RELATE: HEALTH-RELATIVE EVIDENCE FOR STRUCTURED INDUSTRIAL DIAGNOSIS WITH LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Industrial diagnosis spans signals with different sampling rates, fault vocabularies, and diagnostic attributes. We introduce RELATE, a framework that pairs source-rate query features with healthy training references for condition classification and structured record prediction. RELATE-MLP classifies the complete query–reference representation; RELATE-Gen augments the same representation with temporal features and maps the fused evidence to 4 continuous tokens for candidate-conditioned Qwen generation. We evaluate both routes on MOSAIC (Multi-source Operational Signals with Annotated Industrial Conditions), which covers known-setting, closed-set diagnosis across 12 acoustic, vibration, current, and voltage settings under fixed group-level splits. RELATE-MLP reaches 80.60% setting-macro accuracy and 78.36% setting-macro F1 on 3,598 frozen-test groups. RELATE-Gen reaches 78.65/77.01% accuracy/F1 and 68.86% setting-macro Joint EM on records assembled by per-field voting across views. Across three fixed validation runs under the selected autoregressive objective, health-relative evidence improves accuracy by 5.07 points on average; the complementary temporal evidence and learned fusion add 4.56 points.

1 INTRODUCTION

Industrial monitoring produces acoustic, vibration, current, and voltage signals with different sampling rates and fault vocabularies. A bearing record may identify a fault location and damage size, whereas a motor record may describe eccentricity. A diagnosis system must connect signal features to both a local set of conditions and the attributes recorded for that machine. Healthy training examples provide a reference for this process: they describe the signal behavior against which a query can be compared.

We study *known-setting closed-set industrial diagnosis with candidate lists*. Each query arrives with machine context and candidate conditions. The system predicts a condition and the fields supported by the source annotations. Health state, location, and mechanism usually express the condition taxonomy; damage size and eccentricity can vary within a condition and require additional prediction. These two roles motivate separate evaluation of classification and attribute prediction.

Our approach combines a common signal representation with two output routes. A classifier uses the full evidence to select a primary condition. A candidate-conditioned language model receives a compact evidence prefix and generates records for all settings, learning their formats from training targets. We evaluate its field predictions against class-based templates and learned non-language predictors.

We introduce RELATE, a health-relative evidence framework for industrial diagnosis. Its starting point is an ordered subband pair: the query and a retrieved healthy prototype. We augment this pair with its signed and absolute differences to make the comparison explicit in the input representation. RELATE-MLP uses this representation directly for primary diagnosis. RELATE-Gen adds a train-fitted temporal branch and compresses the fused evidence into 4 continuous tokens. These tokens condition a shared decoder together with the candidate set; the training targets define each setting's supported record fields. Our implementation uses MiniROCKET, a frozen FISHER encoder, and Qwen2.5-7B-Instruct. Autoregressive structured targets train the continuous interface and Qwen LoRA parameters. For a fixed schema when latency is the main constraint, the same learned tokens can instead feed a frozen-token readout.

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.

We assemble MOSAIC from eleven RMIS settings and one machine-conditional MIMII-DG setting, retaining source-rate signals and source-supported labels. On its 3,598 test groups, the full-evidence classifier reaches 80.60% setting-macro accuracy, while the generator reaches 78.65% and predicts group-aggregated records at 68.86% Joint EM. Controlled validation experiments test healthy references, the combined temporal branch and fusion, and prefix length. A readout on frozen tokens separates deployment decoding from interface learning; a directly supervised interface provides the corresponding non-language training control.

We make three contributions:

- We introduce RELATE, a health-relative diagnosis framework that builds on FISHER's source-rate signal frontend, preserves ordered query–reference evidence, and supports a full-evidence MLP or a temporally fused 4-token interface for a shared structured decoder.
- We introduce the MOSAIC benchmark protocol for heterogeneous industrial diagnosis with fixed group-level splits, setting-local candidates, source-supported fields, and a shared generated-output scorer.
- Across 12 settings, the full-evidence classifier leads aggregate classification scores, and the generator predicts condition and attribute records through a compact interface. Validation controls show that healthy references and temporal fusion improve the generator, with explicit difference features benefiting the tested Qwen route and the reference pair performing best in the MLP input control.

2 Related Work

Industrial signal pretraining and RMIS. Industrial signal representation has often been tied to one sensor family or one diagnostic task. WDCNN learns raw vibration features for bearing-fault classification, while FISHER uses teacher–student self-distillation on external audio and music to learn representations for industrial signals (Zhang et al., 2017; Fan et al., 2026). The FISHER work also introduces Representation of M5 Industrial Signals (RMIS), a benchmark spanning industrial sound, vibration, current, and voltage datasets. Generic time-series encoders such as TS2Vec and MOMENT broaden the representation-learning setting (Yue et al., 2022; Goswami et al., 2024); MiniROCKET supplies an efficient complementary temporal transform (Dempster et al., 2021). BEATs learns audio representations through acoustic-token prediction (Chen et al., 2022); EAT uses bootstrap self-supervision with utterance- and frame-level objectives (Chen et al., 2024). RELATE-Gen preserves FISHER's ordered subbands and combines them with temporal features for structured diagnosis.

LLM diagnosis and numerical signal conditioning. LLM-based diagnosis has explored both numeric serialization and multimodal conditioning. FD-LLM presents vibration either as tokenized FFT values or named time/frequency statistics and adapts open LLMs for fault classification (Qaid et al., 2024). FD-MVLLM uses multimodal vibration evidence for bearing diagnosis (Li et al., 2025a). DiagLLM combines envelope-spectrum images with expert knowledge (Wang et al., 2025); AeroGPT aligns vibration signals with a large audio model for generative fault classification (Liu et al., 2026). Qwen2-Audio supports general audio instruction following (Chu et al., 2024). RELATE-Gen conditions one generator on continuous signal features and candidate labels to predict the condition and diagnostic attributes in each MOSAIC setting.

Continuous interfaces to language models. Continuous interfaces connect non-text encoders to language models. BLIP-2 uses a lightweight Querying Transformer between a frozen image encoder and a frozen LLM (Li et al., 2023); Time-LLM reprograms time-series patches with text prototypes before a frozen LLM and projects its outputs for forecasting (Jin et al., 2024). TEST similarly aligns time-series features with text prototypes and soft prompts (Sun et al., 2024). RELATE-Gen maps query–reference and temporal evidence into 4 continuous tokens for structured diagnosis. Its language-model base weights remain frozen, while LoRA and the signal interface learn from record targets. Our controls compare generation with a readout on those tokens and compare AR-trained tokens with a directly supervised interface.

Health-relative evidence and references. Healthy references also appear in anomaly scoring. Prototype methods for anomalous-sound detection use normal data to define a reference distribution or prototype and score deviations (Zeng et al., 2023). BearLLM is the closest LLM-based precedent: it pairs a bearing query with a fault-free reference, uses a unified vibration representation with a residual derived from the pair, and feeds the resulting signal features to an LLM for diagnosis and health-management responses (Peng et al., 2025). RELATE adopts query–reference pairing across MOSAIC modalities. A train-only, group-distinct bank yields a prototype r; the interface keeps q and r in corresponding subband positions and, in the full implementation, appends $q - r$ and $|q - r|$ before temporal fusion. The candidate list and source-supported fields define the reporting space. BearLLM evaluates bearing diagnosis and health-management responses on CWRU. Our experiments measure condition and supported-field prediction across the twelve MOSAIC settings. Appendix C also compares a separately adapted RELATE generator with BearLLM on shared CWRU segments.

3 Problem Formulation and MOSAIC

For diagnostic setting s, a benchmark group g contains views $\{x_{g,v}\}$. The candidate set $\mathcal{Y}_s$ and possible field paths $\mathcal{A}_s$ are determined by the source semantics. The fields applicable to a particular record form $\mathcal{A}_g \subseteq \mathcal{A}_s$; this group-specific scoring set is derived from source annotations and is not supplied to the model. A model predicts a primary condition $\hat{y}_g \in \mathcal{Y}_s$ and field values. Provenance identifiers and gold labels remain outside the model-visible input and are joined only during scoring.

The primary condition is evaluated as the classification target. Most record fields realize source taxonomies linked to that condition, such as health state, fault location, and fault mechanism. Damage size and eccentricity are distinct within-condition attributes: their values can change while the primary condition remains fixed. MIMII-DG additionally uses a top-level record `health_state` and a machine-semantic `semantic_factors.health_state`; these are separate scored paths.

MOSAIC combines 11 RMIS settings and the Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization (MIMII-DG) dataset into 12 diagnostic settings covering acoustic leakage, rotating machinery, bearings, planetary gearboxes, motor current and voltage, wind-turbine gearbox states, and machine-conditional acoustic conditions (Fan et al., 2026; Dohi et al., 2022). The benchmark contains 15,477/1,218/3,598 train/validation/test groups and 33,476/2,079/4,430 signal views under fixed group-level splits. These source-defined benchmark groups are the split and scoring units. A group collects the views attached to one source record identifier; depending on the source, its views are signal segments, files, or synchronized channels. Appendix A gives the split construction and source identifiers.

For the 11 RMIS settings, MOSAIC reuses the released signals and original semantic metadata. The protocol adds a frozen group-level split, setting-local candidate spaces, source-supported output schemas, gold-separated generation inputs, and one group-level scorer, together with the MIMII-DG setting. Appendix A.1 details dataset admission and exclusions.

The primary metric is group-level setting-macro accuracy, giving each of the twelve diagnostic settings equal weight. We additionally report class-macro F1 averaged over settings, source-family macro accuracy, and group-micro accuracy, which weights each benchmark group equally. Models that emit candidate logits average them across views before the group-level argmax; models that emit labels or structured records use majority vote, with frozen candidate order breaking ties. From these group predictions onward, all methods use the same scorer, denominators, and metric weights. Each supported record field is voted independently across view-level values; missing values participate in the vote, and a field is emitted only when its winning value is non-missing. The resulting group record can therefore combine fields from different views. A missing field wins only when it has more votes than every non-missing value; tied non-missing values follow canonical JSON lexical order. Group Joint EM requires a correct primary condition and exact matches for every field in $\mathcal{A}_g$. Predicted paths outside $\mathcal{A}_g$ do not change this supported-record metric; paths outside $\mathcal{A}_s$ are also counted in a separate audit. Setting-macro Joint EM first averages group Joint EM within each setting and then weights the 12 settings equally. Field EM computes exact match separately for each of seven field paths over its supported groups, then averages the seven scores equally. Severity EM equally weights damage-size and eccentricity exact match over their respective supported groups. Appendix A provides the complete semantic contract and per-setting counts.

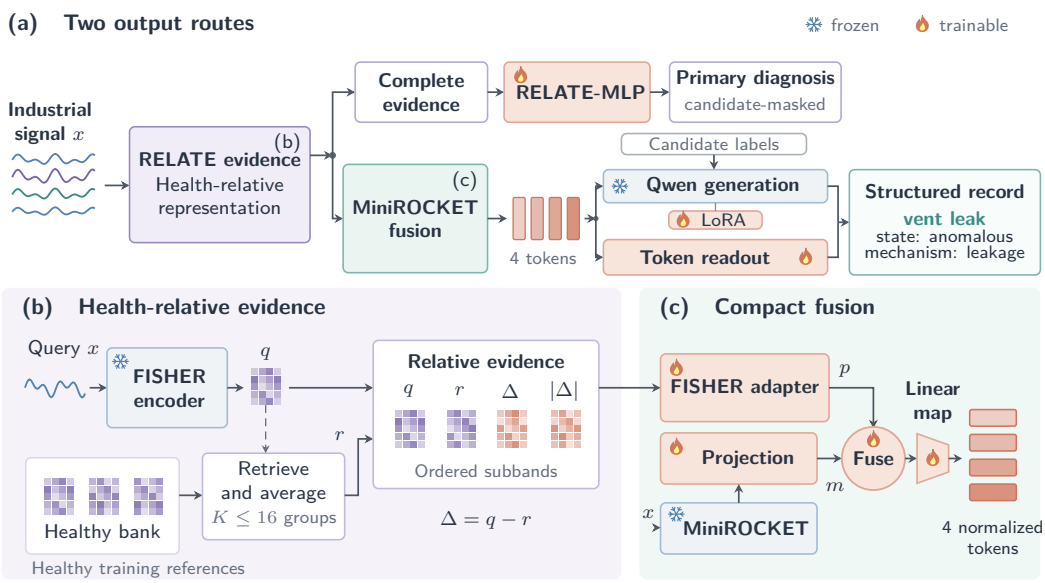

Figure 1: RELATE overview and evidence interface. (a) The full evidence feeds RELATE-MLP, while a MiniROCKET-fused 4-token interface feeds Qwen generation or a token readout. (b) A frozen FISHER encoder supplies query features and a healthy training bank. Query-guided retrieval averages up to 16 distinct healthy groups into the prototype r; the query, reference, and derived differences form $[q, r, \Delta, |\Delta|]$, where $\Delta = q - r$. (c) The relative and MiniROCKET branches are aligned, combined by gated fusion, projected to the Qwen hidden size, and normalized to form 4 evidence tokens. The record card shows selected fields from one generated validation prediction; feature and module glyphs are schematic. Snowflakes denote frozen modules and flames denote trainable modules.

4 Health-Relative Evidence with RELATE

4.1 Overview

Given query signal x, RELATE constructs a health-relative signal representation. RELATE-MLP applies a primary classifier to the complete representation; RELATE-Gen augments it with a train-fitted temporal transform, compresses the fused evidence into 4 continuous tokens, and autoregressively generates a structured diagnosis. Figure 1 summarizes both output routes.

4.2 Healthy-reference construction

Let E be the frozen FISHER encoder and let $q = E(x)$ be the query representation. The healthy bank is built only from views marked healthy or normal in the training annotations. Eligible references match the diagnostic setting, machine axis where applicable, and sample-rate pool. The retrieval rule first removes every view belonging to the query benchmark group, ranks the remaining views by similarity, retains at most one view from each benchmark group, and keeps up to 16 groups. Their uniform mean forms the healthy prototype r. For retrieval, each unpadded FISHER tensor is averaged over its band and time positions to a 384-dimensional vector and L2-normalized; ranking uses cosine similarity between these vectors. The selected full tensors are averaged elementwise before band padding. Retrieval uses representation similarity; operating speed and load are not explicit matching constraints. The training bank remains fixed during evaluation.

4.3 Ordered relative evidence

The FISHER representation retains ordered subband coordinates. RELATE constructs

$$z_E = [q,\ r,\ q - r, |q - r|], \tag{1}$$

where the last two blocks are deterministic functions of q and r that expose signed and magnitude comparisons directly to the learned interface. RELATE then applies the encoder padding mask and averages valid temporal positions within each ordered subband. The flattened representation preserves a fixed coordinate range for each band, allowing the learned projection to distinguish subbands. Temporal averaging pools over the explicit time coordinates of these learned features. MiniROCKET complements them with label-free convolutional statistics computed directly from the waveform, providing temporal-pattern evidence through a fixed convolutional transform. FISHER's rate-scaled 25 ms STFT and 50-bin band split accept source rates of at least 4 kHz without conversion to a common waveform rate. The sources evaluated here span 16–64 kHz and produce 4, 6, 12, or 16 valid subbands, which are aligned by padding and masks before the fixed-size interface.

4.4 Full-evidence classification

RELATE-MLP averages each 384-dimensional component over valid time positions, preserves the ordered bands, and pads each of the four blocks to 16 bands, forming a 24,576-dimensional vector. A 6.30M-parameter MLP with one 256-wide ReLU layer, dropout 0.2, and a shared output layer with 42 label indices predicts candidate-masked logits. The output indices share parameters when the same label string occurs in multiple settings, while each setting's candidate mask and label-to-code map remain local. Each training vector is the mean of its group's view vectors. The seed-111 model trains for 80 epochs with AdamW (learning rate 10^{-3}, weight decay 10^{-4}, batch 256) and inverse-$(\mathrm{setting}, \mathrm{primary})$ sampling; inference averages view logits within each evaluation group. The reported test results use this validation-selected configuration.

4.5 Temporal transform and evidence fusion

MiniROCKET transforms (Dempster et al., 2021) are fitted on training windows without using classifier labels or decision scores. We fit 16 transforms for the setting-local candidate spaces (eleven RMIS settings and five MIMII machine spaces), while the downstream neural interface is shared. The resulting 9,996-dimensional vector is normalized and projected to T bottleneck vectors, producing $m \in \mathbb{R}^{T\times 256}$. Let $p \in \mathbb{R}^{T\times 256}$ denote the aligned FISHER evidence. Fusion is

$$c = [p; m], \qquad g = \sigma(W_g c), \qquad h = \mathrm{LN}(g \odot p + (1 - g) \odot m + W_v(c)), \tag{2}$$

where $[p; m] \in \mathbb{R}^{T\times 512}$ concatenates channels, $W_g : \mathbb{R}^{512} \to \mathbb{R}^{256}$, and W_v is a $512 \to 512 \to 256$ GELU MLP. Fusion layer normalization is followed by a second width-256 output normalization and a linear projection to Qwen's 3,584-dimensional hidden width; trainable prefix normalization then produces the $T = 4$ continuous evidence tokens used by the selected interface.

4.6 Autoregressive structured diagnosis

The 4 signal tokens precede text that lists the modality, machine or component, and setting-local candidates. Qwen2.5-7B-Instruct's 4-bit base weights are frozen; LoRA parameters train with the signal adapter. Training uses a token-level autoregressive loss to predict the setting-specific structured target sequence. At inference, the model generates JSON containing the primary condition and supported fields. Candidate lists condition the prompt; malformed or out-of-candidate records fail Joint EM. Extra unsupported paths are audited separately from the supported-record score. The scorer parses the generated record, aggregates views, and joins gold labels only after prediction generation.

A deterministic renderer presents the structured record as a field-traceable diagnostic note without changing its predicted field values.

5 Experimental Design

All trainable methods use the benchmark train split. Validation selects RELATE configurations; baseline endpoint rules, including the fixed final Qwen2-Audio adapter, are given in Appendix E. RELATE-Gen and its matched generator variants use Qwen2.5-7B-Instruct (Yang et al., 2024) with 4-bit LoRA (Hu et al., 2022) ($r = 8$, $\alpha = 16$), hierarchical sampling over setting, a source-specific balance stratum such as machine identity or operating condition, class, benchmark group, and view,

and 100,424 training microsteps. At each generator endpoint, the model produces JSON diagnoses for all 2,079 validation views; classifier endpoints use their native predictions. Appendix E gives the remaining optimization and decoding details.

RQ1: effectiveness across known settings. We compare signal classifiers, pretrained representations, and language-model diagnosis methods using the same benchmark-group scorer. The Stats15 baseline standardizes 10 time-domain and 5 frequency-domain statistics and fits a setting-local RBF-SVM. The frozen-test comparison covers all 12 settings and the same 3,598 benchmark groups. Model selection uses validation results, fixed training endpoints, or training-only rules, as specified for each method in Appendix E. Evaluation retains each method's training-time sampling rate and window length; Appendix E details these model-specific input configurations. RELATE-MLP is a candidate-masked primary classifier on the complete 24,576-dimensional relative-FISHER representation; RELATE-Gen is the validation-selected 4-token Qwen route shown in Figure 1.

RQ2: record prediction, interface training, and decoding. We compare five record-prediction systems: Qwen generation, readouts on AR-trained and directly supervised tokens, and the full-evidence MLP with either a complete training-derived template or learned attribute heads. Three comparisons address different questions. The two MLP constructions test attribute prediction while preserving the primary classifier. Qwen and its paired readout receive the same frozen tokens, testing deployment decoders. AR-trained and directly supervised interfaces use the same readout architecture and training-only taxonomy rules, testing interface-training routes. The readouts use the generator's sampler and 100,424 view microsteps. Appendix G specifies the record construction and scoring.

RQ3: evidence and compression. Controlled validation runs remove the healthy reference or vary FISHER and temporal evidence, the training objective, and prefix size. Healthy-reference and branch comparisons under the final AR recipe use three fixed runs; token-count selection uses a single run per configuration. We summarize paired uncertainty in Appendix D. Separate supplementary studies compare shared and setting-local classifiers and seven encoder interfaces. The encoder study varies the encoder, preprocessing, and corresponding adapter as a complete configuration (Appendix F).

6 Results

6.1 Known-setting diagnosis

Table 1 reports the full test comparison; Appendix Figure 2 visualizes the per-setting profile. RELATE-MLP uses the full 24,576-dimensional relative-FISHER representation and leads every aggregate classification metric, reaching 80.60/78.36% setting-macro accuracy/F1. RELATE-Gen reaches 78.65/77.01% through a 4-token interface and directly predicts the complete source-supported record. Appendix Table 13 collects uncertainty estimates for the paired comparisons. RELATE-MLP exceeds FISHER 1-NN by 5.36 points and RELATE-Gen by 1.94 points in setting-macro accuracy. Train-majority, a train-prior sample, and a fixed hard-vote ensemble reach 24.66/10.88%, 23.64/21.67%, and 73.93/70.35% setting-macro accuracy/F1, respectively. Qwen3.5-Omni-Plus, Gemini 3.8 Flash, and GPT Audio Mini reach 25.14–26.77% accuracy under fixed candidate-conditioned configurations; Appendix Table 17 reports the full comparison. Setting-macro accuracy weights the twelve diagnostic settings equally; FISHER 1-NN exceeds RELATE-Gen on group-micro accuracy (89.83% versus 88.85%). At the setting level, RELATE-MLP is best or tied on nine settings. RELATE-Gen is strongest on PU current and ties the best result on PU vibration and UMGED voltage. The two routes' complete-record results are compared in Section 6.2.

Within each RELATE route, parameters are shared across all 12 settings. The MLP uses a shared 42-index classifier with setting-local candidate masks; the generator shares one signal adapter and Qwen decoder, conditioned on modality, machine, and candidate labels. Their frontend processes the evaluated 16–64 kHz sources without resampling waveforms to a common rate. Table 1 also includes baselines trained under their respective recipes (Appendix E). Their per-view waveform windows and aggregation rules are listed in Appendix Table 16.

Table 1: MOSAIC frozen-test results over 3,598 benchmark groups. The setting block reports group-level accuracy; the aggregate block reports setting-macro accuracy, setting-local class-macro F1 averaged over settings, source-family macro accuracy, and group-micro accuracy. All values are percentages. Bold marks the best result in each row and underline marks the second best; ties receive the same mark. Detailed per-setting F1 appears in Appendix Table 11. Selection rules for learned methods appear in Appendix E; the proprietary Omni API baseline is zero-shot.

	Signal classifiers		Pretrained representations				Language-model baselines				RELATE	
Setting	WDCNN	MiniR.	MOMENT	TEST	FISHER FT	FISHER 1-NN	FD–LLM	FD–MVLLM	Omni ZS	Qwen2 sup.	MLP (full)	Gen. (4 tok.)
Per-setting accuracy												
IDMT–AIR	34.50	59.75	45.25	29.00	79.75	**89.50**	48.00	32.50	26.25	58.50	**89.50**	85.75
MaFaulDa (sound)	32.41	82.76	22.41	14.14	72.07	80.00	39.31	62.76	10.00	54.83	**85.52**	82.07
MaFaulDa (vib.)	44.70	91.20	34.00	27.90	84.10	97.50	31.70	44.30	10.00	86.00	**97.60**	93.20
PU (current)	36.11	38.89	33.33	33.33	58.33	44.44	50.00	33.33	33.33	47.22	61.11	**75.00**
PU (vibration)	66.67	61.11	44.44	55.56	**83.33**	80.56	50.00	61.11	38.89	69.44	**83.33**	**83.33**
SDUST (bearing)	88.75	94.00	59.00	54.50	98.50	97.25	60.00	83.50	40.25	**99.00**	97.50	97.00
SDUST (gear)	83.86	98.57	62.43	41.86	**100.00**	99.43	40.71	80.00	15.57	98.86	**100.00**	99.29
UMGED (current)	25.00	18.75	25.00	25.00	31.25	50.00	25.00	18.75	25.00	43.75	**62.50**	43.75
UMGED (vibration)	50.00	50.00	46.88	37.50	**75.00**	68.75	34.38	40.63	25.00	65.63	59.38	68.75
UMGED (voltage)	25.00	46.43	28.57	28.57	35.71	39.29	21.43	25.00	25.00	39.29	**60.71**	**60.71**
WT gearbox	80.00	66.67	50.00	56.67	88.33	81.67	58.33	68.33	20.00	86.67	**90.00**	80.00
MIMII-DG	54.00	62.17	59.67	51.67	76.00	74.50	54.00	58.83	52.00	74.83	**80.00**	75.00
Aggregate metrics												
Setting-macro Acc.	51.75	64.19	42.58	37.97	73.53	75.24	42.74	50.75	26.77	68.67	**80.60**	78.65
Setting-macro F1	49.13	62.63	37.48	31.06	73.48	72.13	40.52	48.64	14.79	65.19	**78.36**	77.01
Source-macro Acc.	58.05	69.57	47.12	41.19	79.85	80.78	46.69	56.51	28.15	74.52	**85.21**	82.70
Group-micro Acc.	57.17	81.66	47.30	37.44	85.10	89.83	43.47	58.37	24.26	81.24	**91.83**	88.85

6.2 STRUCTURED RECORD PREDICTION

One decoder serves all 12 settings with their machine context and candidate lists. On 3,598 test groups, RELATE-Gen reaches 79.95% Field EM and 68.86% Joint EM after group aggregation. Table 2 compares generation, both token interfaces under the same readout, and two full-evidence MLP record constructions. Detailed scoring rules appear in Appendix G.

Table 2: Single-run structured test results on 3,598 groups (%). Acc. and F1 are setting-macro; Joint EM is setting-macro after independent per-field voting. Field EM weights seven fields equally and Severity EM weights two fields equally. Both readouts use the same train-only taxonomy rules; MLP records use a complete train-only template or learned severity heads with the template's remaining fields.

Route	Acc.	F1	Field EM	Joint EM	Severity EM
RELATE-Gen	78.65	77.01	79.95	68.86	52.48
AR-token readout + taxonomy rules	77.80	76.19	79.70	69.07	51.78
Direct-token readout + taxonomy rules	76.29	74.49	78.90	66.70	48.00
RELATE-MLP + complete template	80.60	78.36	72.70	64.84	22.23
RELATE-MLP + severity heads	80.60	78.36	77.73	66.85	39.84

The two MLP constructions share the primary classifier, with the severity-head variant predicting attributes from its hidden features. The token interfaces use field supervision and temporal evidence. The MLP routes lead primary accuracy, Gen leads Field and Severity EM, and the AR-token readout leads Joint EM at 69.07%, followed by Gen at 68.86% and the Direct-token readout at 66.70%.

On the test groups supporting each field, Gen scores 77.33% on damage size and 27.63% on eccentricity, compared with 56.00% and 23.68% for the MLP severity heads (Appendix Table 23); paired uncertainty appears in Appendix Table 13.

6.3 INTERFACE LEARNING AND ATTRIBUTE PREDICTION

On the complete 24,576-dimensional relative-FISHER representation, RELATE-MLP reaches 78.38% validation accuracy and 64.84% Joint EM with its train-only complete template (Table 3).

Two linear attribute heads on the frozen primary classifier's hidden features raise Severity EM from 26.49% to 41.73%, Field EM from 72.74% to 77.10%, and Joint EM from 64.84% to 65.33% (Appendix G).

Under the same readout and taxonomy rules, AR- and Direct-trained interfaces reach 65.65% and 65.52% validation Joint EM, respectively (Appendix G).

Across 88 eligible validation groups with within-condition attribute variation, readouts on AR-trained and directly supervised tokens reach 49.43% and 48.96% key-balanced severity score; the primary-conditioned training-set mode template reaches 18.75%. Appendix G gives the field-level results and matched interface-training controls. This analysis covers nine training keys for which a primary condition occurs with multiple damage-size or eccentricity values.

6.4 EVIDENCE INTERFACE AND COMPRESSION

Across three validation runs, adding the MiniROCKET branch and learned fusion improves diagnosis under the final autoregressive objective (Table 3). The paired interface reaches $74.73 \pm 1.20\%$ accuracy, 4.56 points above FISHER-only evidence across the three paired runs; MiniROCKET-only evidence reaches 35.67%.

The frozen-token readout holds the autoregressively learned interface fixed and replaces Qwen at inference. On identical frozen tokens, it closely matches Qwen's condition and field scores (Table 3) and reduces latency and memory in the cached-evidence measurements (Appendix Table 18). This provides a low-cost decoder for fixed-schema deployment; Qwen generates the record directly from the tokens and diagnostic prompt.

Table 3: Validation comparison of compact interfaces and full relative-FISHER baselines. Values are three-run mean $\pm$ standard deviation except the single fixed configurations marked †. Field EM averages exact match across the seven supported fields; Joint EM requires the primary condition and every supported field to match. Bold marks the best value in each column.

Configuration	Acc.	F1	Field EM	Joint EM
FISHER-only generator	70.18 ± 1.90	68.09 ± 1.50	74.82 ± 2.89	60.13 ± 2.43
MiniROCKET-only generator	35.67 ± 1.37	29.99 ± 1.70	53.49 ± 0.63	33.45 ± 1.36
RELATE-Gen	74.73 ± 1.20	72.72 ± 1.38	78.17 ± 1.10	64.64 ± 2.23
Frozen-token readout	74.46 ± 1.31	72.32 ± 1.48	79.19 ± 1.25	64.72 ± 0.87
Full relative-FISHER structured MLP	70.21 ± 1.32	68.02 ± 1.18	$\mathbf{79.78 \pm 0.25}$	59.56 ± 0.99
RELATE-MLP + complete template†	**78.38**	**76.03**	72.74	64.84
RELATE-MLP + severity heads†	**78.38**	**76.03**	77.10	**65.33**

Across three fixed validation runs under the final autoregressive objective, healthy-reference evidence raises setting-macro accuracy from $69.66 \pm 1.79\%$ to $74.73 \pm 1.20\%$, a mean gain of 5.07 points that is positive for every paired run. The no-reference control keeps the temporal branch and input geometry, replacing the FISHER blocks with $[q, q, 0, 0]$.

The input controls show different effects of explicit differences. The MLP reaches 78.38% accuracy with $[q, r]$, versus 75.83% with q and 77.31% with all four blocks. Qwen reaches 75.75% with all four blocks, versus 73.32% when the differences are zeroed (Appendix Tables 19 and 20). Thus, the single-run controls favor the pair for MLP and explicit differences for the compact Qwen route.

Accuracy rises from 70.90% with 1 token to 75.75% with 4 tokens; increasing to 8 tokens yields 75.36% in this single-seed sweep. We select 4 tokens as the compact validation point. The selected interface exceeds an expanded continuous prefix by 4.11 points in the matched single-seed comparison.

Table 4: Fixed-configuration validation controls for training objective, relative evidence, and continuous bottleneck size.

Configuration	Acc.	F1
Fusion + AR objective + auxiliary primary CE	71.79	70.45
Fusion + autoregressive objective	75.75	73.94
No healthy reference + autoregressive objective	67.91	66.53
1-token evidence bottleneck	70.90	68.68
2-token evidence bottleneck	73.43	71.79
4-token evidence bottleneck	75.75	73.94
8-token evidence bottleneck	75.36	73.36
Expanded evidence prefix	71.64	69.89

Under the auxiliary-loss configuration used for the encoder comparison, the replacements rank differently under linear probing and generation. OpenBEATs-large leads the setting-local linear probe at 77.58% accuracy and reaches 61.29% after fusion and generation. Source-rate FISHER-small reaches 71.79%; EAT-base is the best 16 kHz replacement at 69.28%. Appendix F reports the full comparison.

6.5 Deployment from cached evidence

The two decoders offer different inference costs on the same learned tokens. Table 5 measures batch-1 decoding on 92 validation views using one NVIDIA A10. The readout processes 45.95 views per second; autoregressive generation processes 0.20. Measurements include the signal adapter and decoder, with FISHER and MiniROCKET features cached in advance. Appendix E gives the timing protocol and batch-8 results.

Table 5: Batch-1 inference from cached evidence. Memory is peak allocated GPU memory; latency is the median per batch.

Decoder	Latency (ms)	Views/s	Memory (GiB)
RELATE-Gen	4789.82	0.20	5.428
Frozen-token readout	23.61	45.95	0.078

6.6 CWRU-specific adaptation

On 222 shared CWRU segments, RELATE-CWRU reaches 83.33% accuracy / 80.95% macro F1, versus 43.24% / 31.49% for locally adapted BearLLM (Peng et al., 2025). This single-FISHER-branch adaptation adds training-only primary CE for the scored ten-class diagnosis alongside AR record generation (Appendix C).

7 Conclusion

RELATE pairs query signals with healthy references for classification and structured diagnosis across twelve known settings. Its full-evidence MLP reaches 80.60% setting-macro accuracy; its compact Qwen route reaches 68.86% group-aggregated Joint EM. Validation controls show gains from healthy references and from adding the temporal branch with learned fusion. Explicit differences improve the tested Qwen route, while the MLP control favors the pair alone. For attribute prediction, Gen achieves higher damage-size and eccentricity exact match than the MLP severity heads on the evaluated test groups.

Reproducibility Statement

The appendix specifies the benchmark, evaluation, and training configurations, including the group-level scorer and supported fields. Compute details appear in Appendix H.

AI Use Statement

Generative AI tools assisted experimental design, method implementation, results interpretation, reference checks, figure preparation, and manuscript drafting and editing. The authors reviewed the AI-assisted work, made the research decisions, and take responsibility for the code, results, claims, citations, and final text.

References

Shikhar Bharadwaj, Samuele Cornell, Kwanghee Choi, Satoru Fukayama, Hye-Jin Shim, Soham Deshmukh, and Shinji Watanabe. OpenBEATs: A fully open-source general-purpose audio encoder. In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–5, 2025. doi: 10.1109/WASPAA66052.2025.11230965. URL https://doi.org/10.1109/WASPAA66052.2025.11230965.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022. doi: 10.48550/arXiv.2212.09058. URL https://arxiv.org/abs/2212.09058.

Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. EAT: Self-supervised pre-training with efficient audio transformer. *arXiv preprint arXiv:2401.03497*, 2024. doi: 10.48550/arXiv.2401.03497. URL https://arxiv.org/abs/2401.03497.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. doi: 10.48550/arXiv.2407.10759. URL https://arxiv.org/abs/2407.10759.

COPPE/Poli/UFRJ. MaFaulDa: Machinery fault database. Official dataset page, n.d. URL https://www02.smt.ufrj.br/~offshore/mfs/page_01.html. Accessed 2026-08-29.

Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. MiniRocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 248–257, 2021. doi: 10.1145/3447548.3467231. URL https://doi.org/10.1145/3447548.3467231.

Kota Dohi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, Yuki Nikaido, and Yohei Kawaguchi. MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2022. URL https://arxiv.org/abs/2205.13879. arXiv:2205.13879.

Pingyi Fan, Anbai Jiang, Shuwei Zhang, Xinhu Zheng, Zhiqiang Lv, Bing Han, Wenrui Liang, Junjie Li, Wei-Qiang Zhang, Yanmin Qian, Xie Chen, and Jia Liu. FISHER: A foundation model for multimodal industrial signal comprehensive representation. *IEEE Transactions on Industrial Informatics*, 2026. doi: 10.1109/TII.2026.3698554. URL https://doi.org/10.1109/TII.2026.3698554. arXiv:2507.16696.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 16115–16152. PMLR, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

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://openreview.net/forum?id=nZeVKeeFYf9.

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. doi: 10.48550/arXiv.2310.01728. URL https://arxiv.org/abs/2310.01728.

David Johnson, Jakob Kirner, Sascha Grollmisch, and Judith Liebetrau. IDMT-ISA-Compressed-Air Dataset, 2020. URL https://zenodo.org/records/7551606.

Dayang Li, Zhibo Pang, Yongxu Chen, Kun Yang, Jinyan Shao, Yichen Luo, Yanghang Zeng, Chen He, and Yutong Gao. FD-MVLLM: Fault diagnosis based on multimodal vibration data and large language model for bearing system. *Mechanical Systems and Signal Processing*, 239:113226, October 2025a. doi: 10.1016/j.ymssp.2025.113226. URL https://doi.org/10.1016/j.ymssp.2025.113226.

Jiaming Li, Hao Chen, Xian-Bo Wang, and Zhi-Xin Yang. A comprehensive gear eccentricity dataset with multiple fault severity levels: Description, characteristics analysis, and fault diagnosis applications. *Mechanical Systems and Signal Processing*, 224:112068, 2025b. doi: 10.1016/j.ymssp.2024.112068. URL https://doi.org/10.1016/j.ymssp.2024.112068.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Dongdong Liu, Lingli Cui, and Weidong Cheng. A review on deep learning in planetary gearbox health state recognition: Methods, applications, and dataset publication. *Measurement Science and Technology*, 35(1):012002, 2024. doi: 10.1088/1361-6501/acf390. URL https://doi.org/10.1088/1361-6501/acf390.

Jiale Liu, Dandan Peng, Huan Wang, Chenyu Liu, Yan-Fu Li, and Min Xie. Aerogpt: Leveraging large-scale audio model for aero-engine bearing fault diagnosis. *IEEE Transactions on Cybernetics*, pp. 1–14, 2026. doi: 10.1109/TCYB.2026.3668256. URL https://doi.org/10.1109/TCYB.2026.3668256. arXiv:2506.16225.

Paderborn University Bearing Data Center. Bearing data center: Data sets and download. Official dataset page, n.d. URL https://mb.uni-paderborn.de/en/kat/research/bearing-datacenter/data-sets-and-download. Accessed 2026-08-29.

Haotian Peng, Jiawei Liu, Jinsong Du, Jie Gao, and Wei Wang. BearLLM: A prior knowledge-enhanced bearing health management framework with unified vibration signal representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19866–19874, 2025. doi: 10.1609/aaai.v39i19.34188.

Hamzah A. A. M. Qaid, Bo Zhang, Dan Li, See-Kiong Ng, and Wei Li. FD-LLM: Large language model for fault diagnosis of machines. *arXiv preprint arXiv:2412.01218*, 2024. URL https://arxiv.org/abs/2412.01218.

Shandong University of Science and Technology. SDUST dataset for fault diagnosis. Official GitHub repository, n.d. URL https://github.com/JRWang-SDUST/SDUST-Dataset. Accessed 2026-08-29.

Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. TEST: Text prototype aligned embedding to activate LLM's ability for time series. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tuh4nZVb0g.

Jie Wang, Tianrui Li, Yan Yang, Shiqian Chen, and Wanming Zhai. Diagllm: multimodal reasoning with large language model for explainable bearing fault diagnosis. *Science China Information Sciences*, 68(6):160103, 2025. doi: 10.1007/s11432-024-4333-7. URL https://doi.org/10.1007/s11432-024-4333-7.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022. doi: 10.1609/aaai.v36i8.20881. URL https://doi.org/10.1609/aaai.v36i8.20881.

Xiao-Min Zeng, Yan Song, Ian McLoughlin, Lin Liu, and Li-Rong Dai. Robust prototype learning for anomalous sound detection. In *Interspeech*, 2023. doi: 10.21437/Interspeech.2023-1173.

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://doi.org/10.3390/s17020425.

A MOSAIC BENCHMARK CONTRACT

MOSAIC contains eleven RMIS diagnostic settings and one machine-conditional MIMII-DG setting. A *benchmark group* is the unit used for splitting, view aggregation, and evaluation; a *view* is one source file, segment, or synchronized channel associated with that group. Every view from a benchmark group remains in one split. The reported results characterize known-setting closed-set diagnosis at the benchmark-group level. Source identifiers link 2,395 of 2,998 identifiable test groups to acquisition families also present in training; the remaining 600 MIMII-DG groups lack family identifiers. Benchmark groups define the split units; source families can span splits. Here, a family is a source-metadata key: the pre-segmentation record for IDMT; the source stem with channel suffixes removed for MaFaulDa, SDUST, and WTPG; bearing identity, damage origin, speed, torque, and radial force for PU; and serial number, tooth-damage code, eccentricity, drivetrain, and speed for UMGED. Removing channel suffixes links corresponding MaFaulDa sound and vibration records. These source-specific keys identify shared record provenance or operating configurations. The overlap audit matches these metadata keys across splits. Candidate conditions are indexed by diagnostic setting, so identical text in two settings does not merge their classes. Source-family macro accuracy uses eight families: IDMT compressed air (Johnson et al., 2020), MaFaulDa (COPPE/Poli/UFRJ, n.d.), the Paderborn University Bearing Data Center (Paderborn University Bearing Data Center, n.d.), SDUST bearing, SDUST gear (Shandong University of Science and Technology, n.d.), UMGED (Li et al., 2025b), WTPG (Liu et al., 2024), and MIMII-DG (Dohi et al., 2022). SDUST bearing and SDUST gear are separate families in this metric.

A.1 RELATIONSHIP TO RMIS

MOSAIC defines a task and evaluation layer over admitted source data. It reuses the 11 RMIS settings and adds the protocol summarized in Table 6.

MOSAIC uses a class-balanced evaluation subset of the admitted RMIS data. Within each setting, the test-group quota per primary class is the smaller of 100 and the number of groups in the least frequent class of the original test pool. Each selected group contributes at most five views; views are selected round-robin across groups, with at most 100 views per class. Validation uses an equal group quota per class, derived from 20% of the smallest training-pool class and bounded between two and twenty groups. The remaining training groups are retained, with at most five views per group. These rules balance class representation and limit repeated views from the same group, while retaining different setting sizes according to source availability. Together with the setting exclusions below and the addition of MIMII-DG, they explain the count differences between RMIS and MOSAIC.

Table 6: Direct relationship between RMIS and MOSAIC.

Aspect	RMIS source assets used here	MOSAIC protocol
Signals	Eleven admitted RMIS diagnostic settings	The same eleven settings plus MIMII-DG
Source metadata	Original modality, condition, and diagnostic fields	Benchmark-group and view bindings
Split	Source group metadata retained	One frozen train/validation/test assignment at benchmark-group level
Output space	Source-specific diagnostic labels	Setting-local candidates and source-supported structured fields
Evaluation	Source/task-specific evaluation	Gold-separated generated predictions and one group-level scorer
IIEE exclusion	RMIS labels are good, broken, and heavy load	Heavy load is an operating condition rather than a fault class, so IIEE is outside the selected fault/condition taxonomy
UMGED sound exclusion	Eleven severity labels with two validation and two test groups per class	It does not meet the protocol's minimum of four test groups per class
ToyADMOS exclusion	Machine-specific factor combinations form 54, 95, and 55 primary classes	These combinations fall outside the shared setting-local condition/fault task

Table 7: Semantic contract for the 12 diagnostic settings in MOSAIC. Only source-supported fields are scored.

Setting	Modality and source	Candidate conditions	Supported fields
idmt_air	sound; IDMT	normal; tube leak; vent leak; low-pressure vent leak	health state, leak location, leak mechanism
mafaulda_sound	sound; MaFaulDa	normal; rotor imbalance; horizontal/vertical shaft misalignment; underhang/overhang bearing cage, outer-race, or rolling-element fault	health state, fault location, fault mechanism
mafaulda_vib	vibration; MaFaulDa	the same ten conditions as mafaulda_sound	health state, fault location, fault mechanism
pu_cur	current; PU bearing	healthy; inner-race fault; outer-race fault	health state, fault location, fault mechanism
pu_vib	vibration; PU bearing	healthy; inner-race fault; outer-race fault	health state, fault location, fault mechanism
sdust_bearing	vibration; SDUST bearing	healthy; inner-race, outer-race, or rolling-element fault	health state, fault location, fault mechanism, damage size
sdust_gear	vibration; SDUST gear	healthy; planet/sun gear tooth fracture, pitting, or wear	health state, gear location, fault mechanism
umged_cur	current; UMGED	healthy; eccentricity; chipped tooth; compound eccentricity and chipped tooth	health state, fault location, fault mechanism, eccentricity
umged_vib	vibration; UMGED	the same four conditions as umged_cur	health state, fault location, fault mechanism, eccentricity
umged_vol	voltage; UMGED	the same four conditions as umged_cur	health state, fault location, fault mechanism, eccentricity
wt_plane_gearbox	vibration; WTPG	healthy; broken tooth; missing tooth; root crack; tooth wear	health state, fault location, fault mechanism
mimii_dg	sound; MIMII-DG	machine-local normal or anomalous operation for bearing, fan, gearbox, slider, and valve	record health state, machine type, MIMII semantic health state

For idmt_air, the three non-leakage source conditions map to normal operation, while the three leakage conditions retain their device type. For sdust_bearing, the source fault code determines the primary location and the 0.2, 0.4, or 0.6 mm suffix remains a separate field. For UMGED, gear and eccentricity codes jointly define the four candidates; eccentricity magnitude is evaluated as a separate field. MIMII-DG section, domain, and operating attributes remain scorer-side provenance.

Table 8: Frozen train/validation/test counts by diagnostic setting, reported as benchmark groups/views. Every view from one benchmark group remains in one split.

Setting	Train	Validation	Test	Group/view construction
idmt_air	2,716/8,156	80/241	400/400	source record with sound segments
mafaulda_sound	935/935	40/40	290/290	one view per source group
mafaulda_vib	5,653/5,653	200/200	1,000/1,000	one view per source group
pu_cur	58/290	6/30	36/180	five views per source group
pu_vib	58/290	6/30	36/180	five views per source group
sdust_bearing	1,063/4,252	80/320	400/400	four train/validation views; one test view
sdust_gear	1,318/2,720	140/288	700/700	two or four train/validation views; one test view
umged_cur	443/2,215	12/60	16/80	five views per source group
umged_vib	816/4,080	28/140	32/160	five views per source group
umged_vol	547/2,735	16/80	28/140	five views per source group
wt_plane_gearbox	70/350	10/50	60/300	five views per source group
mimii_dg	1,800/1,800	600/600	600/600	one archive member per group
Total	15,477/33,476	1,218/2,079	3,598/4,430	

For the RMIS component, MOSAIC preserves the frozen train/test assignment of the admitted RMIS groups and uses the frozen validation subset drawn from the original training partition with class coverage. Splits are assigned at the benchmark-group level before views are expanded, so every segment or channel bound to a group remains in the same split.

View multiplicity follows the source records. MaFaulDa and MIMII-DG contribute one view per group; PU and UMGED contribute five source-group views; WTPG contributes five segments per

source group; IDMT groups contain source-record segments; and the frozen SDUST test manifests retain one view per group while their train/validation groups retain multiple channels or segments.

The MIMII-DG component is derived from 3,000 official test recordings. Source health labels define strata within each machine–section–domain combination; a deterministic assignment places 30/10/10 groups from every stratum into train/validation/test. These labels are training targets and split strata, while the model input contains only signal evidence, modality, machine description, and candidate labels. This machine-local diagnosis protocol is distinct from the original DCASE Task 2 evaluation.

B Per-Setting Results

Table 11 reports per-setting test F1 over all 3,598 benchmark groups. Per-setting accuracy and aggregate metrics appear in Table 1. Table 11 uses the same twelve methods as the main table; the Omni zero-shot column covers all 3,598 benchmark groups.

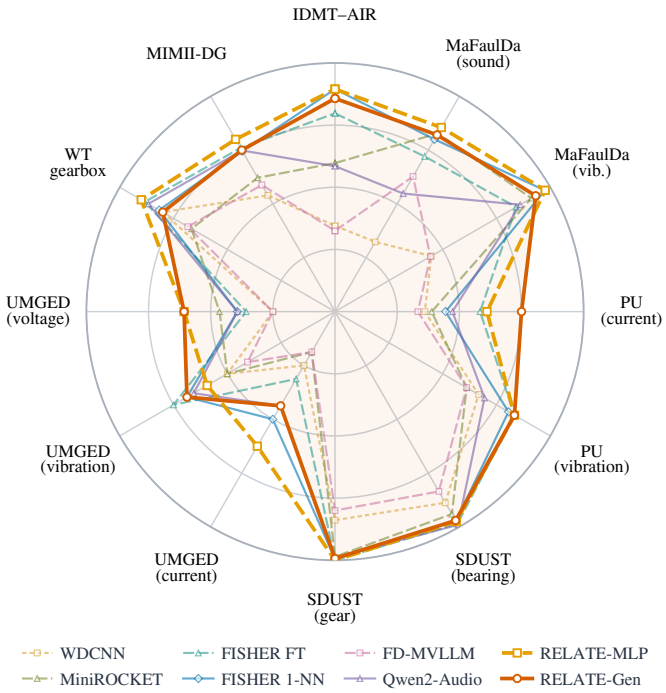

Figure 2: Frozen-test accuracy across 12 settings for eight representative methods spanning the four families in Table 1. The two RELATE routes are emphasized; all spokes span 0–100% with 25-point rings. Table 1 reports all 12 methods.

Table 9: MOSAIC validation on 1,218 benchmark groups. Acc. is setting-macro accuracy, F1 is the mean of setting-local class-macro F1 values, and Src. is source-family macro accuracy. All values are percentages.

Method	Acc.	F1	Src.
Stats15–SVM	45.12	43.50	51.54
WDCNN	57.19	54.73	61.76
MiniROCKET classifier	61.47	58.42	69.72
MOMENT-1-base full fine-tuning	50.08	44.75	55.72
TEST-GPT2 adaptation	40.13	33.00	43.95
FD-LLM-Stats 7B	42.20	37.82	45.78
FD-MVLLM	56.87	54.76	61.42
FISHER setting-local MLP	71.54	68.48	79.56
FISHER 1-NN, all training groups	71.49	66.79	77.53
Supervised Qwen2-Audio	72.56	68.20	78.16
RELATE-MLP	**78.38**	**76.03**	**83.71**
RELATE-Gen	75.75	73.94	80.95

Table 10: Frozen-test results (%) for training-distribution baselines, the fixed three-model hard-vote ensemble, and RELATE-Gen. Prior is the frozen seed-111 train-prior sample. The aggregate rows report setting-macro accuracy, setting-macro F1, and source-family macro accuracy.

Setting	Train maj.	Prior	Ensemble	RELATE-Gen
IDMT–AIR	25.00	24.75	74.25	85.75
MaFaulDa (sound)	10.00	8.62	80.69	82.07
MaFaulDa (vib.)	10.00	10.90	95.70	93.20
PU (current)	33.33	38.89	41.67	75.00
PU (vibration)	33.33	25.00	75.00	83.33
SDUST (bearing)	25.00	24.00	99.00	97.00
SDUST (gear)	14.29	13.29	100.00	99.29
UMGED (current)	25.00	31.25	43.75	43.75
UMGED (vibration)	25.00	25.00	68.75	68.75
UMGED (voltage)	25.00	21.43	42.86	60.71
WT gearbox	20.00	10.00	88.33	80.00
MIMII-DG	50.00	50.50	77.17	75.00
Setting-macro Acc.	24.66	23.64	73.93	78.65
Setting-macro F1	10.88	21.67	70.35	77.01
Source-macro Acc.	25.33	23.77	79.63	82.70

Table 11: MOSAIC frozen-test class-macro F1 by setting (%), computed after view aggregation over the same 3,598 benchmark groups used in Table 1.

Panel A: conventional, pretrained, and general language-model baselines

Setting	WDCNN	MiniR.	MOMENT	TEST	FD–LLM	Omni ZS
idmt_air	32.61	59.20	41.54	21.59	43.06	16.84
mafaulda_sound	30.15	81.53	19.19	7.26	38.48	1.85
mafaulda_vib	44.53	91.26	31.06	25.26	31.19	1.88
pu_cur	34.23	36.40	23.33	16.67	46.31	16.67
pu_vib	66.40	60.71	35.60	50.73	52.20	26.92
sdust_bearing	88.73	94.06	58.67	52.30	60.49	27.43
sdust_gear	83.97	98.58	62.55	36.83	40.96	6.32
umged_cur	16.27	10.00	10.53	17.14	15.97	10.00
umged_vib	41.68	50.58	46.57	25.04	33.38	10.00
umged_vol	17.86	41.02	18.04	23.24	17.24	10.00
wt_plane_gearbox	79.31	66.19	46.05	53.26	55.16	6.67
mimii_dg	53.86	61.99	56.60	43.38	51.82	42.94
Setting macro	49.13	62.63	37.48	31.06	40.52	14.79

Panel B: signal-representation and trained multimodal methods

Setting	FISHER FT	FISHER 1-NN	FD-MVLLM	Qwen2 sup.	RELATE-MLP	RELATE-Gen
idmt_air	79.61	89.59	32.19	57.66	89.54	85.66
mafaulda_sound	71.63	79.57	62.58	52.30	85.05	81.32
mafaulda_vib	84.35	97.48	43.16	85.54	97.61	93.20
pu_cur	58.82	41.79	27.81	44.50	61.68	74.91
pu_vib	83.41	80.94	62.32	68.53	83.16	83.41
sdust_bearing	98.50	97.26	83.62	99.00	97.50	97.00
sdust_gear	100.00	99.43	79.96	98.86	100.00	99.28
umged_cur	31.25	33.33	10.00	29.02	53.33	37.25
umged_vib	74.39	63.54	38.02	59.45	51.13	63.54
umged_vol	35.38	26.36	19.41	26.39	51.11	53.86
wt_plane_gearbox	88.43	81.90	66.56	86.21	90.30	79.75
mimii_dg	75.98	74.40	58.08	74.78	79.91	74.97
Setting macro	73.48	72.13	48.64	65.19	**78.36**	<u>77.01</u>

C CWRU BEARING COMPARISON

We reconstruct BearLLM's ten-class CWRU diagnosis task using non-overlapping one-second, 12-kHz segments. The 7:2:1 split within each source–class stratum contains 1,554 training, 444 validation, and 222 test segments. Both systems use the same segment manifests and candidate labels; checkpoints are selected on validation data and scored on the same test segments.

The CWRU-specific RELATE adaptation uses one frozen FISHER-relative branch and a Normal-K1 healthy reference. Reference matching uses source dataset, sensor, rotation speed, and load; a fixed order selects one training-normal segment per query. Qwen2.5-7B with LoRA learns to generate the diagnostic record with objective $\mathcal{L}_{\mathrm{AR}} + 0.3\mathcal{L}_{\mathrm{primary}}$, where the second term is candidate-masked cross-entropy from an auxiliary head that supervises the ten-class primary condition during training and is removed at inference. Autoregressive training learns the diagnostic record sequence. Generated validation accuracy selects step 4,656.

BearLLM starts from its released MBHM/vibration adapters, official DCN, and Qwen2.5-1.5B, then receives local CWRU LoRA adaptation; validation accuracy selects epoch 44. Table 12 compares these separately adapted systems on the shared test segments.

Table 12: CWRU ten-class diagnosis on the same 222 one-second test segments. Both systems use the same candidate labels and evaluation unit, with their respective encoders and adaptation objectives.

Method	Accuracy (%)	Macro F1 (%)
BearLLM, locally adapted (Peng et al., 2025)	43.24	31.49
RELATE-CWRU, AR + auxiliary CE	**83.33**	**80.95**

D UNCERTAINTY SUMMARY

Table 13 reports paired differences for the main comparisons. The Gen–MLP severity-head Joint-EM interval spans zero, as do the validation intervals for taxonomy-corrected AR–Direct Joint EM and the MLP severity-head/template comparison. These estimates describe the uncertainty in the observed differences conditional on the evaluated groups and fixed trained models.

Table 13: Paired-bootstrap uncertainty for the comparisons discussed in the main text. Differences and 95% intervals are in percentage points and are computed from unrounded results. Validation comparisons aggregate three fixed trained pairs, except the single-run MLP severity-head/template comparison. The two structured-test comparisons use 10,000 setting-stratified paired group resamples, conditional on the fixed predictions and observed benchmark groups.

Comparison	Split and metric	Difference	95% interval
RELATE-Gen − FISHER 1-NN	Test; setting-macro accuracy	$+3.41$	$[0.02, 6.94]$
RELATE-Gen − FISHER 1-NN on PU current	Test; accuracy	$+30.56$	$[13.89, 47.22]$
RELATE-Gen − FISHER 1-NN on UMGED voltage	Test; accuracy	$+21.43$	$[7.14, 35.80]$
RELATE-MLP − FISHER 1-NN	Test; setting-macro accuracy	$+5.36$	$[2.51, 8.43]$
RELATE-MLP − FISHER 1-NN	Test; source-family macro accuracy	$+4.42$	$[2.39, 6.58]$
RELATE-MLP − FISHER 1-NN	Test; group-micro accuracy	$+2.00$	$[1.11, 2.86]$
RELATE-MLP − RELATE-Gen	Test; setting-macro accuracy	$+1.94$	$[-0.56, 4.60]$
RELATE-MLP − RELATE-Gen	Test; source-family macro accuracy	$+2.51$	$[0.80, 4.31]$
RELATE-MLP − RELATE-Gen	Test; group-micro accuracy	$+2.97$	$[2.03, 3.92]$
RELATE-Gen − MLP severity heads	Test; setting-macro Joint EM	$+2.02$	$[-0.89, 4.98]$
AR − direct after equal taxonomy correction	Test; setting-macro Joint EM	$+2.37$	$[0.04, 4.83]$
FISHER+MiniROCKET − FISHER only	Validation; accuracy	$+4.56$	$[2.17, 6.95]$
Healthy reference − no reference	Validation; accuracy	$+5.07$	$[1.82, 8.53]$
AR − direct interface	Validation; setting-macro Joint EM	$+3.31$	$[0.76, 6.13]$
AR − direct after equal taxonomy correction	Validation; setting-macro Joint EM	$+0.13$	$[-2.08, 2.35]$
AR readout − severity mode template	Validation; key-balanced severity score	$+30.68$	$[21.86, 39.97]$
RELATE-MLP severity heads − complete template	Validation; setting-macro Joint EM	$+0.49$	$[-1.99, 3.01]$

Table 14: AR-trained minus Direct-trained readout changes in mutually exclusive validation outcome categories before taxonomy correction, in setting-macro percentage points. Negative values indicate less mass in an error category; the final row reports complete-record correctness. Categories follow the displayed priority order and partition the change in outcome frequencies.

Category	Seed 111	Seed 222	Seed 333	Mean
Primary wrong	$+0.57$	-1.22	$+0.71$	$+0.02$
Primary correct, taxonomy wrong	-3.06	-2.22	-4.27	-3.19
Severity wrong	-1.43	-0.22	$+1.20$	-0.15
Missing or invalid	0.00	0.00	0.00	0.00
Complete record correct	$+3.91$	$+3.66$	$+2.37$	$+3.31$

E IMPLEMENTATION AND TRAINING

E.1 BASELINE RECIPES

Table 15 records the principal training and selection choices for learned baselines. TEST uses the released official pipeline adapted to MOSAIC's candidate-conditioned classification task.

Input preprocessing follows each model's frontend requirements and the configuration used for its MOSAIC training. Qwen2-Audio's feature extractor requires 16-kHz audio; its supervised adapter uses the same 10-s crop at training and evaluation. MOMENT uses 512-point windows, while TEST-GPT2, WDCNN, and the MiniROCKET classifier use the 2,048-point windows configured for their respective adaptations. Evaluation preserves these window lengths and sampling rates, using the aggregation rules in Table 16. The comparison therefore evaluates trained systems on common benchmark groups with model-specific observation durations.

Table 15: Compact recipes for learned baselines.

Method	Input and model	Optimization	Selection
Supervised Qwen2-Audio	16-kHz mono PCM16; centered 10-s crop; candidate-JSON prompt; NF4 LoRA ($r = 64$, $\alpha = 16$, dropout 0.05)	AdamW, 10^{-4}; 1,000-step warmup; batch 1, accumulation 5; 5 epochs; seed 1234	Fixed final adapter
MOMENT-1-base	Three native-rate 512-point start/center/end windows; random train window and mean-logit evaluation; full encoder/embedder tuning	AdamW, 10^{-4}; weight decay 10^{-4}; OneCycle; batch 16; 5 epochs; seed 111	Best validation setting-macro accuracy, epoch 4
TEST-GPT2 adaptation	Three native-rate 2,048-point windows; official 10-layer TCN and 10×768 prototypes; frozen encoder and first six GPT-2 layers; trainable positional embedding, 8 soft prompts, and heads	1,500-step encoder stage with Adam, 10^{-3}; then AdamW, 10^{-3}, weight decay 10^{-4}, cosine; batch 4; at most 30 epochs; patience 5; seed 111	Best validation setting-macro accuracy, epoch 9
WDCNN	Native-rate 2,048-point windows; five-block network; random training window and three-window mean logits at evaluation	AdamW, 10^{-3}; weight decay 10^{-4}; cosine; batch 256; at most 30 epochs; patience 5	Best validation setting-macro accuracy
MiniROCKET classifier	Setting-local transforms of native-rate 2,048-point windows; mean window features	Balanced RidgeClassifierCV, $\alpha \in \mathrm{logspace}(-3, 3, 10)$	Training-only cross-validation of ridge penalty
FISHER FT	Native-rate 10-s crop; full backbone and shared candidate-masked head	AdamW; backbone 10^{-5}, head 3×10^{-4}; weight decay 0.01; batch 2; 10,000 updates	Best validation setting-macro accuracy at 1,000-update intervals
FD-LLM	Stats15 text; Qwen2.5-7B with 4-bit LoRA ($r = 4$, $\alpha = 16$, dropout 0.05; query/value projections)	Learning rate 10^{-4}; cosine; batch 1, accumulation 16; 4 epochs; maximum sequence length 1,024	Fixed final adapter
FD-MVLLM	Waveform patches, CWT images, and Stats15 prompt; frozen GPT-2 with reprogramming and classification layers	AdamW, 10^{-4}; weight decay 10^{-4}; cosine; batch 8; 30 epochs	Best validation setting-macro accuracy, epoch 25

The 512- and 2,048-point window baselines use setting-specific normalization fitted on training data and retain source-native rates. Across the evaluated 16–64 kHz sources, 512 points span 8–32 ms and 2,048 points span 32–128 ms; shorter signals are symmetrically zero-padded.

E.2 FISHER EVIDENCE LAYOUT AND HEALTHY REFERENCES

The frozen FISHER cache retains band, time, and padding masks. The frontend supports source rates of at least 4 kHz: its 25 ms STFT yields at least one complete 50-bin subband, and its 10 ms hop scales with the source rate. The main pipeline evaluates source rates of 16, 25.6, 48, 50, 51.2, and 64 kHz without common-rate waveform conversion. Each signal view contains 20 time tokens and 4, 6, 12, or 16 valid bands depending on the source setting. Every band-time channel has width 384. Concatenating the query, healthy prototype, signed residual, and residual magnitude produces a 1,536-dimensional relative representation at each band-time position.

Eligible references are healthy training views from the same diagnostic setting, machine axis where applicable, and sampling-rate pool. A mean-pooled, L2-normalized FISHER vector supplies cosine-

Table 16: Waveform evidence supplied per evaluation view. Three-window methods use start, center, and end windows. Logit/score methods average within-view window outputs and then average view outputs before the group decision; record-valued methods vote over view-level records. Observation durations differ across the listed model adaptations.

Method	Waveform supplied per view	Within-view output
WDCNN	Three native-rate 2,048-point windows	Mean logits
MiniROCKET classifier	Three native-rate 2,048-point windows	Mean 9,996-D features
MOMENT	Three native-rate 512-point windows	Mean logits
TEST-GPT2	Three native-rate 2,048-point windows	Mean logits
FISHER FT / 1-NN	One native-rate 10-s crop; short views right-padded	One embedding or logit vector
FD-LLM	One centered 2,048-point window serialized as Stats15 text	Qwen2.5-7B LoRA generated label
FD-MVLLM	Three native-rate 2,048-point windows and CWT images	Mean logits
Omni zero-shot	Complete original view as audio	One generated label
Qwen2-Audio supervised	Centered crop of up to 10 s at 16 kHz	One generated label
RELATE-MLP	10-s FISHER query plus up to 16 training healthy references	One logit vector
RELATE-Gen	The same FISHER evidence plus three 2,048-point MiniROCKET windows	One generated record

similarity ranking. The retrieval removes all views from the query benchmark group before ranking, retains the highest-ranked view from each remaining distinct benchmark group, and averages up to 16 groups uniformly. Fewer references are used when the training pool contains fewer eligible groups.

E.3 4-TOKEN EVIDENCE BOTTLENECK

The FISHER adapter averages valid time positions within every ordered band and maps the resulting 1,536-dimensional vectors to width 64. Zero padding extends the list to 32 bands. Flattening, layer normalization, and a linear–GELU mapper produce T vectors of width 256. Learned token embeddings and a final layer normalization complete the FISHER branch.

The MiniROCKET branch fits 16 transforms on training data, indexed by the setting-local candidate space: eleven RMIS settings and five MIMII machine-local spaces. Each 9,996-dimensional transform vector is normalized, mapped to T bottleneck vectors of width 64, and projected to T vectors of width 256. The fusion module then applies the gated residual transformation defined in the main method. A final projection maps the fused tokens to the Qwen hidden size. Validation over $T \in \{1, 2, 4, 8\}$ selects $T = 4$ as the compact point; performance plateaus across 4–8 tokens in this single-seed sweep.

The expanded-prefix control retains a sample token, a global token, the same four fused summaries, and one descriptor plus 20 time tokens per valid FISHER band. Its length is $6 + 21B$: 90, 132, 258, or 342 tokens for $B = 4, 6, 12, 16$.

E.4 OPTIMIZATION AND INFERENCE

Qwen2.5-7B-Instruct is loaded in 4-bit NF4 with double quantization and bfloat16 computation. LoRA uses rank 8, alpha 16, zero dropout, and adapts `q_proj`, `k_proj`, `v_proj`, and `o_proj`. AdamW uses learning rates of 3×10^{-5} for the language-model parameters and 3×10^{-4} for the signal interface, a 5% linear warmup followed by cosine decay, and maximum gradient norm 1.0. Training uses batch size 1, gradient accumulation 8, and 100,424 view microsteps (12,553 optimizer updates). The selected model uses seed 111; the final branch and frozen-token comparisons use seeds 111, 222, and 333. Sampling follows

$$\text{setting} \rightarrow \text{balance stratum} \rightarrow \text{class} \rightarrow \text{benchmark group} \rightarrow \text{view}.$$

The final objective is autoregressive next-token prediction. The auxiliary-objective control adds a candidate-masked primary-condition loss with weight 0.3. Endpoint evaluation generates all 2,079 validation views and aggregates them to 1,218 groups. At inference, majority vote aggregates generated primary conditions within each group; ties follow a fixed candidate order.

The user prompt contains the task instruction, modality, machine/component description, and setting-local candidate labels. It does not contain dataset identifiers, signal UIDs, benchmark-group IDs, or gold fields or a sample-dependent field mask. Decoding is greedy (no sampling), emits at most 256 new tokens, and stops after a complete JSON object or EOS. The scorer joins gold data after generation. Missing or malformed records are incorrect for supported-record Joint EM; out-of-candidate primary labels are errors, and duplicate prediction rows are rejected. Paths outside the setting-level schema are tracked separately.

Prompt and training target example. The following bearing example is rendered from a training row with the candidate-only prompt formatter. Line breaks are added for display. The continuous signal prefix precedes the text; the target is supervised output.

> **System:** You are an industrial signal diagnostic assistant. Use the provided signal embedding and instruction. Return compact JSON only.
>
> **User:** Signal prefix grammar: Four aligned signal-evidence tokens summarize the observed waveform. Use them as diagnostic evidence and output strict JSON.
> Task: `fault_diagnosis`
> Instruction: Diagnose the signal and return the setting-specific natural-language fault class as JSON.
> Diagnostic context:
> `modality`: "vibration"
> `machine_or_component`: "bearing"
> `candidate_primary_labels`: ["healthy bearing", "bearing inner-race fault", "bearing outer-race fault", "bearing rolling-element fault"]
> Return compact JSON only.
>
> **Target:**

```
{"benchmark_primary_code": "inner_race_fault",
 "benchmark_primary_label": "bearing inner-race fault",
 "health_state": "anomalous",
 "semantic_factors": {"fault_location": "bearing inner race",
                      "fault_mechanism": "inner race defect"},
 "severity_fields": {"damage_size_mm": 0.2},
 "task": "fault_diagnosis"}
```

E.5 Frozen-token readout and discriminative classifiers

The direct interface route trains the same FISHER–MiniROCKET adapter and 4-token geometry without loading a language model. Its objective is the equal-weight sum of candidate-masked primary-condition cross-entropy and the mean cross-entropy over the source-supported fields present for each training example. The train-time decoder uses the same 7.74M-parameter readout architecture described below; together with the 6.32M-parameter interface, the route has 14.06M trainable parameters. It uses AdamW at 3×10^{-4}, the same sampler, warmup, cosine schedule, gradient clipping, and fixed 100,424-microstep endpoint as the matched autoregressive route. The two routes share the interface construction and use different initial interface tensors.

For each seed, the frozen-token experiment exports the exact four $3{,}584$-dimensional vectors delivered to Qwen after the trained signal adapter and prefix normalization. The signal adapter stays frozen, and a two-layer GELU MLP with a shared candidate-masked primary head and source-supported field heads trains only on the cache. It follows the RELATE-Gen hierarchical sampler for 100,424 view microsteps (12,553 optimizer updates). Within each autoregressively trained checkpoint, Qwen and its paired readout receive identical token values; only the autoregressive decoder or discriminative readout differs. The readout has 7.74M trainable parameters, compared with 11.37M trainable adapter and LoRA parameters for RELATE-Gen.

We additionally train end-to-end shared and setting-local discriminative classifiers from the same FISHER and MiniROCKET feature sources. These models learn their own signal adapters and

see every training view exactly three times, providing a comparison of complete model families and parameter-sharing choices. The shared model uses 6.35M trainable parameters and reaches $65.83 \pm 2.42\%$ accuracy; the collection of setting-local models uses 75.08M and reaches $67.69 \pm 0.24\%$ under the common scorers.

FISHER FT in the main comparison updates the full FISHER backbone and a shared output head; it differs from the frozen-FISHER setting-local MLP in Table 9. FISHER 1-NN time-pools each subband, pads to the shared band width, and L2-normalizes the result. Its training bank uses one fixed view per benchmark group; each candidate receives its maximum cosine similarity across matching training views. The fixed hard-vote ensemble combines FISHER 1-NN, MiniROCKET, and supervised Qwen2-Audio. Three-way disagreements use the training-group class prior among voted labels, then frozen candidate order.

E.6 Proprietary zero-shot audio APIs

The proprietary Qwen3.5-Omni-Plus, Gemini 3.8 Flash, and GPT Audio Mini baselines receive each frozen-test waveform and the setting-local candidate labels without task training or validation selection. Sound inputs use their native audio representation; vibration, current, and voltage signals use waveform-as-audio adaptation. Original WAV bytes are sent as-is; FLAC inputs are converted in memory to PCM16 WAV without resampling, so the evaluated source rates remain 16–64 kHz. The prompt states the modality, machine/component, and candidate labels and requests exactly one candidate. Predictions are aggregated over views with the same benchmark-group majority rule and candidate-order tie break used by the common scorer. The exact routes are `qwen3.5-omni-plus` through Alibaba Cloud and `google/gemini-3.8-flash` and `openai/gpt-audio-mini` through OpenRouter. All use temperature zero; Gemini additionally uses medium reasoning effort and a 2,048-token output cap. The OpenRouter evaluations ran on September 18–19, 2026. Retryable service failures were retried before scoring. Remaining invalid or out-of-candidate predictions enter the scorer as errors; labels and other models are not used to fill missing predictions.

Table 17: Complete frozen-test results for proprietary zero-shot audio APIs. All metrics are percentages over the same 4,430 views and 3,598 benchmark groups.

Model	Set.-macro Acc.	Set.-macro F1	Source-macro Acc.	Group-micro Acc.
Qwen3.5-Omni-Plus	**26.77**	14.79	**28.15**	**24.26**
Gemini 3.8 Flash	25.14	**19.21**	26.69	22.35
GPT Audio Mini	26.21	15.71	27.03	23.10

All three systems are evaluated on the same complete test release. Qwen3.5-Omni-Plus gives the highest setting-macro accuracy, and Gemini 3.8 Flash gives the highest class-macro F1. Under this fixed candidate-conditioned prompting setup, setting-macro accuracy ranges from 25.14% to 26.77%, below every trained method in Table 1.

E.7 Cached-evidence deployment measurement

We compare the seed-111 generator and paired frozen-token readout on 92 validation views selected from metadata only, with one view per group and at most eight groups per setting. Each system and batch size runs in a separate process on one NVIDIA A10, with three warmup batches and three measured passes. Timing includes cached-feature access, the live signal adapter, prefix normalization, and decoding; it excludes feature extraction, healthy-reference retrieval, and model loading.

E.8 FD-MVLLM adaptation

The FD-MVLLM baseline (Li et al., 2025a) adapts the released implementation, retaining its waveform-patch, CWT-image, word-prototype reprogramming, frozen GPT-2, and classification components. Each input combines a 2,048-sample waveform divided into non-overlapping 64-sample patches with a $4 \times 128 \times 128$ RGBA CWT image. The prompt exposes Stats15 features and setting-local candidate labels. Validation setting-macro accuracy selects epoch 25; test inference uses that fixed checkpoint.

Table 18: Deployment from cached FISHER and MiniROCKET features on 92 views. Median and P95 are batch latencies in ms; throughput is views per second over complete passes; memory is peak allocated GPU memory in GiB.

Decoder	Batch	Median	P95	Views/s	GiB
RELATE-Gen	1	4789.82	6730.92	0.20	5.428
Frozen-token readout	1	23.61	32.75	45.95	0.078
RELATE-Gen	8	7581.58	9615.23	0.98	5.637
Frozen-token readout	8	160.53	196.74	52.56	0.095

F SIGNAL-INTERFACE COMPARISONS

RELATE-MLP and conventional decoders applied to the complete relative-FISHER representation provide a full-evidence validation comparison. On all 1,218 validation groups, candidate-local cosine 1-NN reaches 72.21% setting-macro accuracy and RELATE-MLP reaches 78.38/76.03% accuracy/F1. Expanding its predictions with the complete train-only taxonomy and severity template gives 72.74% Field EM and 64.84% Joint EM. A jointly trained structured MLP on the same full representation reaches 70.21/68.02% accuracy/F1, 79.78% Field EM, and 59.56% Joint EM over three seeds. A primary-only control matched to this structured MLP except for field supervision reaches 69.78/67.34% accuracy/F1. Both controls use view-level 100,424-step training, whereas the main RELATE-MLP uses a group-level 80-epoch recipe. Table 3 places these full-evidence routes alongside the compact interfaces. Both full-representation MLPs consume the 24,576-dimensional $[q, r, q - r, |q - r|]$ representation with up to 16 healthy reference groups. The frozen readout uses 4 tokens and a different training recipe.

Input composition across the two routes. We compare the reference pair and its derived difference blocks separately within each route. Table 19 uses a separately trained primary-MLP control with its own checkpoints. Its three arms share one saved initialization, the main MLP architecture and optimization recipe, sampler, and scorer; they differ only in which input blocks are zero-masked. Adding the healthy prototype raises accuracy/F1 by 2.55/3.18 points over the query alone. Appending the explicit signed and magnitude residual blocks changes the result by $-1.07/-1.71$ points relative to $[q, r]$. The complete four-block arm reaches 77.31% accuracy in this control; the separately trained main RELATE-MLP reaches 78.38% on validation and retains all four blocks on test.

Table 19: Separately trained fixed-configuration validation control for relative-FISHER input blocks on all 1,218 benchmark groups. Omitted blocks are zero-masked; all arms share one saved initialization.

Visible input blocks	Acc.	F1
q	75.83	73.61
$[q, r]$	**78.38**	**76.79**
$[q, r, q - r, \|q - r\|]$	77.31	75.08

Input blocks in the Qwen route. The matched Qwen control replaces the difference blocks with zeros, $[q, r, 0, 0]$, while retaining the healthy references, input width, temporal branch, adapter, LoRA recipe, sampler, and 100,424 training microsteps. Both checkpoints are evaluated on the same 2,079 validation views and 1,218 groups. The complete input improves accuracy and F1 by 2.43 and 2.04 points, respectively, in this single-run comparison. The MLP control favors the pair, so the two routes show different responses to the explicit difference blocks.

The frozen-probe columns use FISHER-small, FISHER-base, BEATs (Chen et al., 2022), OpenBEATs-base and OpenBEATs-large (Bharadwaj et al., 2025), EAT-base, and EAT-STFT-base (Chen et al., 2024) on identical 16 kHz waveforms. Short views are right-zero-padded to ten seconds at batch formation. The linear probe standardizes group embeddings and fits setting-local logistic classifiers with $C = 1$; cosine 1-NN is secondary. The generated comparison uses either relative encoder evidence alone or the same evidence fused with the train-fitted MiniROCKET branch. The

Table 20: Matched single-run Qwen input-block control at the fixed training endpoint. All metrics are setting-macro percentages on validation.

Visible input blocks	Acc.	F1	Joint EM
$[q, r, 0, 0]$	73.32	71.90	66.32
$[q, r, q-r, \lvert q-r \rvert]$	75.75	73.94	66.90

FISHER-small generated entries reuse the main pipeline's source-rate cache, whereas the six replacement encoders use the 16 kHz package. Every generated row uses 4 signal tokens, the same healthy-reference membership, Qwen model, optimizer, training duration, seed, generation procedure, and scorer. Each generated row is a complete encoder–preprocessing interface; the probe columns compare frozen representations under a common 16 kHz linear-probing protocol.

Table 21: Frozen probing and generated validation on the same 1,218 groups. Entries report setting-macro accuracy and setting-macro F1 (%); generated rows use AR loss plus auxiliary primary-label cross-entropy.

Encoder	Params (M)	Probe Acc.	Probe F1	Encoder evidence Acc.	Encoder evidence F1	Fused evidence Acc./F1
FISHER-small	23.9	67.86	64.37	**70.42**	**68.31**	**71.79/70.45**
FISHER-base	323.8	77.09	73.65	60.69	59.87	63.91/62.98
BEATs	90.3	76.08	72.97	62.93	58.34	61.40/59.51
OpenBEATs-base	90.3	74.18	71.48	60.60	59.36	59.99/57.76
OpenBEATs-large	311.4	**77.58**	**75.14**	59.70	57.57	61.29/58.70
EAT-base	88.2	75.00	72.97	66.01	64.30	69.28/68.10
EAT-STFT-base	323.8	67.78	65.00	43.55	37.81	48.00/42.12

Table 22: Single-seed controls on the complete 1,218-group MOSAIC validation split. Values are setting-macro accuracy, setting-macro F1, and source-family macro accuracy (%). All rows use the dual-branch interface.

Configuration	Acc.	F1	Src.
Dual-branch fusion + auxiliary classification	71.79	70.45	77.83
Without healthy-reference evidence + auxiliary classification	71.20	70.21	77.68
Dual-branch fusion + autoregressive objective	75.75	73.94	80.95
Without healthy-reference evidence + autoregressive objective	67.91	66.53	75.74

G Structured-Field Evaluation

Metrics and aggregation. We score seven canonical paths: `health_state`, `semantic_factors.fault_location`, `semantic_factors.fault_mechanism`, `semantic_factors.health_state`, `semantic_factors.machine_type`, `severity_fields.damage_size_mm`, and `severity_fields.eccentricity_mm`. Each path is scored only on benchmark groups whose source schema supports it. Field EM computes group exact match for each path and weights the seven path scores equally. Joint EM is one only when the group primary condition and every source-supported field are exact; setting-macro Joint EM first averages this indicator within each setting and then weights the 12 settings equally. Severity EM weights damage-size and eccentricity exact match equally over their respective supported groups.

Generated and other record-valued predictions are aggregated by majority vote over views, with frozen candidate order breaking primary-label ties. Supported fields are voted independently: missing participates in the vote and wins only when it has more votes than every non-missing value, while tied non-missing values follow canonical JSON lexical order. A field is emitted only when its winning value is non-missing, so the aggregated record can combine values from different views. Predictions outside a group's supported field set do not change Field or Joint EM; paths outside the setting schema are counted in the separate extra-field audit. On the frozen test set, none of the 11,171 emitted group-level field paths fell outside their corresponding setting-level schemas.

This count covers the seven evaluated paths and other paths under `semantic_factors` and `severity_fields`; extra top-level JSON keys are checked for schema validity separately. Validation whole-tuple voting gives the same Joint EM as fieldwise voting for all three RELATE-Gen seeds; seed 111 has three groups whose fieldwise tuple was not emitted by any single view, without a Joint-EM outcome change.

The field paths play different roles. Machine type is provided in the prompt; health state, location, and mechanism are largely determined by the primary taxonomy; damage size and eccentricity can vary within one primary condition. The top-level `health_state` applies across MOSAIC, whereas `semantic_factors.health_state` is the MIMII machine-level attribute. Field EM summarizes these taxonomy and within-condition attributes together.

Training-derived record construction. For the two token readouts, a non-severity field is mapped for a diagnostic setting and primary condition only when it is present in every corresponding training group and has one unique canonical value. The 15,477 training groups yield 204 entries over 68 setting–condition pairs. At inference, the route's predicted primary condition indexes this mapping; a missing setting–primary mapping is an error. Fields without a unique training value keep their route predictions.

The complete RELATE-MLP template instead assigns each supported categorical field its modal value among training groups with the same setting and primary condition, breaking ties by canonical lexical order. It also assigns the modal training value to each supported severity field. At inference, the predicted primary condition indexes this complete record. On the 60 validation groups supporting damage size and the 56 supporting eccentricity, it reaches 33.33% and 19.64% exact match, respectively, or 26.49% when the two field scores are weighted equally. Adding the deterministic taxonomy fields gives 72.74% Field EM and 64.84% Joint EM.

The more restrictive key-balanced analysis includes only $(\text{setting}, \text{primary condition}, \text{field})$ keys with positive training entropy. Nine keys across `sdust_bearing`, `umged_cur`, `umged_vib`, and `umged_vol` cover 88 validation groups. It first averages groups within a key, then keys within a field, and finally the two fields equally. The primary-conditioned mode template reaches 18.75% under this metric. This 88-group analysis and the 60/56-group supported-field analysis above use different denominators.

Frozen-classifier severity heads. We freeze all 6,302,506 parameters of the validation-selected RELATE-MLP, including its 256-dimensional hidden representation, and train linear heads for damage size and eccentricity. Their train-only vocabularies contain 3 and 11 values, and the two heads add 3,598 trainable parameters. Each head uses seed 111 and AdamW for 80 epochs with learning rate 10^{-3}, weight decay 10^{-4}, and batch size 256. Primary predictions remain unchanged on all 1,218 validation groups. On the same 60/56 supported groups, the heads raise Severity EM from 26.49% to 41.73%, Field EM from 72.74% to 77.10%, and Joint EM from 64.84% to 65.33%.

Matched autoregressive and direct interfaces. The AR and Direct controls train the same 4-token interface with autoregressive structured targets or direct field supervision, holding features, sampler, schedule, and 100,424 microsteps fixed. The same 7.74M-parameter readout then decodes both frozen interfaces. Across three validation runs, primary accuracy is 74.46% for AR and 74.48% for Direct. Raw Joint EM is 64.72% versus 61.41%; 3.19 of the 3.31-point difference comes from taxonomy fields. Applying the same train-only taxonomy rule to both leaves primary and severity predictions unchanged and gives 65.65% versus 65.52% Joint EM. The paired interval for the remaining 0.13-point difference is $[-2.08, 2.35]$ (Table 13). On the 88 positive-entropy groups, the key-balanced severity score is 49.43% for AR and 48.96% for Direct, compared with 18.75% for the mode template. Both interfaces score above the primary-conditioned mode template on this 88-group analysis.

For the readouts, the group primary is selected from mean candidate logits and fields are aggregated by the voting rule above. The taxonomy mapping then overwrites only uniquely determined non-severity fields. Before aggregation, the AR readout restores numeric damage-size and eccentricity values from their serialized strings with the unambiguous training vocabularies; this preserves the field values while putting both readouts in the same training-declared scalar types.

Frozen-test within-condition fields. The five-route test comparison appears in Table 2. Table 23 separates the two signal-dependent fields on their supported test groups. Results use the fixed checkpoints reported in the main comparison.

Table 23: Frozen-test exact match (%) for the two within-condition fields. Damage size is scored on 300 supported groups and eccentricity on 76. Their two scores are weighted equally for Severity EM in Table 2.

Route	Damage size	Eccentricity
RELATE-Gen	77.33	27.63
AR-token readout + taxonomy rules	70.67	32.89
Direct-token readout + taxonomy rules	71.00	25.00
RELATE-MLP + complete template	28.67	15.79
RELATE-MLP + severity heads	56.00	23.68

Table 24 compares validation exact match for each field between generation and a readout on the same frozen RELATE-Gen tokens, averaged over three runs.

Table 24: Three-seed validation mean exact match (%) for each supported field using identical frozen RELATE-Gen tokens. The Groups column gives the denominator per seed.

Field	Groups	RELATE-Gen	Frozen-token readout
Record health state (`health_state`)	1,218	85.85	**86.78**
Fault location	618	91.26	**91.53**
Fault mechanism	618	90.51	**91.21**
MIMII semantic health state (`semantic_factors.health_state`)	600	75.50	**77.17**
Machine type	600	99.94	**100.00**
Damage size	60	75.56	**76.11**
Eccentricity	56	28.57	**31.55**
Field macro EM	—	78.17	**79.19**
Setting-macro Joint EM	—	64.64	**64.72**

H Compute Resources

Table 25: Observed per-run compute for trainable MOSAIC methods. GPU rows use one GPU; generated validation is included unless listed separately. Times reflect each method's hardware and preprocessing pipeline. The MLP route bundle includes feature-cache construction, cosine 1-NN, MLP training, and validation scoring.

Method	Training resource	Training time	Separate generation
RELATE-Gen	NVIDIA A10, 6 CPUs	23.45 h	NVIDIA A10, 1.11 h
Direct interface training	NVIDIA A10, 6 CPUs	3.10 h	included
Paired direct-interface readout	NVIDIA A10, 6 CPUs	0.23 h	included
Autoregressive frozen-token readout	NVIDIA A10, 6 CPUs	0.26 h	included
RELATE-MLP route bundle	RTX 4090, 8 CPUs	10.65 min	included
RELATE-MLP severity heads	NVIDIA A10, 6 CPUs	29 s	included
Dual-branch auxiliary-objective control	NVIDIA A10, 6 CPUs	24.50 h	included
Supervised Qwen2-Audio	RTX 4090, 8 CPUs	27.63 h	NVIDIA A10, 3.31 h
FD-LLM-Stats	RTX 4090, 18 CPUs	14.03 h	RTX 4090, 0.08 h
TEST-GPT2 adaptation	NVIDIA A10, 6 CPUs	0.73 h	included
FD-MVLLM	NVIDIA A10, 8 CPUs	1.78 h	included
FISHER FT	NVIDIA A10, 6 CPUs	1.00 h	included
MOMENT-1-base	RTX 4090, 18 CPUs	0.19 h	included
WDCNN	RTX 4090, 6 CPUs	0.03 h	included
MiniROCKET	8 CPUs	0.05 h	included
Stats15–SVM	16 CPUs	13 s	included
FISHER 1-NN	8 CPUs	29 s	included