phase:* • `SOC` rises abnormally fast when charging from a low level (reaches high SOC in a short time). • `MAX_CELL_VOLT` reaches the charging upper limit earlier (compared to normal batteries or historical data). *Discharge Phase:* • `SOC` drops abnormally fast (drops from high SOC to low SOC in a short time). • `MIN_CELL_VOLT` reaches the discharge lower limit earlier (compared to normal batteries or historical data). *Key Point:* Voltage spread (`MAX_CELL_VOLT - MIN_CELL_VOLT`) and temperature (`MAX_TEMP`/`MIN_TEMP`) changes are usually relatively normal. Anomalies are mainly reflected in the SOC change speed and the timing of cell voltages reaching limits.

Table 7 continued

Module	Content & Logic Definitions
	C. Self-Discharge Suspect (self_discharge) *Charge Phase:* • MIN_CELL_VOLT starts at a noticeably lower level, and its rising trajectory remains relatively stable throughout the entire process. • Near the end of charge, even when approaching full capacity, the voltage remains significantly lagging and fails to reach the full-charge plateau. *Discharge Phase:* • MIN_CELL_VOLT is at a lower level from the very beginning, causing an extremely large initial voltage spread (MAX_CELL_VOLT - MIN_CELL_VOLT). *Key Point:* The voltage spread (MAX_CELL_VOLT - MIN_CELL_VOLT) remains large and stable throughout the entire charge/discharge cycle. Temperature changes appear normal. The primary anomaly manifests as a "plateau-level" downward shift of the full-charge voltage. **D. Normal (normal)** In all phases (Charging, Discharging, Idle), parameters change as expected: • Voltage (SUM_VOLTAGE, MAX_CELL_VOLT, MIN_CELL_VOLT) changes smoothly without abnormal surges/drops. Voltage spread stays stable or fluctuates reasonably. • Temperature (MAX_TEMP, MIN_TEMP) changes are within a reasonable range without abnormal heating. • SOC change rate matches the magnitude and direction of SUM_CURRENT. • Cell voltages reach upper/lower limits at expected timings (not prematurely).
Step 3: Final Decision	**Comprehensive Judgment & Output:** Based on the operating phase identified in Step 1 and the signal patterns assessed in Step 2, select the fault label using the following class IDs: **0**: high_resistance, **1**: low_capacity, **2**: normal, **3**: self_discharge

A.3 Implementation details

Model adaptation. The default backbone is Qwen2.5-VL-7B-Instruct. We freeze the vision encoder and the original language-model parameters, attaching LoRA to the query and value projections with rank 8, $\alpha = 16$, and dropout 0.05. The trainable classification head has four output units. In the attention visualization, "response" refers to the hidden representation preceding this head. Inference returns a class label rather than a generated diagnostic response.

Baseline implementations. SVM and KNN perform multiclass classification directly. For PCA, we fit one model to the training samples of each class and assign an input to the class with the smallest reconstruction error. Most representation-learning baselines use an added classification head to predict the four fault labels. These implementations map the learned representations or reconstruction errors to the same four-label task.

Evaluation settings. The limited-data experiment randomly selects the specified proportion of vehicles from the training set, using the segments belonging to those vehicles. Results are averaged over three random selections per ratio. The label "NCM/NCA" in the plot denotes the ternary-battery family; the evaluated dataset subset is NCM. The FFT ablation replaces time-domain curve plots with frequency-domain plots, comparing two visual representations within the framework.

The cost comparison pairs computational measurements with classification scores. Parameter-efficient adaptation reduces the trainable parameter set, whereas inference still involves the frozen backbone. These are distinct aspects of efficiency; the cost table describes the accuracy–cost trade-off rather than establishing suitability for a particular industrial latency target.

B ADDITIONAL EVALUATIONS

B.1 RESOLUTION SENSITIVITY

Table 8: Effect of image resolution on LFP (F1 in percent). We report Macro-F1 and Weighted-F1 scores across different grid image sizes.

Resolution	Macro-F1	Weighted-F1
224×224	87.20	87.50
448×448	90.24	90.79
560×560	89.49	89.67
616×616	89.79	89.90
896×896	88.76	88.62

Macro-F1 ranges from 87.20% to 90.24% across the tested image resolutions, with the highest value at 448×448. Larger images do not improve the result monotonically. This supports the default resolution among the tested settings, but does not establish that all time-series detail remains visible.

B.2 BACKBONE COMPARISON

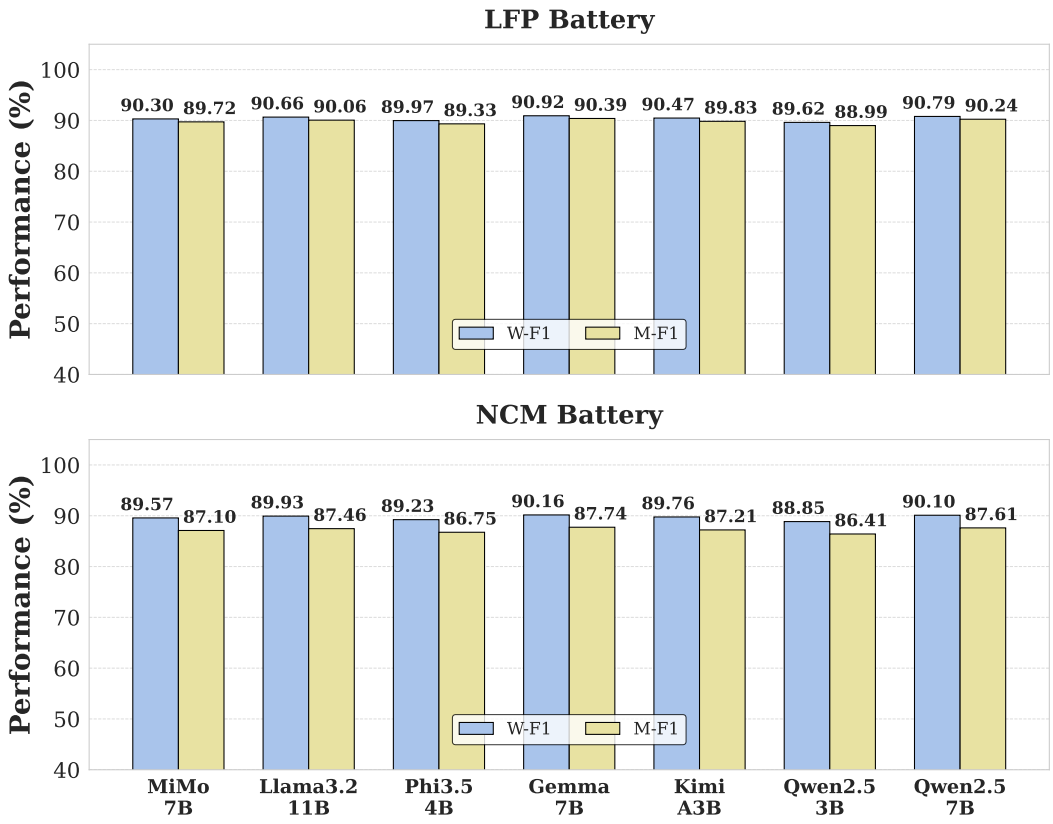


Figure 5: Comparison of multimodal backbones on LFP and NCM. Weighted-F1 (W-F1) and Macro-F1 (M-F1) are reported for adaptation within each dataset.

The evaluated backbones include Qwen2.5-VL (Bai et al., 2025), Kimi-VL (Kimi Team et al., 2025), Llama-3.2-Vision (Meta, 2024), MiMo-VL (Xiaomi LLM-Core Team et al., 2025), Phi-3.5-Vision (Abdin et al., 2024), and LLaVA-Gemma (Hinck et al., 2024). Macro-F1 ranges from 88.99% to 90.39% on LFP and from 86.41% to 87.74% on NCM. The consistently high scores across these backbones show that KnoGML is not tied to a single pretrained model. They support its generality

across the evaluated multimodal backbones, with the same knowledge-guided diagnostic formulation applied to both battery datasets.

B.3 NOISE ROBUSTNESS

Table 9: Performance comparison under original and 10% additive noise conditions on NCM battery. Changes are computed as noisy minus original scores, in percentage points; negative values indicate degradation.

Method	Original		10% Noise		Change (Δ)	
	W-F1	M-F1	W-F1	M-F1	W-F1	M-F1
PCA	38.16	71.03	27.53	52.97	-10.63	-18.06
LOF	80.09	59.83	77.05	53.89	-3.04	-5.94
LSTM	86.48	76.30	78.88	69.88	-7.60	-6.42
KnoGML	**90.10**	**87.61**	**87.07**	**87.61**	**-3.03**	**0.00**

Gaussian noise is added to numerical signal columns in the raw CSV files using seed 42. For column j, $x'_{t,j} = x_{t,j} + \epsilon_{t,j}$, where $\epsilon_{t,j} \sim \mathcal{N}(0, \sigma_j^2)$ and $\sigma_j = 0.1 s_j$. Here, s_j is the sample standard deviation within that CSV (`ddof`=1). If s_j is zero or undefined, $\sigma_j = 0.1|\bar{x}_j|$; if the mean is also zero or undefined, $\sigma_j = 0.01$. Time, charge status, identifiers, and label columns are excluded using case-insensitive column-name matching. Thus, 10% denotes the noise scale relative to within-segment signal variability, rather than the fraction of corrupted observations. Table 9 compares the original and perturbed inputs. Macro-F1 decreases by 18.06 points for PCA, 5.94 for LOF, and 6.42 for LSTM, while KnoGML retains 87.61% at the reported precision. Its Weighted-F1 decreases by 3.03 points. This experiment concerns additive signal perturbations; it does not assess sensor dropout, time shifts, or changes in operating conditions.