applied sciences

MDPI

Article

Acoustic-Array Diagnosis of Coupled Planetary Gearbox Faults Under Controlled 3D Spatial Interference with Physically Motivated Structural Priors

Xingtong Liu [1], Yuqing Wang [2] and Mei Wang [1,*]

[1] School of Mechanical Engineering, Sichuan University, Chengdu 610065, China; 2024223025052@stu.scu.edu.cn
[2] Independent Researcher, Singapore 639798, Singapore
* Correspondence: wangmei@scu.edu.cn

Abstract

Non-contact acoustic arrays offer a complementary route for planetary-gearbox monitoring, but multiple sound sources introduce spatial ambiguity and masking. We propose a multi-modal mixture-of-experts framework (MM-MoE) with physically motivated structural priors for 16-channel acoustic signals. Four learned prototype waveforms replace per-channel wavelet transforms, and an adapted Vision Mamba encoder produces channel-aligned spectral embeddings. A waveform branch models temporal patterns, a coupling prior learned from inter-channel spectral similarity biases cross-modal attention, and task-aware expert routing supports simultaneous scene parsing and fault diagnosis. Test-bench gearbox recordings were replayed through two loudspeakers at controlled positions, heights and sound-pressure levels in a semi-anechoic chamber. Over five training runs, MM-MoE achieves window-level accuracies of 89.4% for 16-class spatial-energy scene parsing and 76.5% for 16-class coupled fault diagnosis, with 69.2% joint exact match (JEM), 13.4 percentage points above the strongest evaluated baseline (95% confidence interval: 12.4–14.4). Aggregating the windows of each recording raises JEM to 87.1%. When source regions are excluded from training, ordered fault accuracy on them is 68.3%, compared with 75.5% on seen regions. The findings are limited to controlled loudspeaker replay; validation on operating machinery remains necessary.

Keywords: faultdetection; deep learning; time-series classification; multi-modal learning; mixture-of-experts; planetary gearbox

Check for updates

Academic Editor: Bernardo Sotto-Maior Peralva

Received: 16 September 2026
Revised: 22 October 2026
Accepted: 23 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Reliable condition monitoring is essential for intelligent manufacturing and industrial safety [1,2]. Planetary gearboxes are widely used in wind turbines, rotorcraft drive systems, and other heavy-duty transmission platforms because of their high power density and load-bearing capacity [2]. Under heavy loads and variable speeds, however, their compact gear trains are vulnerable to pitting, tooth breakage, and related compound failures. These failures can cause costly downtime and safety risks, making robust gearbox diagnosis a practical requirement [2].

Most vibration-based monitoring systems rely on contact sensors, which provide limited spatial information and can be difficult to deploy on inaccessible equipment. Airborne-acoustic microphone arrays offer a complementary sensing route by recording machine sound across spatially separated channels. This spatial awareness is valuable in factory environments where multiple machines operate simultaneously [3]. Yet realistic layouts often

introduce vertical height differences, complex 3D interference, and strong acoustic masking. Height differences alter propagation delays and inter-channel phase relationships [4], while competing sources can mask fault signatures [3].

For an acoustic array, the design challenge is to combine waveform and time–frequency (TF) information while retaining channel alignment and controlling the cost of TF encoding. Cross-attention can learn input-dependent channel interactions; an explicit coupling bias is an additional architectural option (Section 2.1). Vision Mamba provides efficient long-range modeling [5], and anisotropic patches with dual-axis traversal adapt it to acoustic TF maps. Because the two required outputs share parameters, they may also compete during training [6], which motivates task-dependent expert routing.

To investigate this design, this paper proposes a Multi-Modal Mixture-of-Experts (MM-MoE) framework with physically motivated structural priors for coupled fault diagnosis in planetary gearboxes using 16-channel acoustic signals. The study is evaluated on a controlled replay-based 3D spatial-interference dataset, where real planetary-gearbox fault recordings are first collected from a test bench and then replayed by loudspeakers at controlled positions, heights, and sound-pressure levels. The resulting task requires simultaneous spatial-energy scene parsing and coupled fault diagnosis under strong masking. This protocol isolates the spatial-energy factors of interest; it is not a validation on operating machinery.

The proposed framework combines three structural ideas. First, a topology-guided low-rank continuous wavelet transform (TG-LR-CWT) maps multichannel acoustic windows into a small set of prototype time–frequency representations, and a Spectral–Temporal Vision Mamba (ST-Vim) encoder preserves impulse and harmonic geometry. Second, a Condition-Guided Temporal Clustering Module (CG-TCM) learns latent signal-adaptive routing from raw waveforms, and a Spectral-Coupling Channel Clustering Module (SC-CCM) estimates a soft channel-coupling prior without explicit operating-condition supervision. Third, this learned prior biases Topology-Masked Cross-Attention (TMCA), and a Task-Aware Mixture-of-Experts (TA-MoE) decoder combines always-active shared experts with task-specific routing to produce the two outputs.

These choices are motivated by the microphone-array setting and the acoustic structure of gearbox signals, but the framework does not impose a physical propagation model. Section 3.1 separates the fixed, acoustically motivated design choices from the components learned from data.

2. Related Work

2.1. Acoustic-Array Diagnosis and Multimodal Fusion

Deep learning methods extract diagnostic features from raw vibration signals [7] and time–frequency representations [8,9], although gearbox signals vary with speed and load [10]. Microphone arrays add spatial information. Beamforming uses inter-channel relationships to localize or focus on sources [11], and co-prime circular arrays have been applied to bearing diagnosis [4]. A recent preprint by Li et al. steers multiple beams toward gear and bearing regions and fuses the extracted signals for health-state classification [12]. These approaches exploit spatial selectivity explicitly, whereas MM-MoE learns channel relationships without a beamforming stage or microphone coordinates.

The closest precedent is Deng's gearbox diagnosis model [13], whose published abstract describes beamforming-based noise reduction followed by cross-attention between time-domain vibration features (queries) and frequency-domain acoustic features (keys and values). Array-assisted cross-attention is therefore an established direction. MM-MoE differs in its inputs and outputs: it uses waveform and TF views of the same acoustic recording and jointly predicts spatial-energy scene and fault labels. Our standalone plain

cross-attention (Standalone Plain CA) baseline and internal V5 variant compare fusion configurations on the present benchmark and are not reproductions of Deng's pipeline.

Noise-aware acoustic diagnosis has also been studied beyond supervised classification. MIMII provides eight-channel machine recordings mixed with factory background noise [14]; Lee et al. combine variational-autoencoder anomaly detection with domain adaptation for environmental noise [15]; and MC-ABDS addresses overlapping interference in diagnosis at a low signal-to-noise ratio (SNR) [3]. Their tasks and noise protocols differ from the controlled two-source scene-and-fault classification considered here.

Fusion mechanisms. Waveform and TF views can be combined within one sensing modality [16]. Other methods fuse acoustic and vibration signals with vision transformers and domain adaptation [17], acoustic and infrared measurements at the data level [18], or vibration and thermal information across operating conditions [19]. Qiu et al. denoise vibration signals and fuse waveform and Gramian-angular-field features [20]. Attention bottlenecks restrict multimodal exchange to a few latent tokens [21], and multi-view graph networks combine time- and frequency-domain sensor information [22]. Neither adaptive fusion nor explicit sensor relationships are therefore new in themselves.

Structural attention and positioning. Graphormer adds graph-derived structural terms to attention logits [23], and DUET learns frequency-domain channel relationships for multivariate forecasting [24]. Conventional cross-attention is already input-dependent and, with channel encodings, can learn channel relationships. The question addressed here is practical: how to retain channel-aligned interaction at a reduced TF-encoding cost, add an explicit spectral-coupling preference, and serve two outputs with different needs. MM-MoE computes continuous wavelet transforms (CWTs) of only four prototype waveforms, lifts their features to 16 channel-aligned tokens, adds a learned per-window coupling bias to cross-attention, and routes the fused features to task-specific experts. Section 4.5 evaluates the corresponding configurations.

2.2. Multi-Task Learning and Mixture-of-Experts

Multi-task learning (MTL) for diagnosis shares information across related outputs [3,25]. Shared parameters may face competing objectives [6], but negative transfer must be established against suitable single-task controls. MMoE uses task-specific gates over shared experts [26], and sparse expert architectures activate selected experts to increase capacity at limited computation [27]. PLE organizes shared and task-specific experts with progressive routing [28]; DeepSeekMoE combines always-active shared experts with sparsely routed experts [29]; and Switch Transformers uses an auxiliary objective to balance expert allocation [30]. In machinery diagnosis, M4oE builds a diagnosis task for each sensor stream, combines shared and task-specific features, and fuses decisions to tolerate sensor unavailability [31]. TA-MoE adapts these sharing and routing ideas with separate routers for scene parsing and fault diagnosis; it does not claim to introduce shared experts or task-specific gating.

2.3. State Space Models in Vision and Time-Series

State space models have emerged as efficient alternatives to attention for modeling long-range dependencies. Architectures such as structured state space sequence models and Mamba build on continuous-time system representations and achieve linear complexity with respect to sequence length [32,33]. This efficiency has encouraged their adaptation from sequence modeling to visual representation learning. Vision Mamba and VMamba process two-dimensional images using sequence-based scanning strategies [5,34].

In machinery TF representations, the time and frequency axes have different meanings: sustained harmonics and short transients call for different aggregation patterns [9].

Anisotropic patches and dual-axis traversal adapt state-space encoders to this structure; their analytical scales are discussed in Appendix F.

2.4. Physics-Informed Learning and Physically Motivated Architecture

Physics-informed learning can incorporate physical knowledge through data, model structure or training objectives [35], and its applications to anomaly detection and condition monitoring span several forms of such integration [36]. Enforcing a governing equation or a measured physical relation is a stronger claim than choosing an architecture motivated by signal characteristics. We use *physically motivated structural priors* in the latter sense: phase-related inputs, patch geometry and waveform processing have acoustic rationales, while the coupling matrix and routers are learned from data. No acoustic propagation equation constrains the loss or outputs, and the learned channel graph is not a measured propagation topology (Section 3.1 and Table A23).

3. Materials and Methods

3.1. Method Overview

The proposed MM-MoE framework jointly uses prototype-level time–frequency representations and raw multi-channel acoustic waveforms to solve two supervised tasks: 16-class spatial-energy scene parsing and 16-class coupled fault diagnosis. Each input sample is a segmented 16-channel acoustic window $X \in \mathbb{R}^{N \times L}$, where $N = 16$ is the number of microphone channels, and $L = 4096$ is the window length. The model uses only the two task labels during training. No source-height, sound-pressure-level, or operating-condition label is used to supervise the latent routing modules.

As shown in Figure 1, the framework has three stages. First, the image branch maps X into $K = 4$ topology-guided acoustic prototypes, applies complex CWT to obtain phase-aware time–frequency maps, and encodes them with a shared Spectral–Temporal Vision Mamba (ST-Vim) encoder. The resulting prototype features are lifted back to channel-aligned spectral embeddings, so the model keeps channel-level alignment without computing a full CWT for every microphone. Second, the waveform branch processes the original 16-channel signals through latent temporal experts and learns a soft channel-coupling prior from frequency-domain channel relationships. Third, the coupling prior biases cross-modal attention. A task-aware mixture-of-experts decoder then routes the fused representation to scene-specific and diagnosis-specific experts. Table 1 summarizes the role of each module, and Table 2 gives the dimensions of the main representations. Further implementation details are provided in Appendices A.7, B and C.

Table 1. Module roles, expected contributions, and related comparisons. Expected contributions are design motivations; the related comparisons are discussed in Section 4.5.

Module	Purpose	Expected Contribution	Related Comparisons
TG-LR-CWT	Project 16 channels into $K = 4$ learned waveforms, compute phase-aware CWT maps, and lift prototype features to 16 channel-aligned embeddings.	Reduce CWT operations to 4/16 = 25% while retaining spectral and channel-dependent information (total cost in Section 4.2).	Section 4.5.2; Tables A15 and A16.

Table 1. *Cont.*

Module	Purpose	Expected Contribution	Related Comparisons
ST-Vim	Encode prototype TF maps with anisotropic overlapping patches and frequency-major/time-major scans.	Reduce truncation of narrow transients and capture complementary spectral and temporal dependencies.	Section 4.5.3 (encoder).
CG-TCM	Encode each raw channel with $K_t = 4$ temporal experts weighted by a signal-adaptive router.	Adapt waveform processing to latent signal regimes and supply raw temporal information.	Section 4.5.3 (temporal routing).
SC-CCM	Learn a soft coupling matrix M from channel amplitude spectra using a quadratic spectral dissimilarity.	Supply a task-adaptive channel-relation prior to channel lifting and attention.	Table A15 (C3 vs. C4); Section 4.5.3 (attention).
TMCA	Fuse spectral queries with temporal keys/values, using $\log(M + \epsilon)$ as an attention bias.	Improve cross-modal alignment by down-weighting weakly coupled channel pairs.	Table A10 (V5 vs. V6).
TA-MoE	Decode fused features with always-active shared experts and two task routers, each selecting 3 of 8 routed experts.	Retain common acoustic features while allowing task-specific specialization to reduce interference between the two tasks.	Table A10 (V6 vs. V7); Section 4.4; Section 4.5.3 (decoder).

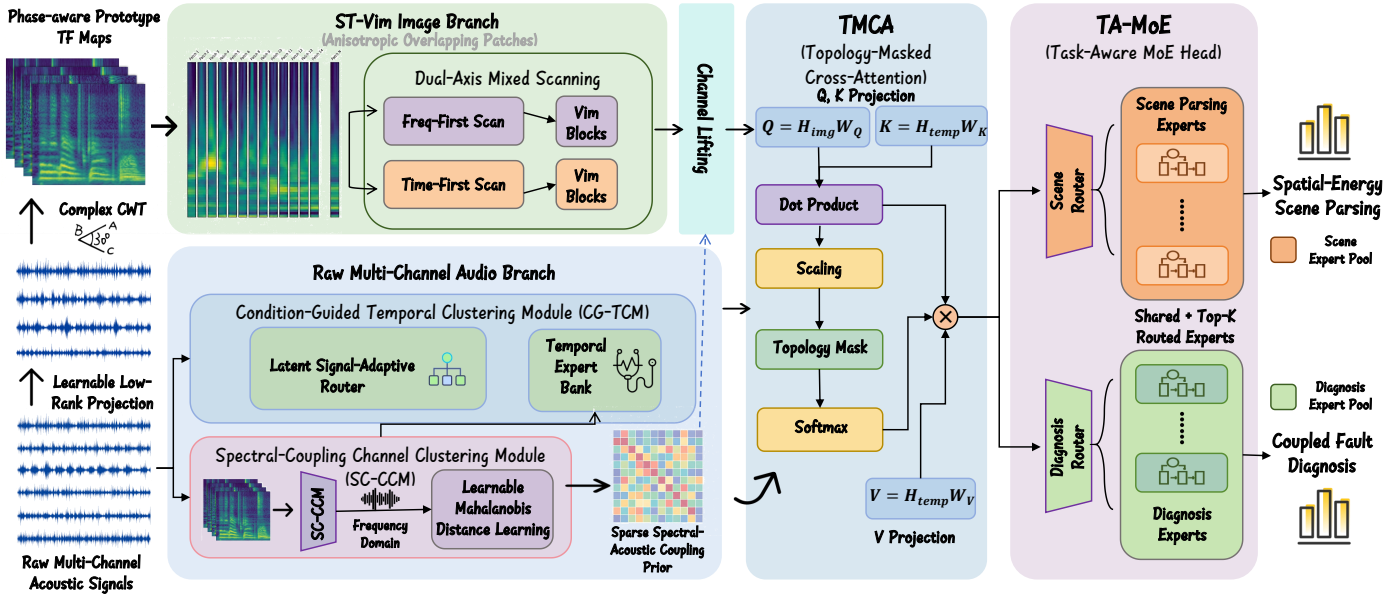


Figure 1. Architecture of MM-MoE. The image branch encodes phase-aware prototype time–frequency maps with ST-Vim and lifts the features to the channel axis. The raw-waveform branch combines latent signal-adaptive temporal routing with spectral channel-coupling estimation. The coupling prior supplies soft biases to channel lifting and TMCA; it does not impose hard edge pruning or measured array adjacency. The scene and diagnosis routers use a shared expert bank with task-dependent routing weights; the colored expert groups indicate task-dependent contributions. Table 2 gives the representation dimensions.

Table 2. Concrete tensor shapes for the implementation specification. Shapes omit the batch axis; $N = 16$, $L = 4096$, $K = K_t = 4$, $d = 256$, $r = 64$, and $F = 2049$.

Symbol	Representation/Operation	Shape	Fixed or Learned Dependence
X	Raw channel waveforms	16×4096	Observed input; fixed channel indexing.
P, $U = PX$	Projection and prototype waveforms	4×16; 4×4096	P learned; U computed.
I_k	Phase-aware TF map of prototype k	$224 \times 224 \times 3$	CWT, log/phase construction and resizing fixed.
T_{in}, $\mathcal{T}_{in}$	Patch sequence and grid	1443×256; $13 \times 111 \times 256$	32×4 patches; 16×2 strides; learned projection and positions.
Z_{tf}	Pooled prototype features	4×256	Shared ST-Vim encoder; mean over 1443 patch tokens.
A, ηMP^T	Lifting weights and coupling bias	16×4 each	Learned embeddings, W_a, P, M and scalar η.
H_{img}	Channel-aligned spectral features	16×256	$AZ_{tf} + E_{ch}$; E_{ch} learned.
X_n^{freq}	One channel's amplitude spectrum	2049	Full one-sided rFFT magnitude; DC and Nyquist included.
$G(x_n)$	Temporal routing weights	4	Dense softmax over four temporal experts.
H_{temp}	Temporal features	16×256	Learned channelwise temporal experts and router.
A_q, R	Spectral factor and quadratic form	64×2049; 2049×2049	A_q learned; $R = A_q^T A_q$ computed.
D, M	Dissimilarity and coupling	16×16 each	Computed symmetric arrays; M dense and soft.
Q_{att}, K_{att}, V	Query, key and value	16×256 each	Three learned $256 \to 256$ projections; one attention head.
S, $\log(M + \epsilon)$, A_{masked}	Scores, bias and attention	16×16 each	Row softmax over key channels; $\epsilon = 10^{-6}$.
Z	Residual fused features	16×256	$A_{masked}V + H_{img}$.
H_{scene}, $H_{diagnosis}$	Task expert outputs	16×256 each	Two shared and eight routed experts; Top-3 per task/window.
h_{scene}, $h_{diagnosis}$	Channel-mean task vectors	256 each	Mean of each H_t over its 16 channel rows.
$\hat{Y}_{scene}$, $\hat{Y}_{diagnosis}$	Class probabilities	16 each	Separate learned $256 \to 16$ heads and class softmax.

The two scan orders are inverse-permuted to the same 13×111 grid before feature concatenation and $512 \to 256$ fusion. Fixed channel indexing does not prescribe a microphone-geometry adjacency. Routing probabilities and selected weights are distinct; Appendix C.4 fixes the normalization.

The modules are coupled through their data flow. In particular, the matrix M produced by SC-CCM both refines the lifting weights (Section 3.2) and biases the TMCA scores (Section 3.4). The V5–V6 step of the ablation therefore evaluates SC-CCM and TMCA together, and the gains of related comparisons should not be added as independent module effects.

Physically motivated and learned components. The framework imposes no governing acoustic equation or physical constraint on its loss or outputs, and measured transfer functions, propagation delays and microphone coordinates are not model inputs. Its physically motivated elements are four fixed design choices:

- A complex CWT with log-magnitude and sine/cosine phase channels, which retains energy and phase information;
- Anisotropic overlapping patches with frequency-major and time-major scan orders, which represent narrow transients and harmonic continuity;
- A raw-waveform path, which retains inter-channel timing;
- Channel amplitude spectra, obtained by the real-input fast Fourier transform (rFFT), as the inputs for estimating channel relationships.

All other components are learned from data. These include the prototype projection P, the channel and prototype embeddings, the lifting projection W_a, the spectral metric factor A_q and the threshold γ. They also include the patch projections, state-space weights, attention projections, experts and routers. The coupling matrix M and the lifting weights A are computed for each input window from these parameters (Sections 3.2 and 3.3). Training uses the two task losses and an auxiliary expert-balancing term (Section 3.4), without a physical-law residual or propagation supervision. Accordingly, "topology" in TG-LR-CWT and TMCA denotes the learned channel-coupling graph, not the array geometry. Appendix F (Table A23) gives the acoustic rationale and calculated scales of the fixed choices.

3.2. Spectral–Temporal Representation Learning

The image branch is designed to preserve time–frequency geometry without computing CWT for every microphone channel. Given $X \in \mathbb{R}^{N\times L}$, TG-LR-CWT first projects the 16 channels into K acoustic prototypes:

$$U = PX, \quad P \in \mathbb{R}^{K\times N}, \quad U \in \mathbb{R}^{K\times L}, \tag{1}$$

where P is a learnable low-rank projection matrix, $K = 4$ is the default number of prototypes (its accuracy–cost rationale is discussed in Section 4.5.2), and $U = \{u_1, \ldots, u_K\}$ contains the projected prototype waveforms. Each $u_k \in \mathbb{R}^L$ represents a learned spatial response pattern formed from the original microphone channels.

For each prototype u_k, we compute a complex CWT and construct a phase-aware time–frequency map:

$$I_k = [\log|C_k|, \cos(\angle C_k), \sin(\angle C_k)], \quad C_k = \mathrm{CWT}(u_k), \tag{2}$$

where C_k denotes the complex wavelet coefficients of the k-th prototype, and $I_k \in \mathbb{R}^{H\times W\times 3}$ is the resulting TF map with frequency resolution H, temporal resolution W, and three channels. The log-magnitude channel captures spectral energy, while the sine and cosine phase channels retain phase-related structure that is useful for spatial-energy parsing. The K maps are encoded by a shared ST-Vim encoder:

$$z_k = E_{TF}(I_k), \quad Z_{tf} = [z_1, \ldots, z_K] \in \mathbb{R}^{K\times d}. \tag{3}$$

Here, $E_{TF}(\cdot)$ is the shared image encoder, $z_k \in \mathbb{R}^d$ is the encoded feature of prototype k, and d is the embedding dimension.

ST-Vim adapts Vision Mamba to acoustic time–frequency images [5,32,33]. It uses anisotropic overlapping patches to reduce truncation of narrow transient impulses and dual-axis scanning to separately model harmonic continuity and temporal periodicity. The patch tokens form a grid $\mathcal{T}_{in} \in \mathbb{R}^{F'\times T'\times d}$. Frequency-first serialization visits all F' frequency patches at one time position before advancing in time; time-first serialization visits all T' time patches in one band before advancing in frequency. The two streams are combined as

$$\begin{aligned} Y_f &= \mathrm{Unflat}_f\big(\mathrm{SSM}_{freq}(\mathrm{Flat}_f(\mathcal{T}_{in}))\big), \\ Y_t &= \mathrm{Unflat}_t\big(\mathrm{SSM}_{time}(\mathrm{Flat}_t(\mathcal{T}_{in}))\big), \\ \mathcal{Z}_{out} &= \mathcal{T}_{in} + \mathrm{Linear}_{2d\to d}([Y_f \| Y_t]). \end{aligned} \tag{4}$$

For zero-based patch indices (f, t), Flat_f maps a patch to sequence index $tF' + f$ and Flat_t to $fT' + t$. The inverse permutations Unflat_f and Unflat_t restore both outputs to the same grid, so $Y_f[f, t, :]$ and $Y_t[f, t, :]$ refer to the same patch. Their features are concatenated ($\|$, giving $2d$ features), projected back to d by a shared linear layer, and added to the input token.

Finally, the prototype features are lifted back to 16 channel-aligned spectral embeddings:

$$H_{img} = AZ_{tf} + E_{ch}, \quad A = \mathrm{Softmax}(E^a_{ch} W_a E^T_{proto} + \eta M P^T), \tag{5}$$

where $H_{img} \in \mathbb{R}^{N \times d}$ is the channel-aligned spectral embedding, $A \in \mathbb{R}^{N \times K}$ is the lifting matrix, and $E_{ch} \in \mathbb{R}^{N \times d}$ is the microphone-channel embedding. The lifting matrix depends on the channel embeddings E^a_{ch}, the prototype embeddings E_{proto} and a learnable projection W_a. It also depends on the coupling matrix M learned by SC-CCM, whose strength is controlled by a scalar weight η. The softmax is taken over the K prototypes for each channel, so the lifting logits and $\eta M P^T$ are 16×4. Because P is not constrained to be nonnegative, $\eta M P^T$ is a signed coupling term rather than a probability of prototype membership. Further details of TG-LR-CWT are provided in Appendix B.

Information retained by prototype lifting. P has rank at most four, so the prototype waveforms cannot determine arbitrary 16-channel inputs, and Equation (5) constructs channel embeddings rather than reconstructed signals. Three properties limit the loss of channel information. First, full-channel information bypasses the compression. $M(X)$ is computed from all 16 original spectra, and the waveform branch processes all 16 channels and supplies the TMCA keys and values (Section 3.4). Second, P is trained with the task losses, so it retains task-relevant components. Third, the empirical evidence supports the design. The frozen learned projection retains 92.5% of the held-out waveform energy, close to the 92.9% retained by the optimal rank-4 singular value decomposition (SVD) projection fitted on training data (Appendix B.6). Four prototypes reach 69.2% JEM, compared with 69.4% for per-channel CWT (Section 4.5.2). Replacing channel-specific lifting with a broadcast of pooled TF features lowers JEM to 60.1% (Table A15).

3.3. Temporal Encoding and Channel-Coupling Prior

The waveform branch keeps the original 16-channel acoustic signals, preserving inter-channel timing, phase-related, and amplitude-difference cues that may be weakened by prototype-level CWT compression. It contains two components: a latent temporal routing module for waveform heterogeneity and a spectral-coupling module for channel relationships. CG-TCM adapts the temporal-clustering idea of DUET [24], and SC-CCM adapts its frequency-domain channel clustering. DUET is therefore included as a waveform-branch baseline (Appendix A.9).

3.3.1. Latent Temporal Routing

The Condition-Guided Temporal Clustering Module (CG-TCM) processes each channel waveform $x_n \in \mathbb{R}^L$ using a bank of temporal experts. The word "condition" refers to latent operating-regime patterns inferred from the signal, not to explicit height, intensity, or scene labels. The router computes:

$$G(x_n) = \mathrm{Softmax}(W_r \phi(x_n) + b_r), \quad H_n = \sum_{k=1}^{K_t} G(x_n)_k E_k(x_n), \tag{6}$$

where $\phi(\cdot)$ is a lightweight feature extractor, W_r and b_r are router parameters, and $G(x_n)_k$ is the routing weight assigned to the k-th expert. $E_k(\cdot)$ denotes the k-th temporal expert, $K_t = 4$ is the number of temporal experts, and $H_n \in \mathbb{R}^d$ is the temporal embedding for channel n. Stacking all channel outputs gives $H_{temp} \in \mathbb{R}^{N \times d}$. The routing weights are optimized through the downstream objective without scene-related metadata as privileged supervision for fault diagnosis.

3.3.2. Spectral-Coupling Channel Clustering

CG-TCM treats channels independently during temporal expert assignment, so an additional module is needed to model cross-channel response relationships. SC-CCM estimates a soft channel-coupling prior from learned comparisons of the channel amplitude spectra. For each channel,

$$X_n^{freq} = |\,\mathrm{rFFT}(x_n)| \in \mathbb{R}^F, \qquad F = \lfloor L/2 \rfloor + 1 = 2049, \tag{7}$$

where X_n^{freq} is the one-sided amplitude spectrum of channel n, including the DC and Nyquist bins. The learned quadratic dissimilarity between channels i and j is

$$D_{ij} = (X_i^{freq} - X_j^{freq})^T R (X_i^{freq} - X_j^{freq}), \quad R = A_q^T A_q, \tag{8}$$

where A_q is a learnable factor. The factorization gives $D_{ij} = \|A_q(X_i^{freq} - X_j^{freq})\|_2^2 \geq 0$, a symmetric, non-negative spectral dissimilarity; it is a mathematical property and imposes no physical law.

Relation to physical similarity. An ideal pure delay changes phase but not spectral magnitude, whereas attenuation, source-level differences and interference between sources or propagation paths change the amplitude spectra (Table A23). This motivates comparing amplitude spectra, but A_q mixes frequency bins and is learned through the task objective. D_{ij} is therefore a task-weighted spectral dissimilarity, not a calibrated physical distance. A small D_{ij} does not imply adjacent microphones.

The soft coupling prior is then generated as:

$$M_{ij} = \mathrm{Sigmoid}(\gamma - D_{ij}), \quad M \in (0,1)^{N \times N}, \tag{9}$$

where γ is a single learnable scalar shared by all channel pairs and windows. It is initialized at 1 and optimized with the network (Section 3.4); it is not set from microphone geometry or a sparsity target. γ sets the dissimilarity at which $M_{ij} = 0.5$. Because $D_{ii} = 0$, M is symmetric and satisfies $0 < M_{ij} \leq \mathrm{Sigmoid}(\gamma) = M_{ii}$ (Appendix C.3).

For finite arguments, M is dense, with entries strictly between zero and one; no thresholding or pruning is applied. We therefore describe it as a soft channel-coupling prior rather than a sparse mask. M is recomputed for each input window and reused in channel lifting and TMCA, where it acts as a relative attention bias (Section 3.4).

3.4. Topology-Masked Fusion and Task-Aware Decoding

3.4.1. Topology-Masked Cross-Attention

Topology-Masked Cross-Attention (TMCA) aligns the channel-aligned spectral embeddings H_{img} with the waveform-derived temporal embeddings H_{temp}. The spectral embeddings are used as queries, and the temporal embeddings are used as keys and values:

$$Q_{att} = H_{img} W_Q, \quad K_{att} = H_{temp} W_K, \quad V = H_{temp} W_V, \tag{10}$$

where Q_{att}, K_{att}, and V are the query, key, and value matrices, and W_Q, W_K, and W_V are learnable projections. The learned coupling prior M biases attention between channel pairs:

$$\begin{aligned} S &= \frac{Q_{att} K_{att}^T}{\sqrt{d}}, \quad A_{masked} = \mathrm{Softmax}(S + \log(M + \epsilon)), \\ Z &= A_{masked} V + H_{img}. \end{aligned} \tag{11}$$

Here, S is the raw attention score matrix, A_{masked} is the masked attention weight matrix, Z is the fused multi-modal representation, and ϵ is a small constant for numerical stability. Adding a structural term to attention logits follows Graphormer [23]; here, the term comes from the learned spectral-coupling matrix rather than from distances on a supplied graph.

Two uses of the coupling prior. SC-CCM computes one matrix $M(X) \in (0,1)^{16\times16}$ per input window. In channel lifting (Equation (5)), the term $\eta MP^T \in \mathbb{R}^{16\times4}$, with entries $\eta \sum_j M_{ij}P_{kj}$, is added to the channel–prototype logits before the softmax over prototypes. In TMCA, $\log(M+\epsilon) \in \mathbb{R}^{16\times16}$ is added to the spectral-query/waveform-key scores before the softmax over the 16 key channels, which is equivalent to

$$(A_{masked})_{ij} = \frac{\exp(S_{ij})(M_{ij}+\epsilon)}{\sum_\ell \exp(S_{i\ell})(M_{i\ell}+\epsilon)}.$$

The prior therefore multiplies the unnormalized attention weights. Weakly coupled pairs receive less relative weight, but they are not removed when their raw scores are high. Although M is symmetric, A_{masked} need not be, because queries and keys come from different representations.

3.4.2. Task-Aware Mixture-of-Experts

The fused representation must support two related objectives with potentially different feature requirements. Spatial-energy scene parsing depends strongly on phase and amplitude structure, whereas fault diagnosis depends on diagnostic transients and spectral harmonics. The TA-MoE decoder therefore combines always-active shared experts with a bank of routed experts and uses one router per task. This organization follows the shared and task-specific experts of PLE [28] and the shared-plus-routed design of DeepSeekMoE [29]. For task $t \in \{scene, diagnosis\}$,

$$\begin{aligned} H_t &= H_{shared}(Z) + H_{routed,t}(Z), \\ H_{shared}(Z) &= \sum_{i\in\Omega_{shared}} E_i(Z), \\ H_{routed,t}(Z) &= \sum_{j\in\mathcal{K}_t(Z)} w_{t,j}(Z)E_j(Z), \qquad \mathcal{K}_t(Z) = \mathrm{Top3}(g_t(Z)). \end{aligned} \tag{12}$$

Here, Ω_{shared} contains the two shared experts, and each router g_t selects three of the same eight routed experts; the two tasks may select overlapping subsets. The selected gate weights $w_{t,j}$ are renormalized to sum to one, and each shared expert has coefficient one. No additional coefficient mixes the two sums; their balance is learned through the expert outputs and gates. Shared experts receive gradients from both task losses and carry common features. A routed expert receives a task gradient only when that task selects it, which allows task-specific specialization. These routers are distinct from the CG-TCM router and receive no operating-condition labels.

The task-specific representations are classified by two lightweight heads. Each MLP_t below consists of channel-mean pooling and a $256 \to 16$ linear map.

$$\begin{aligned} \hat{Y}_{scene} &= \mathrm{Softmax}(\mathrm{MLP}_{scene}(H_{scene})), \\ \hat{Y}_{diagnosis} &= \mathrm{Softmax}(\mathrm{MLP}_{diagnosis}(H_{diagnosis})). \end{aligned} \tag{13}$$

The total objective is:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{CE}(\hat{Y}_{scene}, Y_{scene}) + \lambda_2\mathcal{L}_{CE}(\hat{Y}_{diagnosis}, Y_{diagnosis}) + \lambda_3\mathcal{L}_{aux}, \tag{14}$$

where Y_{scene} and $Y_{diagnosis}$ are the two supervised labels; λ_1, λ_2, and λ_3 are loss weights; and $\mathcal{L}_{aux}$ is a load-balancing term on the routed experts, following Switch Transformers [30] (Appendix C.4). We use $(\lambda_1, \lambda_2, \lambda_3) = (1, 1, 0.1)$. Equal task weights give equal priority to two 16-class tasks with the same chance-level cross-entropy ($\log 16 \approx 2.77$). $\mathcal{L}_{aux}$ equals 2 under uniform routing, so with $\lambda_3 = 0.1$, it regularizes routing without dominating the task losses. A systematic sensitivity study of these weights was not performed (Section 4.7).

4. Results and Discussion

4.1. Experimental Setup and Task Formulation

To evaluate MM-MoE under controlled replay-based 3D acoustic masking, we use the 3D Spatial-Interference Gearbox Dataset. Real planetary-gearbox fault recordings are first acquired from a test bench and then replayed by loudspeakers at controlled positions, heights, and sound-pressure levels. The source recordings were acquired from a single-stage planetary gearbox at a constant input speed of 1500 rpm and a load of 10 N m. The sun gear was either intact or had a prepared pitting, crack or missing-tooth defect. Each health state was recorded in six separate 60 s runs (Appendix A.1, Table A1). A 16-channel microphone array records the resulting mixed acoustic field at 10 kHz. The dataset contains 4 replayed source-region combinations, 4 height–intensity source configurations, and 16 coupled gearbox health-state combinations, giving $4 \times 4 \times 16 = 256$ experimental configurations. Each configuration contains 5 raw recordings, and each recording is segmented into 24 non-overlapping windows of length $L = 4096$. The final dataset therefore contains $256 \times 5 \times 24 = 30,720$ segmented samples.

Unlike single-label diagnosis, we formulate the benchmark as two coupled 16-class tasks:

- **Task 1: Spatial-Energy Scene Parsing (16 Classes):** The model identifies the replayed source-pair region and its elevation–intensity configuration.
- **Task 2: Coupled Fault Diagnosis (16 Classes):** The model classifies the joint health state of the two replayed gearbox fault recordings.

We use a chronological file-level split. For every configuration, files 1–3 are used for training, file 4 for validation and file 5 for testing. All windows of a recording therefore stay in one partition, and the source segments are disjoint across partitions. This split tests new recordings of configurations seen during training, within the same replay apparatus, array and room; transfer to excluded source positions is evaluated separately in Section 4.3.3.

Performance is reported as Top-1 accuracy and Macro-F1 for each task, and as Joint Exact Match (JEM), the fraction of test windows for which both task predictions are correct (Equation (A4)). Results with $\pm$ are means and sample standard deviations (SDs) over five independent training runs.

Label Usage Protocol

Although the dataset metadata includes source region, height, sound-pressure level, and fault state, the model is trained only with the two target labels above. The CG-TCM router does not use HB/HS/LB/LS metadata or pair-level height–intensity configurations as auxiliary supervision. During both training and inference, routing weights are inferred only from the raw waveform input. Hardware, partitioning, evaluation metrics, and hyperparameters are detailed in Appendix A.

4.2. Comparison with Evaluated Baselines

We first compare MM-MoE with traditional, single-modal, multi-modal-fusion, and multi-task baselines under the same dual 16-class formulation (Table 3). All models use the same chronological file-level split, 16-channel windows of 4096 samples and two 16-class targets. Each deep model was tuned with 30 validation-only trials of the tree-structured Parzen estimator (TPE) over the same search space and then trained in five independent runs (Appendix A.9). The methods were not matched in model size, which is reported in Table 4.

Table 3. Performance on the dual 16-class spatial–acoustic tasks. Values with ± are means and sample standard deviations over five independent training runs on the same file-level split. The final column gives the JEM advantage of MM-MoE over each comparator and its unadjusted 95% Welch confidence interval, in percentage points.

Method	Scene		Fault		JEM	ΔJEM (pp)
	Acc. (%)	F1 †	Acc. (%)	F1 †	(%)	95% CI
Traditional						
SVM + Spectral	31.5 ± 1.2	0.285	28.4 ± 1.5	0.251	10.2 ± 1.1	+59.0 57.6–60.4
RF + Wavelet	35.8 ± 1.4	0.321	32.1 ± 1.3	0.298	12.5 ± 1.0	+56.7 55.4–58.0
Single-modal						
1D-CNN (Time)	68.4 ± 1.1	0.652	51.2 ± 1.2	0.495	36.8 ± 1.3	+32.4 30.8–34.0
DUET (Time)	74.5 ± 0.9	0.721	58.7 ± 1.0	0.564	45.2 ± 1.1	+24.0 22.6–25.4
ResNet-50 (Image)	52.3 ± 1.3	0.491	64.5 ± 0.8	0.621	35.4 ± 1.2	+33.8 32.3–35.3
Vanilla Vim (Image)	55.6 ± 1.0	0.512	66.8 ± 0.7	0.645	38.9 ± 1.0	+30.3 29.0–31.6
Multimodal fusion						
Standalone Plain CA	76.2 ± 0.8	0.745	68.5 ± 0.9	0.662	54.1 ± 0.9	+15.1 14.0–16.2
MC-ABDS (MTL)	78.5 ± 0.7	0.768	67.4 ± 0.8	0.651	55.8 ± 0.8	+13.4 12.4–14.4
Proposed						
MM-MoE (Proposed)	**89.4 ± 0.5**	**0.882**	**76.5 ± 0.6**	**0.751**	**69.2 ± 0.6**	Reference

Italic rows denote method groups. Underlining marks the best baseline for each metric; bold marks MM-MoE. † Macro-F1 on a 0–1 scale, reported as point values. ΔJEM is MM-MoE minus the comparator. Intervals describe training variability on the fixed split; Holm-adjusted tests for all comparisons are given in Appendix E.

Traditional spectral or wavelet features reach JEM below 13%. The single-modal models show complementary task profiles. The waveform-based DUET [24], adapted from forecasting to classification, reaches 74.5% scene accuracy but 58.7% fault accuracy. The image-only models reach up to 66.8% fault accuracy but at most 55.6% scene accuracy. This pattern is consistent with waveforms retaining inter-channel timing cues and TF maps emphasizing fault-related structure, although these models also differ in architecture. The fusion baselines, Standalone Plain CA and MC-ABDS [3], reach 54.1% and 55.8% JEM. MM-MoE reaches 89.4% scene accuracy, 76.5% fault accuracy and 69.2% JEM. MM-MoE differs from these baselines in several components, so this framework-level comparison does not attribute the gain to individual modules. Component effects are examined in Section 4.5.

Significance. The strongest baseline depends on the metric. Relative to MC-ABDS, MM-MoE improves scene accuracy by 10.9 percentage points (95% Welch confidence interval (CI): 10.0–11.8; Holm-adjusted $p_H = 1.90 \times 10^{-7}$) and JEM by 13.4 points (12.4–14.4; $p_H = 1.19 \times 10^{-7}$). Relative to Standalone Plain CA, the strongest fault-accuracy baseline, the improvement is 8.0 points (6.9–9.1; $p_H = 5.28 \times 10^{-6}$). All 24 comparisons with the eight baselines remain significant after Holm correction over the full family of 30 tests

(Appendix E). Macro-F1 improves by 0.114 for scene parsing and 0.089 for fault diagnosis over the best baseline values.

Interpretation of JEM. A window counts as a joint exact match only when both outputs are correct. Task 1 must identify the source-pair region and height–intensity configuration, and Task 2 the health states of both sources. A correct joint prediction therefore states which faults are present and where they come from. If only Task 1 is wrong, a correctly recognized fault is assigned to the wrong location or configuration, and a region error could send an inspection to the wrong machine. If only Task 2 is wrong, at least one health state is misclassified, causing a missed fault or a false alarm. Relative to MC-ABDS, JEM rises from 55.8% to 69.2%: the share of windows with at least one wrong output falls from 44.2% to 30.8%, a relative reduction of about 30%.

Recording-level evaluation. The 24 windows of each test recording are correlated subsamples of one acquisition. We therefore also average the 16-class probabilities of all windows of a recording separately for each task (soft voting); recording-level JEM requires both aggregated predictions to be correct (Appendix A.6.1). Each trained model is evaluated separately. At the recording level, MM-MoE reaches 96.1% scene accuracy, 89.8% fault accuracy and 87.1% JEM (95% CI over recordings: 85.0–89.1%), compared with 88.7%, 79.7% and 73.8% for MC-ABDS. Aggregating the first six windows (2.46 s) already gives 97.9% of the full-recording JEM. We also used a bootstrap that resamples both training runs and test recordings, keeping the windows of each recording together. It gives JEM differences between MM-MoE and MC-ABDS of 13.4 points (95% CI 10.5–16.4) at the window level and 13.3 points (9.0–17.7) at the recording level.

Computational cost. Table 4 and Figure 2 summarize the accuracy–cost trade-off. The model sizes were not matched: total parameter counts range from 1.18 to 23.57 M. Relative to MC-ABDS, MM-MoE uses 42.7% more parameters (13.76 vs. 9.64 M), 34.9% more floating-point operations (FLOPs; 6.38 vs. 4.73 G) and 39.5% more latency (31.8 vs. 22.8 ms per window). It thus adds 9.0 ms of latency for a gain of 13.4 JEM points. MM-MoE has an active-parameter count close to that of Standalone Plain CA (10.92 vs. 10.86 M) and 15.1 points higher JEM. However, it has 26.7% more total parameters and 24.6% more FLOPs, so this comparison is not capacity-matched either. Compared with the larger ResNet-50, MM-MoE uses 41.6% fewer parameters and has 66.6% lower latency.

Table 4. Predictive performance and computational cost of the evaluated implementations for one 16-channel, 4096-sample window.

Method	JEM (%)	Params, Total/Active (M)	FLOPs (G)	Inference Memory (GB)	Peak GPU Memory (GB)	Latency (ms)
1D-CNN (Time)	36.8	1.18/1.18	0.31	0.14	0.9	1.6
DUET (Time)	45.2	2.74/2.74	0.58	0.20	1.4	2.9
ResNet-50 (Image)	35.4	23.57/23.57	130.79	0.46	8.0	95.1
Vanilla Vim (Image)	38.9	7.92/7.92	3.86	0.58	5.6	18.7
Standalone Plain CA	54.1	10.86/10.86	5.12	0.74	7.2	24.6
MC-ABDS (MTL)	55.8	9.64/9.64	4.73	0.68	6.8	22.8
MM-MoE ($K = 4$)	**69.2**	**13.76/10.92**	**6.38**	**0.92**	**8.4**	**31.8**
MM-MoE + per-channel CWT	69.4	13.76/10.92	14.92	1.75	19.7	78.6

Bold marks MM-MoE with the default $K = 4$. M and G denote 10^6 parameters and 10^9 operations. Inference memory is measured at batch size one; peak GPU memory is measured during training (Appendix A.8). Latency includes time–frequency preprocessing, where applicable, and network inference. ResNet-50 uses a shared Torchvision v1.5 encoder over the 16 maps; its FLOPs count convolution and classifier arithmetic at two FLOPs per multiply–accumulate operation.

At the point estimates, the accuracy–latency frontier in Figure 2 consists of 1D-CNN, DUET, MC-ABDS, MM-MoE ($K = 4$) and MM-MoE with per-channel CWT. No evaluated method achieves higher JEM than MM-MoE ($K = 4$) at equal or lower latency. DUET remains the better option when latency is the binding constraint: it is about 11 times faster (2.9 ms per window) at 24.0 points lower JEM. Within MM-MoE, per-channel CWT raises FLOPs from 6.38 to 14.92 G, latency from 31.8 to 78.6 ms and inference memory from 0.92 to 1.75 GB. In return, JEM changes by only 0.2 points, which is not significant (Section 4.5.2). All latencies are well below the 409.6 ms duration of one window. They were, however, measured on the experimental hardware and do not by themselves establish streaming or edge-device performance (Section 4.7).

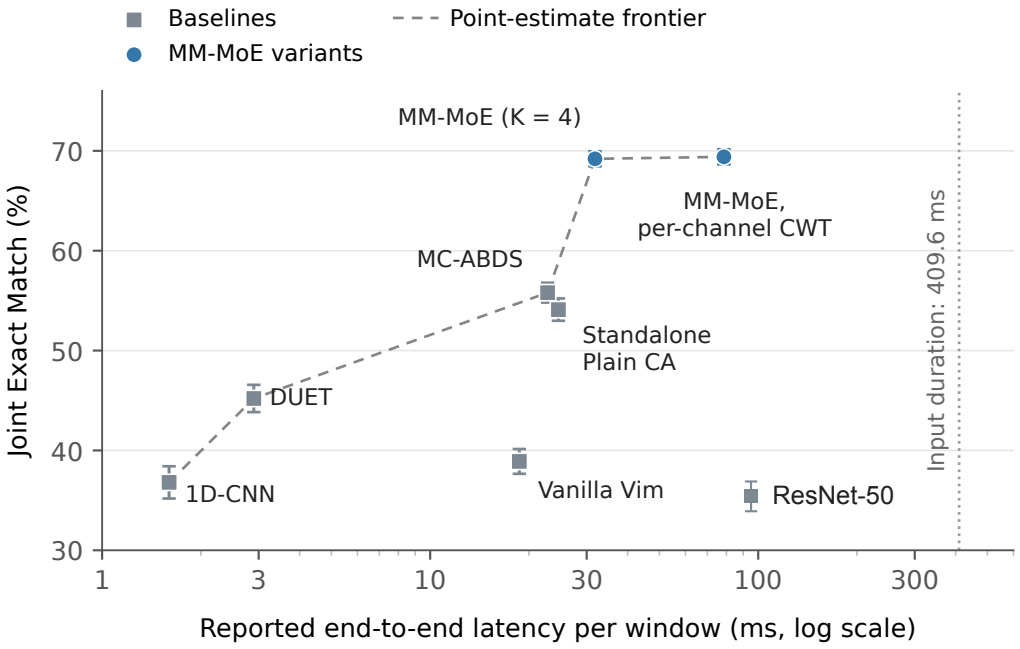


Figure 2. JEM–latency trade-off for the eight evaluated implementations in Table 4. Latency includes time–frequency preprocessing, where applicable, and network inference, and it is shown on a logarithmic axis. Markers show mean JEM; vertical bars are 95% Student-t intervals for the mean over five training runs. The dashed line connects non-dominated point estimates, and the dotted line marks the 409.6 ms window duration.

4.3. Error Analysis Under Spatial-Energy Ambiguity and Masking

We next examine whether the aggregate gains hold under the two main sources of difficulty in the dataset: joint spatial-energy ambiguity and strong acoustic masking. We then evaluate a configuration-level holdout of source regions.

4.3.1. Deconstructing Spatial-Energy Parsing Errors

Task 1 requires the model to jointly resolve source-pair region, vertical height, and playback intensity. Figure 3 decomposes scene-parsing errors into location-only, energy/height-only, and simultaneous errors. The main bottleneck for the evaluated baselines was energy–height classification rather than horizontal localization alone. MC-ABDS produced a 5.1% location-only error rate but a 13.5% energy/height-only error rate. In contrast, MM-MoE reduced the total error rate to 10.6% and the energy/height error to 6.6%. In the ordered ablation (Section 4.5.1), the largest scene-accuracy increment (6.8 percentage points) occurs when the coupling prior and TMCA are introduced (V5–V6). This is consistent with learned channel relationships helping to resolve energy and height.

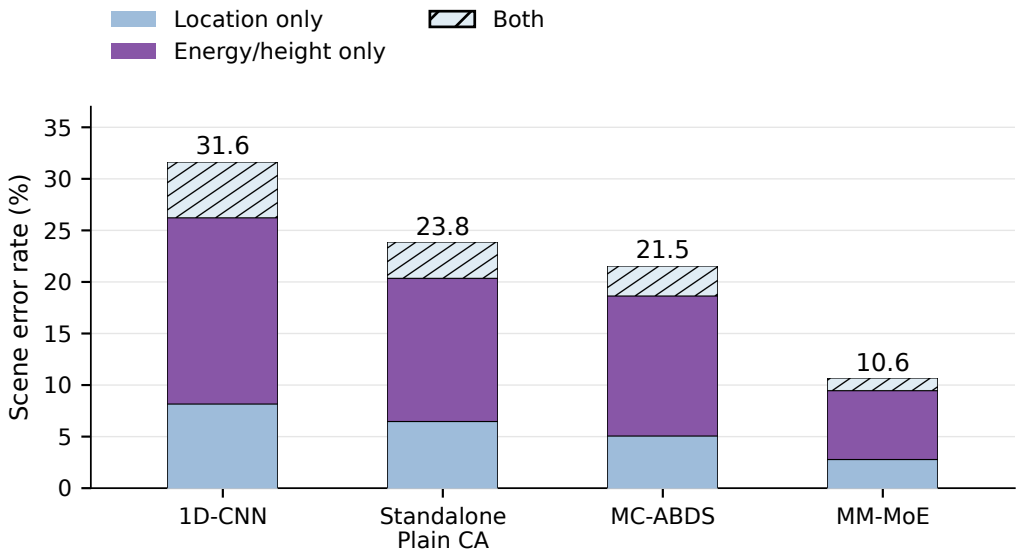


Figure 3. Fine-grained error decomposition for Task 1 (Spatial-Energy Scene Parsing). The total error rate is partitioned into location-only misclassifications, energy/height-only misclassifications, and instances where both are predicted incorrectly. Standalone Plain CA denotes the baseline, not internal variant V5.

4.3.2. Fault Diagnosis Under Strong Acoustic Masking

Figure 4 compares the favorable condition (60 dB target playback setting) with the strong-masking condition (40 dB target and 60 dB interferer settings; Appendix A.3). The 16-class fault accuracy of MC-ABDS decreases from 75.1% to 59.7% (−15.4 points), whereas that of MM-MoE decreases from 79.2% to 73.8% (−5.4 points; ±0.6 over five runs, Table A11). The MM-MoE advantage thus grows from 4.1 points under the favorable condition to 14.1 points under strong masking. The interferer is another replayed gearbox recording, so this result concerns masking by a competing gearbox source rather than heterogeneous plant noise (Section 4.7).

Most affected fault types. A source-resolved analysis with four health states per source (Table A12) shows that MM-MoE accuracy drops from 88.5% to 81.5% under strong masking, compared with 84.0% to 68.0% for MC-ABDS. Pitting is the most affected fault (−12 points for MM-MoE, −23 for MC-ABDS) and missing tooth the least affected (−2 and −11 points). Most masked-condition errors are faults misclassified as normal (Table A13), i.e., missed detections rather than confusions between fault types.

Transient versus harmonic cues. Under strong masking, perturbing the phase common to all channels reduces the source-level accuracy of MM-MoE from 81.5% to 63.0%, whereas notching the mesh-harmonic comb reduces it to 75.0%. For MC-ABDS, the corresponding accuracies are 53.0% and 62.0%, compared with 68.0% without perturbation. The masked-condition decisions therefore rely more on phase and timing information, which carries the transient and inter-channel structure, than on harmonic amplitudes. This experiment shows which cues the models use; it does not by itself establish which features physical masking removes.

Role of the coupling prior. SC-CCM computes M from the amplitude spectra of each window (Equations (7)–(9)), so the prior adapts to the recorded mixture. Through $\log(M + \epsilon)$, channel pairs with low learned coupling receive less relative attention. Consistently, introducing the coupling prior with TMCA (V5–V6) improves source-level accuracy by 1.6 points under favorable conditions but by 6.5 points under strong masking. A fixed geometric adjacency lies between the two (86.8% and 76.0%). Because this step changes lifting and attention together, we attribute the gain to the combined SC-CCM/TMCA configuration. The image branch also contributes. Under masking, TG-LR-CWT ($K = 4$) loses 5.4 points, close

to per-channel CWT. PCA-CWT, which encodes a single waveform obtained by principal component analysis (PCA), loses 8.6 points (Table A11).

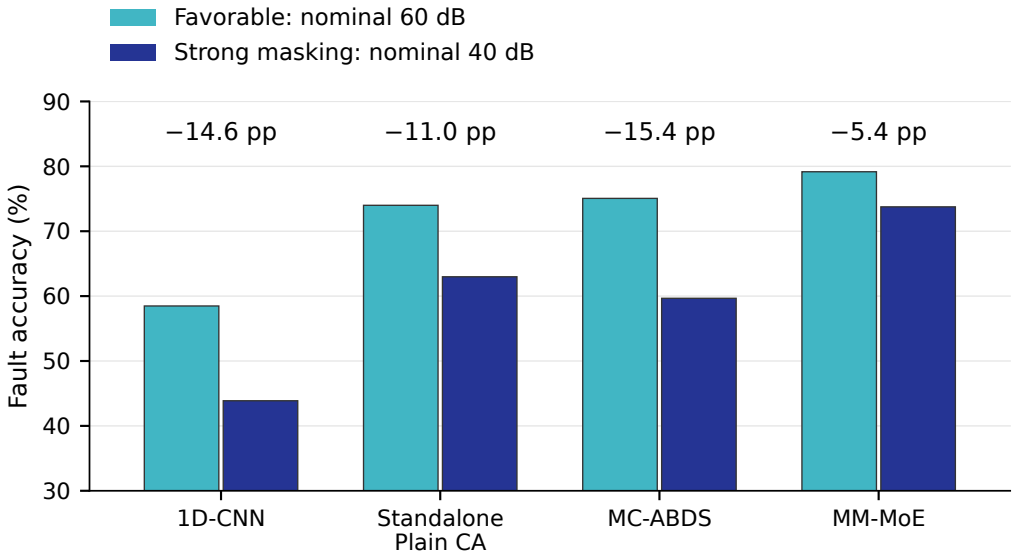


Figure 4. The figure shows 16-class fault diagnosis accuracy under the favorable and strong-masking playback conditions. The 60 dB and 40 dB values are playback-level settings, not sound-pressure levels at the array (Appendix A.3). Annotations give the accuracy decreases in percentage points (pp); run-level variability is reported in Tables A11 and A20. Standalone Plain CA is the baseline, distinct from V5.

4.3.3. Configuration-Level Generalization

The file-level split tests new recordings of configurations seen during training. To evaluate unseen source positions, we performed a four-fold leave-one-source-pair-region-out experiment, holding out Regions 1–2, 3–4, 5–6 and 7–8 in turn. In each fold, files 1–3 of the remaining 192 configurations were used for training (13,824 windows) and file 4 for validation and checkpoint selection (4608 windows). Normalization statistics were fitted on the training files only. All five files of the 64 excluded configurations formed the unseen-region test set (7680 windows), and file 5 of the training regions formed a seen-region reference (4608 windows). MM-MoE and MC-ABDS were trained from scratch with five runs per method and fold.

Withholding a region removes four of the 16 scene classes from training, so full scene accuracy and JEM are not meaningful endpoints. We therefore report three endpoints decoded from the original heads. Ordered fault-pair accuracy is the Task 2 accuracy. Order-invariant fault-pair accuracy also accepts the predicted pair with its two sources swapped. Height–intensity accuracy is obtained by marginalizing the scene probabilities over region, $p(h \mid x) = \sum_r p_{\text{scene}}(r, h \mid x)$.

Averaged over the four folds, the ordered fault accuracy of MM-MoE decreases from 75.5% in seen regions to 68.3% in excluded regions (−7.2 points). For MC-ABDS, it decreases from 66.2% to 56.8% (−9.4 points), so MM-MoE remains 11.5 points ahead on unseen source positions. For MM-MoE, order-invariant accuracy changes from 80.2% to 73.4% and height–intensity accuracy from 90.5% to 82.3%. The SD across the four held-out regions is about two points for each endpoint (Table A14). This experiment varies source positions within one chamber, replay apparatus and array; transfer to other measurement conditions requires separate validation (Section 4.7).

4.4. Comparison with Single-Task and Hard-Sharing Training

Figure 5 compares MM-MoE with independently trained single-task models and a hard-sharing multi-task model. The single-task models reach 84.5% scene and 70.2% fault accuracy; their combined JEM, computed from joint correctness on the same test windows rather than from the product of the two marginal accuracies, is 59.3%. Hard-sharing MTL reaches 55.4% JEM, 3.9 points below the single-task reference, and MM-MoE reaches 69.2%. The single-task and hard-sharing models are independently trained, unpaired configurations that are not capacity-matched, so the −3.9-point gap is a descriptive difference rather than a test of negative transfer. Within the internal ablation, the variant with shared feed-forward network (FFN) heads, V6 (85.3% scene, 72.8% fault) already exceeds the single-task accuracies. Replacing its shared heads with TA-MoE (V6–V7) adds 4.1 and 3.7 points (Section 4.5.1).

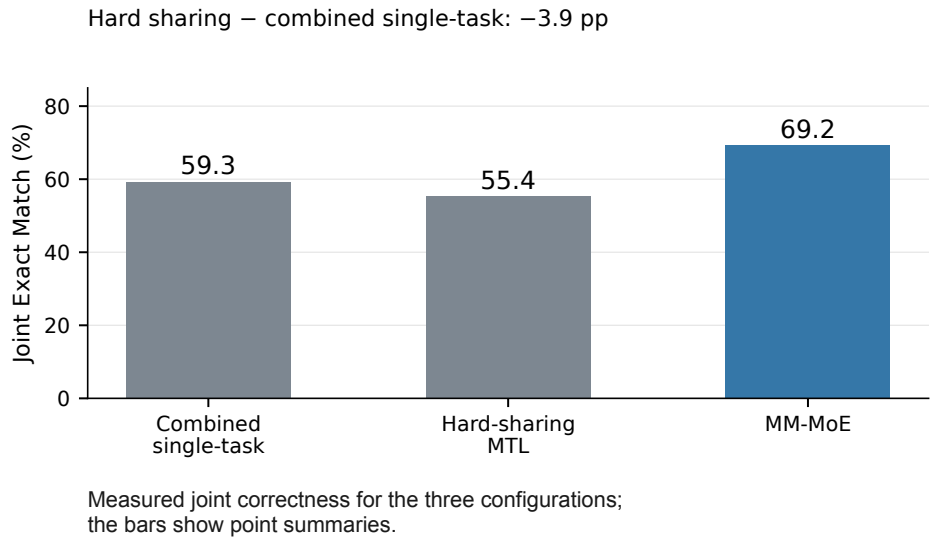


Figure 5. JEM of the combined single-task reference, hard-sharing MTL and MM-MoE. Combined single-task JEM requires both single-task predictions to be correct on the same window. The annotation gives the difference between hard-sharing MTL and the single-task reference in percentage points (pp).

Expert Routing

Figure 6 shows the router probabilities before Top-3 selection, averaged over the evaluation windows of one trained model. The three largest weights occur at E1–E3 for scene parsing and E5–E7 for fault diagnosis, indicating task-dependent expert preferences. Expert indices are arbitrary across independently trained runs. We therefore assessed stability with permutation-invariant statistics of the per-window Top-3 selections, computed within each of five runs on the same 2048 test windows (definitions in Appendix A.6.2; values in Table A5).

The statistics are stable across runs. Normalized load entropy is 0.813 ± 0.005 for the scene router and 0.861 ± 0.004 for the diagnosis router. The Jensen–Shannon divergence between the selected loads of the two routers is 0.305 ± 0.008 nats, so the two tasks use the expert bank differently in every run. The entropies are also similar across the four source-region groups (0.807–0.818 and 0.855–0.867) and the two masking conditions. Each router has a preferred expert, which takes 31.2% (scene) and 26.8% (diagnosis) of the routed slots, compared with 12.5% under uniform use. The remaining selections are spread across the bank, with on average 41 and 48 distinct selected sets per run. No expert receives less than 1% of the routed slots in any run, so no expert collapse is observed at the final checkpoints.

The auxiliary loss ($\lambda_3 = 0.1$) and the always-active shared experts are the design elements intended to prevent collapse; routing dynamics during training were not analyzed.

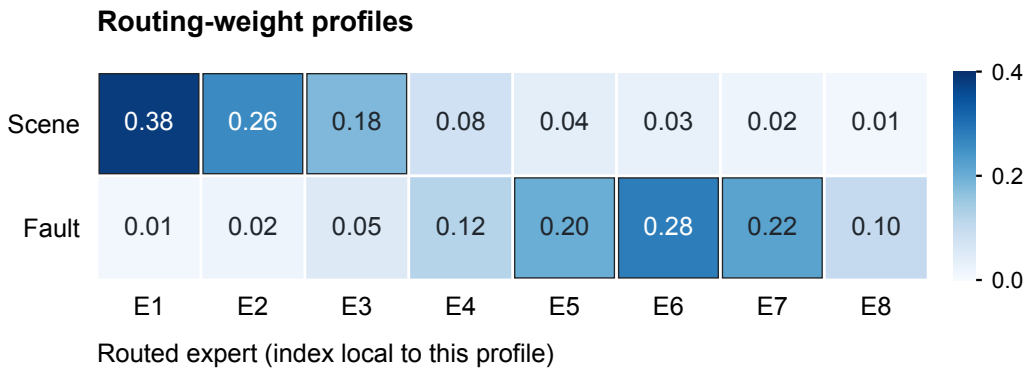


Figure 6. Routing-weight profiles of the scene and diagnosis routers over the eight routed TA-MoE experts: softmax probabilities before Top-3 selection, averaged over the evaluation windows of one trained model. Outlined cells mark the three largest weights of each row. Repeated-run statistics are given in Table A5.

4.5. Ablation Studies

The ablations compare architectural choices within the proposed framework. Ordered comparisons describe how performance changes as components are added; they do not control for model capacity or isolate interactions between components. Extended results are provided in Appendix D.

4.5.1. Module-Level Ablation

Figure 7 and Table A10 follow the ordered dual-branch path V3–V7, with the single-modal variants V1 (temporal) and V2 (image) as references. V5, "Plain CA (internal)", uses the internal dual branches and shared FFN heads with plain cross-attention (CA). It differs from the Standalone Plain CA baseline in Table 3 (78.5%/69.4% vs. 76.2%/68.5% scene/fault accuracy), and the two labels are kept distinct throughout.

The dual-branch model without interaction (V3) reaches 75.2% scene and 67.5% fault accuracy, close to the better single-modal model for each task (74.5% for V1 on scene parsing and 66.8% for V2 on fault diagnosis). Concatenation (V3–V4) adds +1.6/+0.7 points (scene/fault), and plain cross-attention (V4–V5) a further +1.7/+1.2 points. Introducing the coupling prior with TMCA (V5–V6) gives the largest step, +6.8/+3.4 points. It mainly benefits scene recognition, which is consistent with channel relationships being most informative for spatial parsing. Replacing the shared FFN heads with TA-MoE (V6–V7) gives the most balanced step, +4.1/+3.7 points and the largest one for fault diagnosis. The full model reaches 89.4% and 76.5%. JEM increases from 50.8% (V3) to 62.1% (V6) and 69.2% (V7).

These increments depend on the components already present and on the order in which they are added, so they are not independent contributions. The last two steps account for 10.9 of the 14.2 scene points and 7.1 of the 9.0 fault points along this path. V5–V6 introduces the SC-CCM prior in both channel lifting and attention and therefore evaluates SC-CCM and TMCA together. V6–V7 also increases the parameter count from 11.39 to 13.76 M, so it does not separate architecture from capacity. It is therefore interpreted as an improvement over the internal shared-FFN configuration. Only V7 has repeated-run estimates. The two largest steps are large relative to its run-level SD (0.5–0.6 points), whereas the first two increments (0.7–1.7 points) are close to run-to-run variation and are not interpreted further.

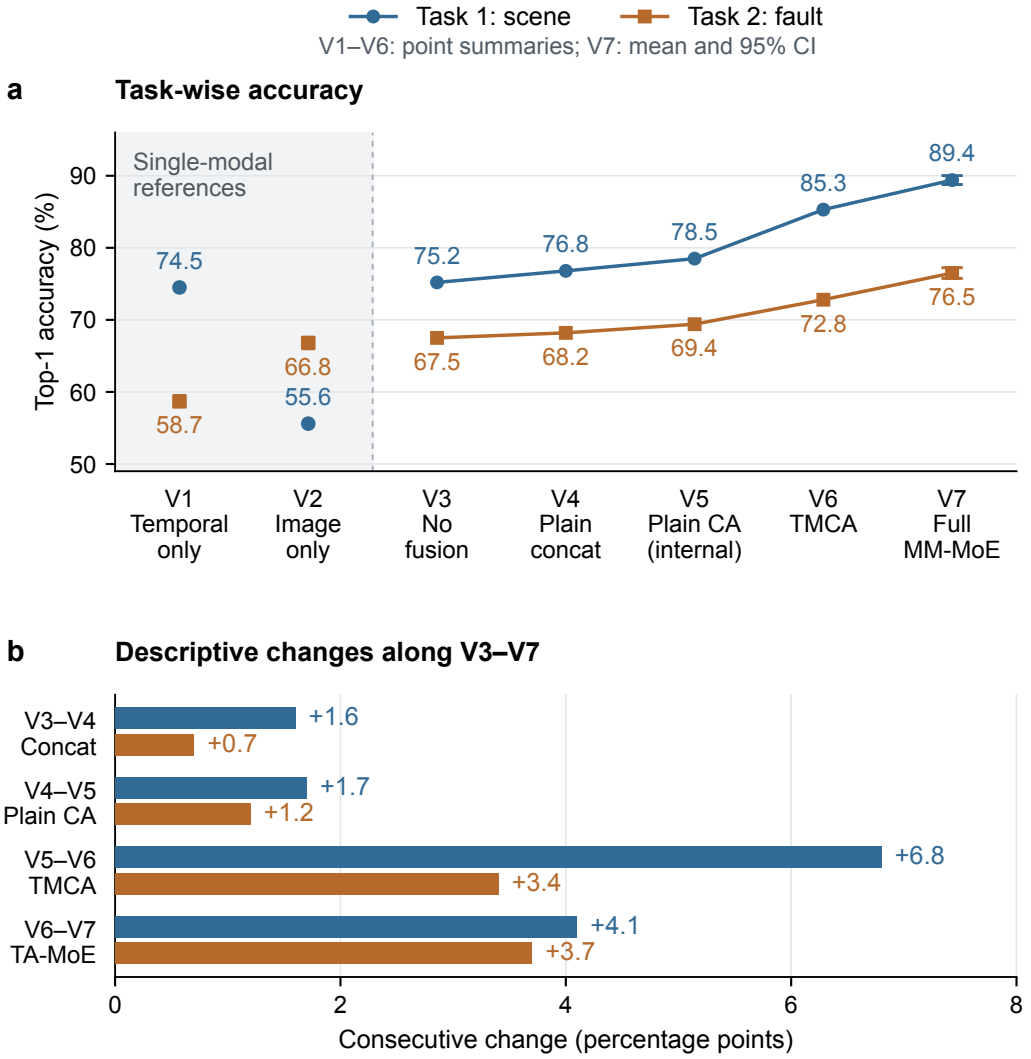


Figure 7. Module-level ablation under the dual 16-class setting. (**a**) Scene and fault accuracies (Table A10); V1 and V2 are single-modal references, and lines follow the ordered path V3–V7. Error bars on V7 are 95% t intervals over five training runs (Table A18); V1–V6 are point values. V5 is the internal Plain CA variant, distinct from the standalone baseline. (**b**) Consecutive accuracy changes in percentage points.

4.5.2. Image-Branch Encoding Strategy

Table 5 compares image-branch encodings within the MM-MoE backbone. PCA-CWT (A1) uses a fixed projection onto one dominant waveform, and the LR-CWT variant with $K = 2$ (A3) omits topology refinement. These references therefore differ from TG-LR-CWT in more than the prototype count. Within TG-LR-CWT, $K = 4$ reaches 69.2% JEM, compared with 69.3% for $K = 8$ (twice the CWT operations) and 69.4% for per-channel CWT (16 operations). Neither difference is significant: $K = 4$ minus $K = 8$ is -0.1 points (95% Welch CI: -1.0 to 0.8), and $K = 4$ minus per-channel CWT is -0.2 points (-1.1 to 0.7), both with $p_H = 1.00$ (Table A22).

Table 5. Image-branch encoding comparisons under MM-MoE. Panel (**a**) reports predictive performance and CWT operation counts; panel (**b**) reproduces the computational measurements for the default and per-channel configurations from Table 4.

(**a**) **Encoding configurations and performance**					
Variant	**Encoding**	**CWT Ops**	**Task 1 Acc. (%)/F1**	**Task 2 Acc. (%)/F1**	**JEM (%)**
A1	PCA-CWT, single dominant waveform	1	84.1 ± 0.7 0.827	72.0 ± 0.8 0.704	62.7 ± 0.8
A2	Per-channel CWT	16	89.7 ± 0.5 0.884	76.8 ± 0.6 0.754	69.4 ± 0.6
A3	LR-CWT, $K = 2$, no topology refinement	2	86.9 ± 0.6 0.855	73.9 ± 0.7 0.724	65.9 ± 0.7
A4	**TG-LR-CWT, $K = 4$ (default)**	**4**	**89.4 ± 0.5 0.882**	**76.5 ± 0.6 0.751**	**69.2 ± 0.6**
A5	TG-LR-CWT, $K = 8$	8	89.6 ± 0.5 0.883	76.6 ± 0.6 0.752	69.3 ± 0.6
(**b**) **Computational comparison**					
Encoding		**CWT Ops**	**Relative CWT Operation Count**	**FLOPs (G)**	**End-to-End Latency (ms/window)**
TG-LR-CWT, $K = 4$		**4**	**25%**	**6.38**	**31.8**
Per-channel CWT		16	100%	14.92	78.6

Bold marks the default configuration (A4). F1 denotes Macro-F1. Entries with ± are mean ± sample SD over five training runs. CWT ops counts time–frequency transforms per window; the relative count is normalized to 16 transforms and is not a ratio of total inference cost. Latency includes time–frequency preprocessing and network inference (Table 4).

Rationale for the default K. Among the evaluated TG-LR-CWT settings, $K = 4$ is therefore the lowest-cost one that matches per-channel accuracy. It uses $4/16 = 25\%$ of the CWT operations. Its total cost is higher, at 42.8% of the per-channel FLOPs (6.38 vs. 14.92 G) and 40.5% of the latency (31.8 vs. 78.6 ms per window), because the temporal branch, fusion and task heads do not shrink with K. We present $K = 4$ as a practical accuracy–cost default for the present 16-channel array. Intermediate counts ($K = 3, 5, 6$) were not evaluated, and the preferred count may differ for other arrays or datasets. The four learned prototypes are not interpreted as physically identifiable acoustic modes. Additional comparisons of channel lifting, phase-aware TF construction and encoding cost are reported in Appendix D.

4.5.3. Component Configuration Comparisons

Figure 8 summarizes encoder, temporal-routing, attention and decoder configurations. In each panel, the full configuration reaches the highest fault accuracy among its alternatives. In Figure 8c, for example, the learned coupling prior outperforms full attention, a random sparse mask and a fixed geometric adjacency. Because these configurations were not designed as controlled single-factor comparisons, we treat the panels as descriptive and use the ordered ablation above for attribution.

4.6. Qualitative Interpretability and Representation Analysis

Finally, we examine attention patterns and learned representations. Attention maps and t-SNE projections are qualitative views of internal model behaviour. We complement them with quantitative readouts in the original feature space and do not use either as evidence of a physical mechanism.

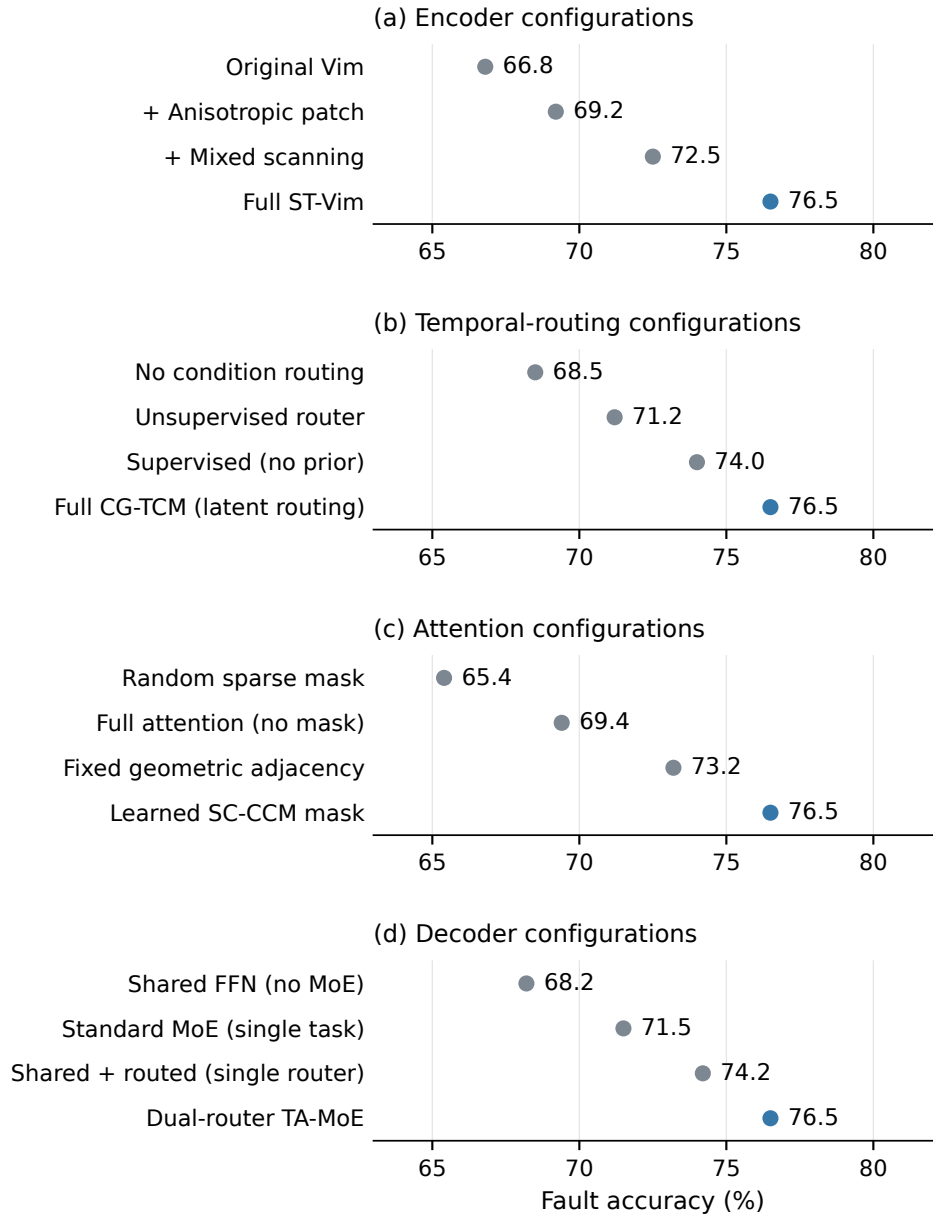


Figure 8. Configuration comparisons for Task 2 (coupled fault diagnosis) under the dual 16-class setting: (**a**) ST-Vim encoder, (**b**) temporal routing, (**c**) attention, and (**d**) decoder configurations. Blue markers denote the full MM-MoE configuration, and grey markers denote the alternatives. The full CG-TCM uses latent routing; condition-label supervision is used only by the supervised comparator in (**b**). Point values; see Section 4.5.3 for their interpretation.

4.6.1. Coupling-Prior-Guided Attention Patterns

TMCA precedes the task-specific routers, so its attention (Equation (11)) is shared by both tasks for a given input. The two panels of Figure 9 are therefore attention summaries rather than separate task-specific attention maps. Figure 9a contains stronger weights in several near-diagonal blocks, whereas Figure 9b contains more scattered high-weight entries. The channel indices do not encode physical distance, and M is a learned task-adaptive prior. These patterns are therefore not evidence of microphone adjacency, propagation paths or suppression of noise-dominated channels. Such interpretations would require comparison with measured geometry or channel-level SNR.

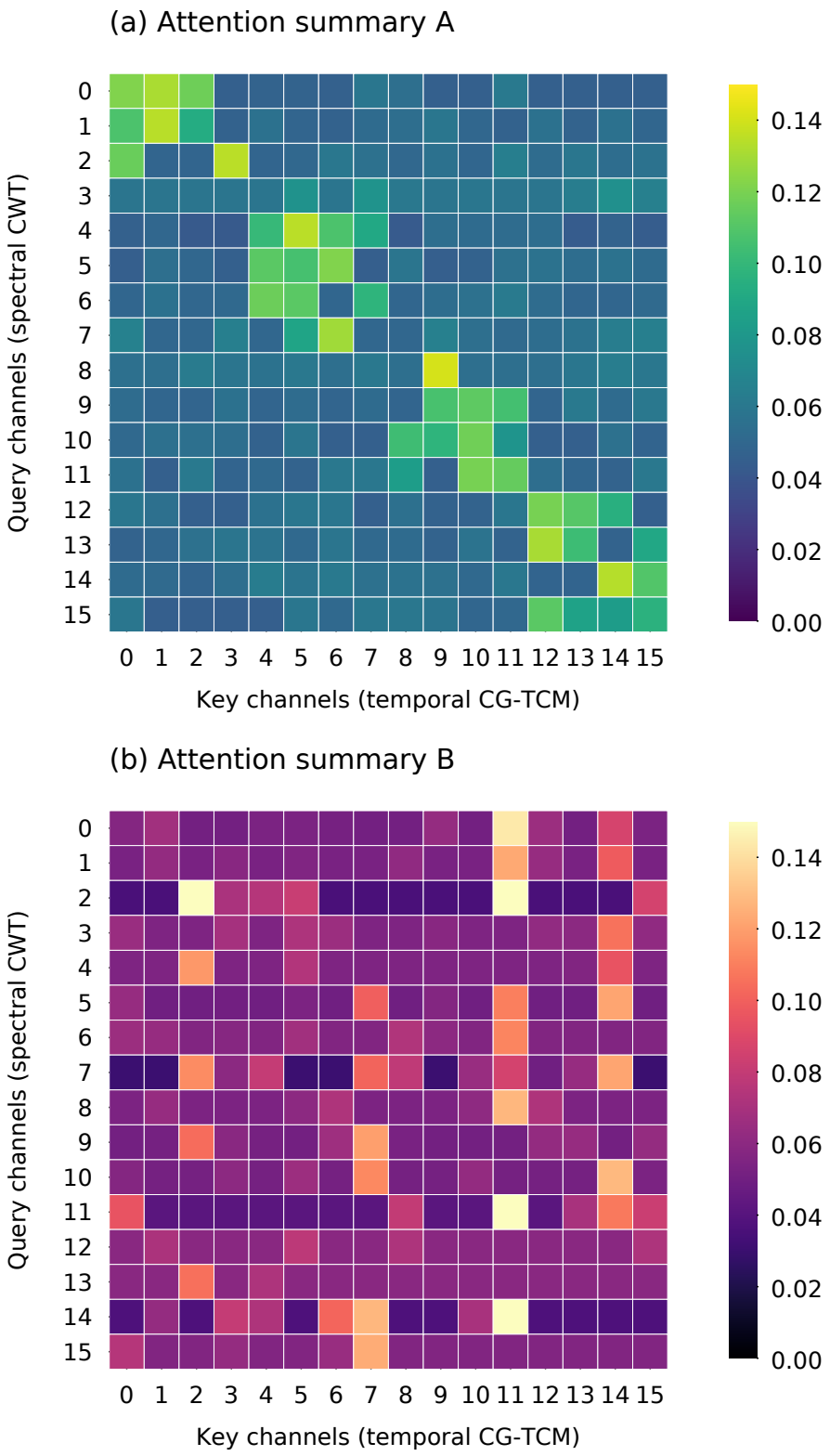


Figure 9. Attention summaries of TMCA. Queries are the 16 channel-aligned spectral embeddings constructed from four prototype TF features; keys and values are derived from the waveform branch. Because TMCA precedes the task routers, panels (**a**,**b**) are not separate task-specific attention computations.

4.6.2. Task-Specific Representations After TA-MoE

Figure 10 shows t-SNE projections of the shared representation Z and the routed representations H_{scene} and $H_{diagnosis}$, colored by the corresponding task labels. The projections

use perplexity 30, PCA initialization, an automatic learning rate and 1000 iterations on up to 3000 test windows sampled by joint label, with seeds 0–4. The routed panels appear more separated, but t-SNE geometry depends on the projection settings. We therefore quantified class separation in the original feature space, which does not depend on the t-SNE seed (Table 6). Standardization, a linear probe ($C = 1$) and a 10-nearest-neighbor (10-NN) classifier were fitted on training features and evaluated on held-out test recordings. Cosine silhouettes were computed in the original coordinates. As a control for proximity to the classifier, we include the same-depth representation of the shared-FFN variant V6.

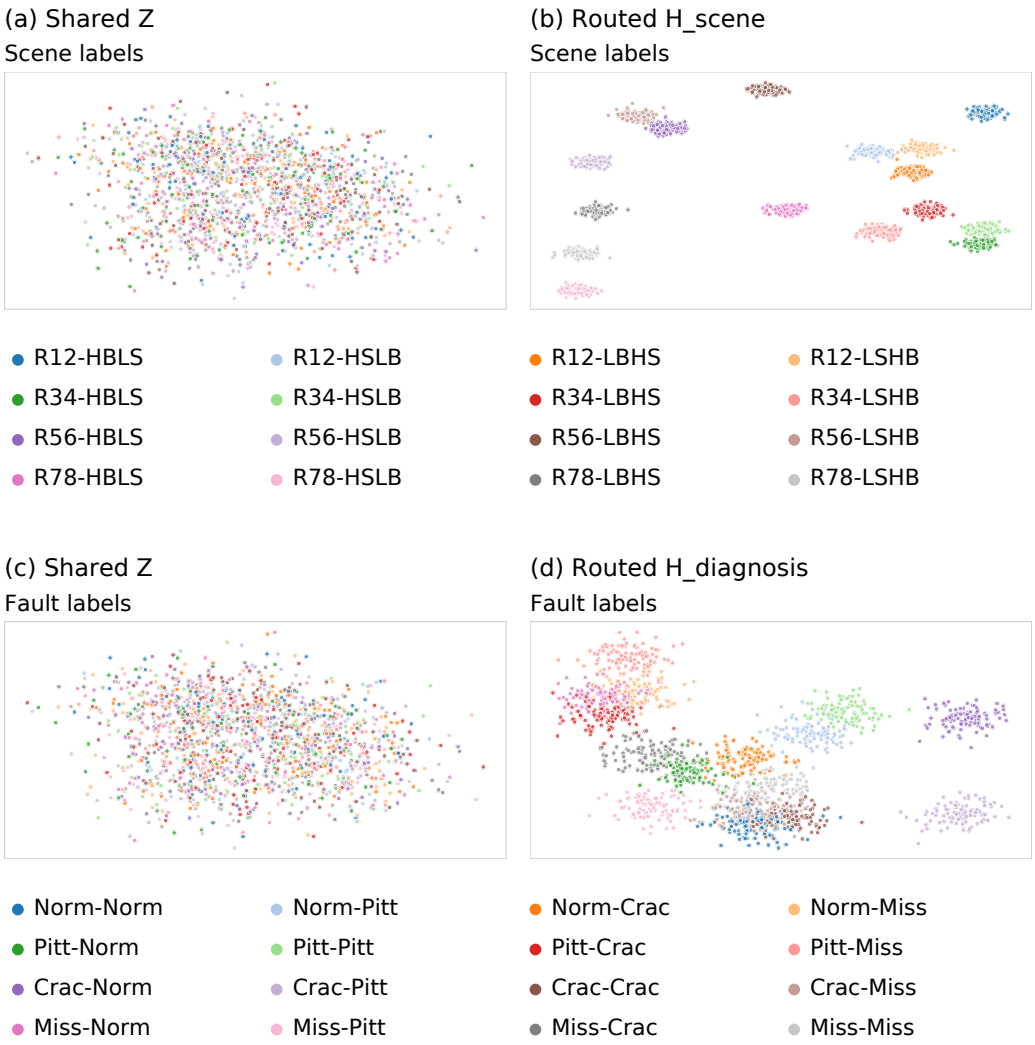


Figure 10. t-SNE views of shared and routed representations. Panels (**a**,**c**) show Z colored by scene and fault labels, respectively; panels (**b**,**d**) show H_{scene} and $H_{diagnosis}$ colored by their own task labels (16 classes each). Quantitative separation is reported in Table 6.

Own-task linear-probe accuracy is 91.0% for H_{scene} and 80.2% for $H_{diagnosis}$, compared with 85.0% and 72.5% for Z and 86.5% and 74.1% for V6. The own-task silhouettes (0.34 and 0.27) also exceed those of Z (0.18 and 0.12) and V6 (0.23 and 0.16). Cross-task information is much less linearly accessible after routing: fault accuracy decreases from 72.5% on Z to 24.8% on H_{scene} and scene accuracy from 85.0% to 30.4% on $H_{diagnosis}$. The routed representations are thus more task-specific than both the shared features and a same-depth representation without routing. Because they feed the task heads directly, we interpret this as task-specific class separation, not as evidence of physically distinct acoustic subspaces or of reduced gradient conflict.

Table 6. Readout accuracy and cosine silhouette of the shared, routed and same-depth V6 representations. Probes are fitted on training features and evaluated on test recordings.

Representation	Label	Linear Probe (%)	10-NN (%)	Silhouette
Shared Z	Scene	85.0	82.6	0.18
Shared Z	Fault	72.5	69.8	0.12
H_{scene}	Scene	91.0	89.2	0.34
H_{scene}	Fault	24.8	22.9	0.03
$H_{diagnosis}$	Scene	30.4	28.1	0.05
$H_{diagnosis}$	Fault	80.2	77.9	0.27
V6 shared FFN	Scene	86.5	84.2	0.23
V6 shared FFN	Fault	74.1	71.6	0.16

4.7. Scope of Validation and External Validity

All results were obtained on the controlled replay dataset of Section 4.1. Replay allows the same recorded source signals to be presented at prescribed positions, heights and levels, independently of the gearbox operating state. Reproducing these configurations with operating gearboxes would require several installed units under controlled mechanical conditions. The file-level split separates recording files, but all partitions share the configuration space, replay apparatus, array and room. Segment-level separation also does not make the parent recordings or fault specimens independent (Appendix A.1). The following limitations apply.

Replayed sources versus operating machinery. A directional loudspeaker differs from a gearbox radiating through its housing and attached structures, which has spatial extent and frequency-dependent directivity. Replay does not reconstruct structure-borne transmission through supports and foundations or re-radiation by surrounding structures. The replayed signals also pass through the original recording chain and the loudspeaker response, and the playback levels do not reproduce operating sound-pressure levels. Such differences can alter the inter-channel cues used by Task 1 and the spectral and transient signatures used by Task 2. The source recordings were made at a single speed (1500 rpm) and load (10 N m), and machined defects do not represent the full variability of naturally developed damage.

Interference from operating machinery. The competing source is a second replayed gearbox recording. Plant interference also contains broadband, tonal and impulsive components from motors, fans, pumps and impacts, and machines sharing a support can be mechanically coupled. The masking results (Section 4.3.2) therefore do not quantify robustness to heterogeneous plant noise or mechanically coupled interference.

Reverberation. The semi-anechoic chamber with a reflective floor is not representative of reverberant plants. Additional reflections would change inter-channel timing and phase and smear transients, and the channel relationships learned by SC-CCM may differ between rooms.

Unknown and simultaneous sources. Both tasks are closed-set: two known sources in one of four region pairs, each with one of four recorded health states. The model has no unknown-class rejection or source-count estimation, so additional sources, unseen fault types or multiple faults within one gearbox are outside the validated scope.

Environmental variability and hardware. Array placement and orientation, microphone response, channel availability and background noise were not varied, and channel failures or reduced channel counts were not evaluated. The four-prototype result concerns compression within the image branch only. The waveform branch still requires all 16 synchronized channels, with an aperture of about 0.6 m, calibration and a 10 kHz sampling rate (5 kHz

Nyquist limit). Task 1 labels are defined relative to this installation and would need to be redefined for a new one.

Optimization sensitivity. The results use fixed loss weights $(\lambda_1, \lambda_2, \lambda_3) = (1, 1, 0.1)$, and their sensitivity was not studied. The routing statistics in Section 4.4 characterize the final checkpoints at this setting.

In summary, all results concern controlled loudspeaker replay in one chamber, not operating industrial machinery, and the applications discussed in Section 5 are potential rather than demonstrated. External validation requires independently acquired operating-machinery data; open-set evaluation with unknown or additional sources; and tests across equipment, operating conditions, array configurations and sites. The region-holdout experiment (Section 4.3.3) addresses transfer to excluded source positions within the present setup only.

5. Conclusions

This study presented MM-MoE, a multimodal mixture-of-experts framework with physically motivated structural priors for joint scene parsing and coupled planetary-gearbox fault diagnosis from 16-channel acoustic signals. It combines prototype-level time–frequency encoding, waveform modeling, a learned channel-coupling prior that biases cross-modal attention, and task-specific expert routing.

All results were obtained with controlled loudspeaker replay of test-bench gearbox recordings in one semi-anechoic chamber, not on operating industrial machinery. Under these conditions, MM-MoE achieved scene and fault accuracies of 89.4% and 76.5% and a JEM of 69.2%. This is 13.4 percentage points above the strongest evaluated baseline (95% Welch CI: 12.4–14.4; five training runs). Aggregating the windows of each recording raised JEM to 87.1%. On source regions excluded from training, ordered fault accuracy was 68.3%, compared with 75.5% on seen regions. Four prototypes required 25% of the per-channel CWT operations and 40% of its latency without a significant loss in JEM. The largest ablation gains came from the coupling prior with TMCA and from TA-MoE. The learned coupling matrix is a task-adaptive prior, not a measured propagation topology.

Because the validation is limited to controlled replay, practical use remains a potential application rather than a demonstrated one. Candidate settings are fixed installations with stable machine and array layouts where contact sensors are difficult to mount. Such use would require validation on operating machinery, installation-specific scene definitions, labeled recordings of the relevant health states, alarm criteria and unknown-fault rejection. The method also needs 16 synchronized channels, and its 31.8 ms latency per 409.6 ms window was measured on the experimental GPU only. Future work should therefore evaluate the framework on operating machinery, under reverberant and heterogeneous industrial interference, and in open-set and cross-site settings (Section 4.7).

Author Contributions: Conceptualization, X.L. and Y.W.; methodology, X.L.; software, X.L.; validation, X.L. and Y.W.; formal analysis, X.L.; investigation, X.L.; resources, X.L.; data curation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, Y.W. and M.W.; visualization, X.L.; supervision, M.W.; project administration, M.W.; funding acquisition, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Provincial Open Fund of Key Laboratory, Sichuan University, under Grant No. 2024LF3016(2024RZY01)-1.

Data Availability Statement: The experimental data and source code supporting the findings of this study are available from the corresponding author upon reasonable request. They are not publicly available because they form part of an ongoing research project.

Acknowledgments: The authors thank Sichuan University for providing an excellent academic platform and Haocheng Yin for guidance and assistance in writing.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in study design, data collection, analysis or interpretation, manuscript preparation, or the decision to publish the results.

Appendix A. Dataset Construction and Experimental Protocol

This appendix describes the dataset construction, data split, evaluation protocol and implementation. The dataset was generated in two stages: real acoustic recordings of four gearbox health states were first collected on a test bench (Appendix A.1) and then replayed by two loudspeakers at controlled positions, heights and sound-pressure levels while a 16-channel microphone array recorded the mixed acoustic field.

Appendix A.1. Test-Bench Source Recordings and Replay Provenance

The source recordings were acquired on a test bench comprising a three-phase induction motor, a single-stage planetary gearbox, and a magnetic powder brake (Figure A1). The stage has a 28-tooth sun gear, four 36-tooth planet gears, and a 100-tooth ring gear. The ring is fixed, the sun is the input, and the carrier is the output, giving a reduction ratio of $(28 + 100)/28 = 4.5714$. All source recordings were made at a constant input speed of 1500 rpm and a constant load of 10 N m.

Let $f_{\text{sun}} = 1500/60 = 25$ Hz, $Z_s = 28$, $Z_r = 100$, and $N_p = 4$. The ideal kinematic carrier and mesh frequencies are

$$f_{\text{car}} = f_{\text{sun}} \frac{Z_s}{Z_s + Z_r} = 5.46875 \text{ Hz}, \qquad f_{\text{mesh}} = Z_r f_{\text{car}} = 546.875 \text{ Hz}.$$

The contact-repetition frequency of a localized sun-gear defect is $N_p f_{\text{mesh}}/Z_s = 78.125$ Hz. Mesh orders 1–9 (up to 4921.875 Hz) lie below the 5 kHz Nyquist frequency of the array acquisition; these are theoretical values derived from the gear geometry and input speed.

The four source-level health states are normal condition, pitting, crack, and missing tooth. The three fault states were produced on interchangeable sun gears in the same gearbox, with realignment after each reassembly. Pitting was represented by electrical-discharge-machined pits on one tooth flank, the crack by a wire-EDM slot at the tooth root, and the missing-tooth condition by removal of one complete tooth. Table A1 gives the dimensions. The labels were assigned from the prepared specimens, and the gears were inspected before and after recording. Figure A2 shows the three fault specimens.

Sound was recorded with a G.R.A.S. Type 40BE microphone (GRAS Sound & Vibration A/S, Holte, Denmark) placed 0.5 m from the gearbox housing at a sampling rate of 25.6 kHz. Each health state was recorded in six separate 60 s runs, with the bench stopped and restarted between runs, giving 24 source recordings in total. The repeated runs are separate acquisitions on the same gearbox and gear specimen, not independent specimen replicates.

For replay, the source recordings were divided into non-overlapping 10 s segments. The five replay files of each configuration used different source segments, no source segment was used in more than one data partition, and when both sources had the same health state, the two loudspeakers played segments from different recordings of that state. An original bench recording (60 s at 25.6 kHz), a replay acquisition file (10 s, 100,000 samples per channel at 10 kHz) and a model-input window (4096 samples) are therefore different units, and the five replay files per configuration are distinct from both the six bench runs per health state and the five training runs used for statistical reporting.

Table A1. Health states, fault preparation, and original test-bench recordings. EDM denotes electrical discharge machining. The six recordings for each state are separate stopped-and-restarted runs on the same gearbox and corresponding gear specimen.

Health State	Faulty Part	Preparation	Fault Dimensions	Separate Runs	Duration per Run
Normal	None	Unmodified reference	Not applicable	6	60 s
Pitting	Sun gear	EDM pits on one tooth flank	3 pits; diameter 1.0 mm; depth 0.5 mm	6	60 s
Crack	Sun gear	Wire-EDM slot at tooth root	Depth 2.0 mm; width 0.2 mm	6	60 s
Missing tooth	Sun gear	Wire-EDM removal	One complete tooth	6	60 s

Appendix A.2. Acoustic Replay Environment

The complete acquisition scene is shown in Figure A1. The setup was deployed in a large semi-anechoic industrial chamber to reduce external noise and uncontrolled wall reflections. Two source heights were used to introduce vertical acoustic variation. The low-position source was placed directly on the reflective floor, and the high-position source was placed 1.5 m above the floor on a rigid platform. The microphone array was kept at a fixed distance and a fixed center height of 1.0 m.

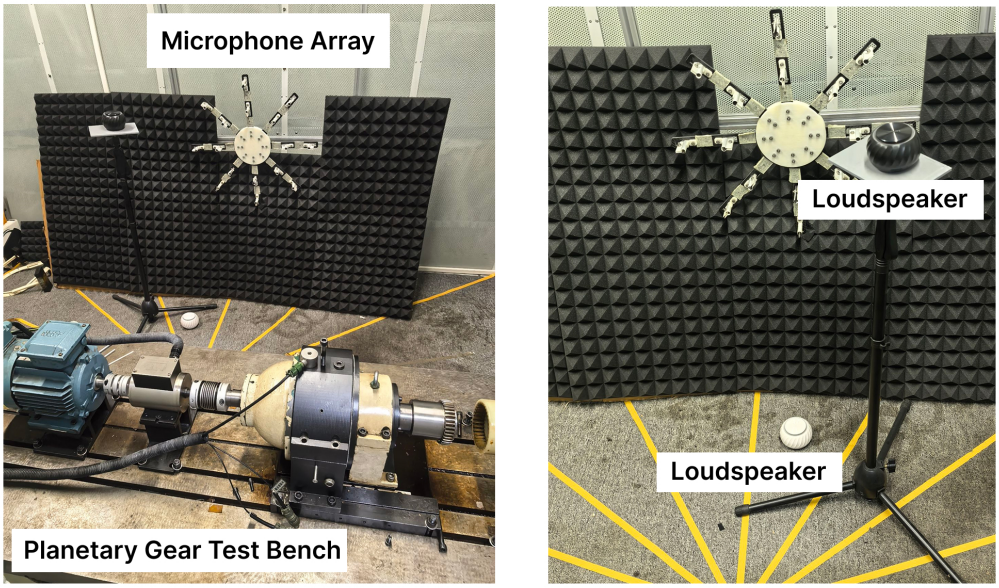


Figure A1. Overview of the two-stage dataset construction process. Real planetary-gearbox fault sounds were first recorded on a test bench and then replayed by two loudspeakers at controlled positions, heights, and sound-pressure levels during microphone-array acquisition.

Appendix A.3. Microphone Array and Acoustic Sources

The acquisition system used a rigid custom 16-channel planar circular array with G.R.A.S. Type 40BE microphone capsules and a maximum radius of 30 cm. A wideband directional loudspeaker system with a nominal frequency range of 0.1–20 kHz was used for replay. Strong and weak source levels were produced by adjusting the loudspeaker driving voltage, giving a 20 dB playback range.

The 60 dB and 40 dB labels in Figure 4 denote these high and low playback-level settings; they are not sound-pressure levels measured at the array. Because the interferer

is a second gearbox source, masking is described by the target-to-interferer ratio at a receiver, SIR $= 10\log_{10}(q_t/q_i)$, which is distinct from the target-to-background-noise ratio SNR $= 10\log_{10}(q_t/q_b)$; q_t, q_i and q_b are the target, interferer and background mean-square pressures over the same bandwidth and interval, evaluated before z-score standardization. For the weak-target condition (−20 dB playback contrast), inverse-distance spreading gives a nominal SIR at the array of −26.0, −20.0 and −14.0 dB for interferer-to-target distance ratios of 0.5, 1 and 2. These are calculated values: source spectra, heights, loudspeaker directivity and floor reflections also affect the received levels, and the received SIR was not measured separately.

The specimens associated with the three fault labels are illustrated in Figure A2; the normal condition provides the fourth source-level class. These recordings were treated as real gearbox fault acoustic sources. The dataset combines 4 replayed source-pair regions, 4 height–intensity source configurations, and 16 joint fault-state combinations. Their Cartesian product gives $4 \times 4 \times 16 = 256$ experimental configurations.

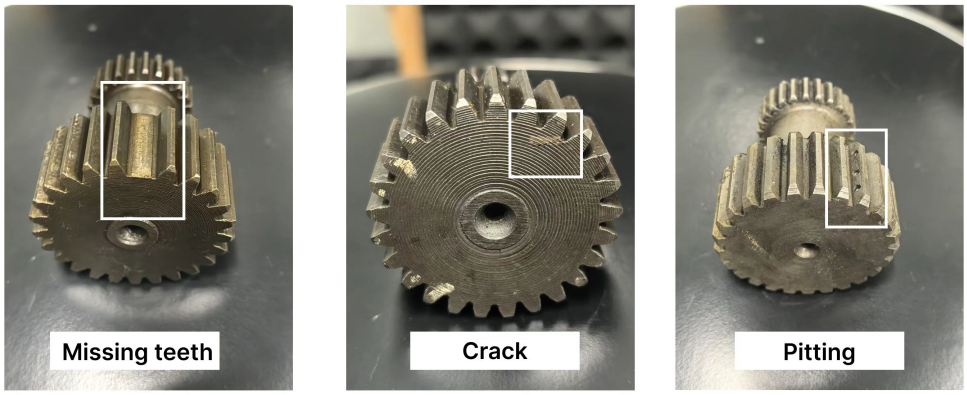


Figure A2. Photographs of the gear specimens associated with the missing-tooth, crack, and pitting source labels, shown from left to right.

Appendix A.4. Dataset Hierarchy and Label Definition

Each of the 256 configurations was stored as one raw-data directory rather than as one training sample. Each directory contains five raw recording files. The hierarchy is defined by three factors:

1. **Replayed source-pair region:** Four source-pair regions, denoted as Regions 1–2, 3–4, 5–6, and 7–8.
2. **Source-level condition metadata:** Each replayed source has height and sound-pressure-level metadata, denoted as HB, HS, LB, or LS. These metadata describe the dataset hierarchy and are not used as auxiliary labels for the CG-TCM router.
3. **Joint fault state:** Each replayed source belongs to one of four health states. The two-source combination gives $4 \times 4 = 16$ coupled fault classes.

The total number of raw-data directories is therefore

$$N_{dir} = 4_{\text{pairs}} \times 4_{\text{cond}} \times 16_{\text{states}} = 256. \quad \text{(A1)}$$

The multi-task label for each segmented sample is $Y = \{y_{scene}, y_{diagnosis}\}$. Task 1, spatial-energy scene parsing, identifies the source-pair region and pair-level height–intensity configuration. Task 2, coupled fault diagnosis, identifies the joint health state of the two replayed gearbox fault recordings. The dataset is balanced across both task dimensions: each replayed source-pair region under every operating condition contains 1920 segmented samples, and each of the 16 joint fault classes contains 480 segmented samples per condition.

Appendix A.5. Segmentation and Data Split

Each raw recording file contains 100,000 sampling points per channel. We use non-overlapping windows of length $L = 4096$, yielding

$$N_{seg} = \left\lfloor \frac{100,000}{4096} \right\rfloor = 24 \tag{A2}$$

windows per raw recording file. The remaining 1696 points are discarded. With five raw files per configuration, the final dataset contains $256 \times 5 \times 24 = 30,720$ segmented samples.

The split was performed at the raw-file level within each configuration, so all windows of a replay file stay in one partition, and the source segments are disjoint across partitions (Appendix A.1). The first three raw files were used for training, the fourth for validation, and the fifth for testing (Table A2). Z-score standardization used statistics computed only from the training set.

Table A2. Dataset partitioning under the chronological file-level split.

Subset	Raw Files per Configuration	Windows per Recording	Segmented Samples	Ratio
Training set	3	24	18,432	60%
Validation set	1	24	6144	20%
Test set	1	24	6144	20%
Total	5	24	30,720	100%

Appendix A.6. Evaluation Metrics

The two 16-class tasks are evaluated using Top-1 accuracy and Macro-F1. For a task with $C = 16$ classes, Macro-F1 is defined as

$$Macro\text{-}F1 = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c}. \tag{A3}$$

Joint Exact Match (JEM) measures whether both task labels are correct for the same sample:

$$JEM = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(y^{(i)}_{scene} = \hat{y}^{(i)}_{scene} \wedge y^{(i)}_{diagnosis} = \hat{y}^{(i)}_{diagnosis}\right), \tag{A4}$$

where N is the number of test samples, $\mathbb{I}(\cdot)$ is the indicator function, $y^{(i)}$ denotes the ground-truth label, and $\hat{y}^{(i)}$ denotes the predicted label.

Repeated runs and statistical reporting.

Each result reported with $\pm$ is the mean and sample standard deviation (divisor $R-1$) over $R = 5$ independent training runs with different random seeds on the same file-level split; each metric is computed on the full test set (6144 windows from 256 recordings) separately for each run. Accuracy and JEM are in percent, and their standard deviations in percentage points. Macro-F1 uses the fixed set of 16 classes of each task and is reported as a point value on the 0–1 scale. The training runs quantify training variability conditional on the fixed dataset and split; windows from one recording are correlated subsamples, which is addressed by the recording-level analyses below.

Confidence intervals and between-model tests.

Two-sided 95% confidence intervals for run-level means are $\bar{m} \pm t_{0.975,4}\, s/\sqrt{5}$ with $t_{0.975,4} = 2.7764$ (Python 3.9.6, SciPy 1.13.1). Training runs were independent and unpaired across methods, so differences were tested with two-sided Welch tests, $t = (\bar{m}_a - \bar{m}_b)/\sqrt{s_a^2/5 + s_b^2/5}$ with Welch–Satterthwaite degrees of freedom. The test family comprises MM-MoE against the eight baselines on scene accuracy, fault accuracy and JEM (24 tests) and the default $K = 4$ encoding against $K = 8$ and per-channel CWT on the same

metrics (6 tests); Holm correction controls the family-wise error rate over all 30 tests at $\alpha = 0.05$. Difference intervals are unadjusted, and significance decisions use the adjusted p values (p_H). All results are reported in Appendix E.

Appendix A.6.1. Recording-Level Evaluation and Joint Uncertainty

Recording-level aggregation.

For training run r, test recording j, task t, class c and chronological window w, let p_{rjtwc} be the predicted class probability. Soft voting averages the probabilities of the first k windows,

$$\bar{p}^{(k)}_{rjtc} = \frac{1}{k}\sum_{w=1}^{k} p_{rjtwc}, \qquad \hat{y}^{(k)}_{rjt} = \arg\max_{c} \bar{p}^{(k)}_{rjtc},$$

with $k = 24$ for full recordings (9.83 s of signal). Recording-level JEM requires both aggregated task labels of a recording to be correct. Aggregation is performed separately for each trained model, without ensembling across runs. Exact probability ties are broken by the smallest class index. For the observation-duration analysis, the first $k \in \{1, 2, 3, 4, 6, 8, 12, 24\}$ consecutive windows ($0.4096k$ s) are aggregated. Table A3 reports the results; with only one test file per configuration, residual errors cannot establish systematic failure of a configuration.

Recording-composition uncertainty.

The 256 test-recording IDs are resampled with replacement 10,000 times, using the same sampled IDs for all methods and retaining all windows and all five runs of each sampled recording; metrics are computed per run and averaged. The 2.5th and 97.5th percentiles give 95% intervals conditional on the five trained models.

Combined run and recording uncertainty.

To account for both sources of variation, a crossed bootstrap (10,000 replicates, seed 20261012) independently resamples the five runs within each method and resamples the 256 recording IDs, shared across methods, keeping the 24 windows of each recording together. Basic bootstrap intervals are formed from the centered deviations, $[\hat{\theta} - q_{0.975}(e), \hat{\theta} - q_{0.025}(e)]$ with $e_b = \theta^*_b - \hat{\theta}$, and method differences are computed within each common recording resample (Table A4). For MM-MoE minus MC-ABDS, the JEM difference is 13.4 points (95% CI 10.5–16.4) at the window level and 13.3 points (9.0–17.7) at the recording level, both with Holm-adjusted $p = 0.006$. The recording bootstrap treats recording files as exchangeable; reuse of parent source recordings may introduce additional dependence.

Equivalence sensitivity.

Post hoc two one-sided tests (TOST) with margins of ± 1 and ± 2 JEM points give 90% Welch intervals of -0.81 to 0.61 points for $K = 4$ minus $K = 8$ ($p = 0.023$ and 0.0005) and -0.91 to 0.51 points for $K = 4$ minus per-channel CWT ($p = 0.034$ and 0.0007). Because the margins were not specified in advance, we do not claim practical equivalence.

Table A3. Window- and recording-level performance (soft voting over 24 windows). Intervals resample test recordings conditional on the five trained models.

Model	Window Scene/Fault/JEM (%)	Recording Scene/Fault/JEM (%)	Recording JEM 95% CI (%)
DUET	74.5/58.7/45.2	84.0/70.3/61.7	59.1–64.3
Standalone Plain CA	76.2/68.5/54.1	85.5/80.5/71.9	69.5–74.2
MC-ABDS	78.5/67.4/55.8	88.7/79.7/73.8	71.5–76.2
MM-MoE	89.4/76.5/69.2	96.1/89.8/87.1	85.0–89.1

Table A4. JEM with crossed run-by-recording bootstrap intervals. SD is over five training runs.

Model	Window JEM ± SD (%)	Window 95% CI (%)	Recording JEM ± SD (%)	Recording 95% CI (%)
DUET	45.2 ± 1.1	43.0–47.4	61.7 ± 0.6	58.1–65.3
Standalone Plain CA	54.1 ± 0.9	52.0–56.3	71.9 ± 0.6	68.7–75.2
MC-ABDS	55.8 ± 0.8	53.7–57.9	73.8 ± 0.6	70.5–77.1
MM-MoE	69.2 ± 0.6	67.4–71.1	87.1 ± 0.6	84.6–89.8

Appendix A.6.2. Routing Statistics

Routing statistics are computed within each trained model from the per-window softmax probabilities before Top-3 selection and the selected expert sets, on the same 2048 test windows for every run, and then summarized as mean and sample SD over five runs. For a subset of n windows, let $p_{wj}^{(t)}$ be the pre-selection probability of expert j and $S_w^{(t)}$ the selected set for task t, with $E = 8$ experts and $k = 3$ selections. We distinguish the mean soft probability, the selection frequency and the normalized slot share,

$$\bar{p}_j^{(t)} = \frac{1}{n}\sum_w p_{wj}^{(t)}, \qquad f_j^{(t)} = \frac{1}{n}\sum_w 1\{j \in S_w^{(t)}\}, \qquad u_j^{(t)} = \frac{f_j^{(t)}}{k},$$

so that $\sum_j u_j = 1$. The normalized load entropy is $H(u)/\log E$ (1 for uniform use), the maximum slot share is $\max_j u_j$ (uniform value $1/E = 12.5\%$), and an expert with $u_j < 0.01$ is counted as low load. Input dependence is described by the soft-probability information $I_p = H(\bar{p}) - n^{-1}\sum_w H(p_w)$ and the selection information $I_{\text{sel}} = H(u) - \log k$, together with the number of distinct selected sets and the fraction of windows using the most frequent set. Task differences are measured by the Jensen–Shannon divergence $J(a,b) = H((a+b)/2) - [H(a) + H(b)]/2$ (natural logarithms, in nats) between the two routers' soft profiles and between their slot-share profiles, using the common expert indexing of each model. All statistics are invariant to a common relabeling of the experts, which makes them comparable across independently trained runs. They are also computed for the four source-region groups and the two masking conditions. Table A5 reports the per-router results; the Jensen–Shannon divergence between the two routers is 0.2165 ± 0.0021 nats for the soft profiles and 0.3048 ± 0.0082 nats for the slot shares. Across the four source-region groups, load entropy ranges from 0.807 to 0.818 (scene) and from 0.855 to 0.867 (diagnosis), and the two masking conditions give 0.803/0.822 and 0.855/0.867. For the four CG-TCM temporal experts, which use dense soft routing, the soft-load entropy is 0.9918 ± 0.0008. The aggregate profile in Figure 6 has normalized entropies of 0.763 (scene) and 0.852 (diagnosis), a Jensen–Shannon divergence of 0.351 nats and Top-3 masses of 0.820 and 0.700.

Table A5. Routing statistics of the two TA-MoE routers over five training runs (2048 test windows per run; mean ± sample SD).

Metric	Scene Mean ± SD	Diagnosis Mean ± SD
Normalized load entropy	0.8129 ± 0.0048	0.8614 ± 0.0036
Maximum slot share	0.3123 ± 0.0016	0.2683 ± 0.0033
Maximum selection frequency	0.9370 ± 0.0049	0.8050 ± 0.0100
Low-load experts (<1% slots)	0	0
Soft-probability information (nats)	0.2469 ± 0.0029	0.2536 ± 0.0021
Selection information (nats)	0.5918 ± 0.0099	0.6925 ± 0.0075
Modal selected-set fraction	0.4147 ± 0.0100	0.2502 ± 0.0132
Distinct selected sets	41.4 ± 1.3	48.2 ± 1.6

Appendix A.7. Implementation Details

MM-MoE was implemented in PyTorch 2.1 (Python 3.9.6) and trained on one 80 GB NVIDIA A100 GPU (NVIDIA Corporation, Santa Clara, CA, USA). The TG-LR-CWT branch used $K = 4$ acoustic prototypes by default. Each prototype TF map was resized to 224 × 224 × 3 and processed by a shared ST-Vim encoder with anisotropic patch size $P_h = 32, P_w = 4$. Compared with full channel-wise CWT over $N = 16$ channels, TG-LR-CWT therefore required $K/N = 25\%$ of the CWT operations.

The network was optimized with AdamW, a weight decay of 1×10^{-4}, an initial learning rate of 1×10^{-4}, and cosine annealing to 1×10^{-6} over 100 epochs. The batch size was 64. Early stopping used validation JEM with a patience of 15 epochs. The loss weights were $\lambda_1 = 1.0$ for Task 1 cross-entropy, $\lambda_2 = 1.0$ for Task 2 cross-entropy, and $\lambda_3 = 0.1$ for the TA-MoE auxiliary load-balancing loss. No coefficient was used for condition-router supervision because CG-TCM was trained as a latent signal-adaptive router. TA-MoE used eight routed experts, with Top-K gating activating three experts per task-specific forward pass.

Model configuration.

The embedding width is $d = 256$, and the spectral-factor width is $r = 64$. For a 224 × 224 × 3 map, a 32 × 4 patch contains 384 values; strides $(16, 2)$, zero padding and unit dilation give $13 \times 111 = 1443$ tokens per map (5772 over four maps), each projected from 384 to 256 features, and mean pooling over the encoded tokens gives each prototype vector. The channel and prototype lifting embeddings are 16 × 256 and 4 × 256, with $W_a \in \mathbb{R}^{256\times256}$; the trainable scalars η and γ are initialized at 1 and unconstrained. Each temporal expert is a biased 4096 → 256 linear map, and the channelwise router is a 4096 → 64 → 4 MLP with GELU that produces dense softmax weights. TMCA uses one attention head of width 256 (score divisor 16) and $\epsilon = 10^{-6}$. The decoder is specified in Table A6.

Table A6. TA-MoE decoder and readout specification.

Field	Value or Operation
Expert bank	2 shared and 8 routed experts; each a tokenwise 256 → 512 → 256 MLP
Expert activation	GELU, then dropout 0.1 between the two linear layers
Router input	Mean of Z over the 16 channel rows (width 256)
Task routers	Two independent 256 → 8 linear maps followed by softmax
Selection	Top-3 per task and input window; ties broken by the smaller index
Selected weights	$p_{t,j} / \sum_{l\in\mathcal{K}_t} p_{t,l}$, summing to one
Expert aggregation	Shared experts with coefficient 1 plus weighted routed experts; 16 × 256
Task pooling and heads	Channel mean 16 × 256 → 256; two separate 256 → 16 linear heads
Auxiliary loss	Defined in Appendix C.4; summed over the two routers
Loss weights	$(\lambda_1, \lambda_2, \lambda_3) = (1, 1, 0.1)$

Appendix A.8. Computational Profiling and Training Cost

Table 4 reports model size, FLOPs, inference and peak GPU memory, and end-to-end latency for one 16-channel window with $L = 4096$; latency includes time–frequency preprocessing, where applicable, and network inference. Table A7 gives the per-epoch training time and peak GPU memory, measured with PyTorch 2.1 on one 80 GB NVIDIA A100 GPU at batch size 64. Absolute runtimes depend on hardware, CWT implementation, batch size and runtime optimization.

Table A7. Reported training time and peak GPU memory for the computational comparison.

Method	Training Time/Epoch (min)	Peak GPU Memory (GB)
1D-CNN (Time)	0.42	0.9
DUET (Time)	0.65	1.4
Vanilla Vim (Image)	1.84	5.6
Standalone Plain CA	2.54	7.2
MC-ABDS (MTL)	2.31	6.8
MM-MoE ($K = 4$)	3.12	8.4
MM-MoE + per-channel CWT	8.27	19.7

Appendix A.9. Baseline Implementation and Hyperparameter Tuning

Common conditions.

All methods use the chronological split in Table A2, the same 16-channel, 4096-sample windows, the two 16-class targets of Section 4.1 and training-only standardization. Each deep model receives the same number of tuning trials over the same search space, which equalizes search effort but not model capacity or computation (Table 4); for example, MM-MoE and Standalone Plain CA have similar active parameter counts (10.92 and 10.86 M) but different total counts (13.76 and 10.86 M) and FLOPs (6.38 and 5.12 G).

Published methods and their adaptation.

DUET [24] was introduced for multivariate time-series forecasting, using temporal clustering, frequency-domain channel clustering and masked channel fusion. We include it because these two clustering mechanisms underlie CG-TCM and SC-CCM (Section 3.3); the DUET entry in Table 3 denotes our classification adaptation described below, not a result of the original study. The temporal-only ablation variant V1 has the same scene and fault accuracies as DUET (74.5% and 58.7%) but is a separate configuration with its own JEM and parameter count. MC-ABDS [3] uses its published backbone with the classification interface below. Standalone Plain CA is distinct from the internal ablation variant V5, which uses the internal branches and shared FFN heads (Section 4.5.1).

Hyperparameter tuning.

The validation-selection procedure used for each deep model comprises 30 completed validation-only TPE trials, with failed trials logged separately, a search seed of 0, at most 50 epochs per trial and a validation-JEM patience of 10. Ties in validation JEM are broken by lower validation cross-entropy and then earlier trial number. Learning rate is searched log-uniformly over $[10^{-5}, 3 \times 10^{-3}]$, weight decay over $[10^{-6}, 10^{-2}]$, dropout over $\{0, 0.1, 0.2, 0.3\}$, and batch size over $\{32, 64, 128\}$ (Table A8). The selected configuration is then trained in five independent runs (at most 100 epochs, patience 15). Each trial and final run uses AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$ and optimizer epsilon 10^{-8}; cosine learning-rate decay reaches 10^{-6}. Initialization is from scratch. The held-out test set is used only for final evaluation.

Classification interfaces.

Every model has two heads for the two 16-class labels, with cross-entropy coefficients of one; auxiliary losses are used only by DUET and MM-MoE.

The DUET classification interface retains a B × 16 × 128 fused tensor. Fixed channel-order flattening gives 2048 features, followed by a biased 2048 → 128 projection, GELU and two 128 → 16 heads. Four temporal experts use Top-2 with selected-probability renormalization. Its balancing term is $0.1 \times 4 \sum_{j=1}^{4} f_j \bar{p}_j$, where the routing units are the $16B$ channel series and f_j divides selection counts by $2 \times 16B$. No test labels enter normalization or selection.

For ResNet-50 (Torchvision v1.5) and Vanilla Vim, each of the 16 channels supplies a 224 × 224 × 3 phase-aware CWT map to a shared image encoder. Encoded tokens or spatial positions are mean-pooled within each map, then the 16 channel vectors are averaged before the two classification heads. The MC-ABDS classification interface globally averages non-feature spatial or temporal axes, projects the resulting encoder feature to 256 dimensions, and uses two 256 → 16 heads; its backbone is the cited architecture. The standalone Plain CA baseline combines the specified temporal and prototype-image branches, with eight attention heads of width 32, no coupling bias in attention or lifting, and channel-mean readout.

Table A8. Starting settings for validation-based hyperparameter selection and classical-model grids.

Method	Specified Starting Setting
SVM + Spectral	Flatten 16 × 2049 rFFT magnitudes; RBF kernel; $C = 10$, $\gamma = 0.01$; separate classifiers per task
RF + Wavelet	db4 level-5 wavelet decomposition; energy in 6 bands per channel (96 features); 500 trees, depth 40, feature count $\lfloor\sqrt{96}\rfloor = 9$
1D-CNN	Six length-preserving convolutions, width 64, kernel 15, stride 1, padding 7, GELU; temporal mean; two 64 → 16 heads; learning rate 3×10^{-4}
DUET classification	Width 128, 4 experts, Top-2, 2 layers, 4 heads of width 32, FFN width 256; flatten/project adapter above; learning rate 5×10^{-4}
ResNet-50	Shared encoder, 2048-feature pooled readout per map; channel mean and two 2048 → 16 heads; learning rate 3×10^{-4}
Vanilla Vim	12 layers, width 192; 16 × 16 non-overlapping patches, 196 tokens/map; token and channel means; two 192 → 16 heads; learning rate 2×10^{-4}
Standalone Plain CA	Branch width 256; 1 attention layer, 8 heads of width 32; two 256 → 16 heads; learning rate 2×10^{-4}
MC-ABDS adaptation	Published backbone with the 256-feature classification interface above; learning rate 3×10^{-4}
MM-MoE	Table 2 implementation specification; learning rate 10^{-4}; TA-MoE auxiliary coefficient 0.1

Deep-model starting settings use weight decay 10^{-4}, dropout 0.1 and batch size 64. Dropout is placed on the pooled classifier input for the image baselines, after GELU for 1D-CNN and the DUET adapter, and between expert linear layers for MM-MoE. SVM searches $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$; RF searches trees $\{200, 500, 1000\}$, depth $\{10, 20, 40\}$ and feature counts $\{9, 24, 48\}$. Each classical-model setting is fitted separately for the two tasks; taskwise validation accuracy selects its setting, with earlier grid order breaking ties. Epoch and AdamW settings apply only to deep models.

Appendix B. TG-LR-CWT Preprocessing Details

This appendix gives the mathematical details of the TG-LR-CWT preprocessing pipeline.

Appendix B.1. Low-Rank Channel Projection

Given the raw multi-channel acoustic sample $X \in \mathbb{R}^{N \times L}$, where $N = 16$ and $L = 4096$, the TG-LR-CWT module first applies a learnable low-rank spatial projection:

$$U = PX, \quad P \in \mathbb{R}^{K \times N}, \quad U \in \mathbb{R}^{K \times L}. \tag{A5}$$

In this work, $K = 4$ is used by default. The projected matrix $U = \{u_1, \ldots, u_K\}$ contains four acoustic prototype waveforms that summarize multiple spatial response patterns without reducing the 16-channel signal to a single principal component.

Appendix B.2. Prototype-Level Complex CWT

For each prototype waveform u_k, the complex Continuous Wavelet Transform (CWT) is utilized for multi-resolution time–frequency analysis. Unlike the Short-Time Fourier Transform (STFT), CWT provides high temporal resolution for high-frequency components and high frequency resolution for low-frequency components. The transformation is defined as:

$$C_k(s, \tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} u_k(t) \psi^* \left(\frac{t - \tau}{s} \right) dt \tag{A6}$$

where $C_k(s, \tau)$ denotes the complex wavelet coefficients of the k-th acoustic prototype. The parameter s represents the scale factor (which is inversely proportional to frequency), τ represents the translation parameter (time shift), and $\psi^*(\cdot)$ is the complex conjugate of the mother wavelet function.

Appendix B.3. Phase-Aware TF Construction

The resulting coefficients are converted into phase-aware prototype TF maps:

$$I_k = [\log|C_k|, \cos(\angle C_k), \sin(\angle C_k)], \quad I_k \in \mathbb{R}^{H\times W\times 3}. \tag{A7}$$

This representation retains spectral energy and phase-related information for each acoustic prototype.

Appendix B.4. Shared ST-Vim Encoding

The K prototype TF maps are encoded by a shared ST-Vim encoder:

$$z_k = E_{TF}(I_k), \quad Z_{tf} = [z_1, z_2, \ldots, z_K] \in \mathbb{R}^{K\times d}. \tag{A8}$$

Sharing the ST-Vim encoder avoids replicating its weights across prototypes, although the projection $P \in \mathbb{R}^{K\times 16}$ and the prototype embeddings still grow with K.

Appendix B.5. Topology-Guided Channel Lifting

Finally, the prototype-level TF features are mapped back to channel-aligned spectral embeddings:

$$H_{img} = AZ_{tf} + E_{ch}, \quad H_{img} \in \mathbb{R}^{N\times d}, \tag{A9}$$

where $A \in \mathbb{R}^{N\times K}$ is a topology-guided lifting matrix and $E_{ch} \in \mathbb{R}^{N\times d}$ denotes learnable microphone-channel embeddings. The resulting H_{img} provides the 16 channel-aligned spectral queries used by TMCA without requiring full per-channel CWT computation.

Appendix B.6. Linear Reconstruction Diagnostic for Prototype Compression

To quantify how much multichannel waveform energy the four-prototype projection retains, independently of task accuracy, we fit a linear decoder to the frozen learned projection P. Using the normalized windows that enter the network, the uncentered second-moment matrix C_{tr} of the training windows defines the least-squares decoder

$$C_{tr} = \frac{1}{n_{tr}L}\sum_{b\in tr} X_b X_b^T, \qquad W^* = C_{tr}P^T(PC_{tr}P^T)^\dagger,$$

and held-out windows from disjoint test recordings are reconstructed as $\widehat{X}_b = W^*PX_b$. The reconstruction score is

$$R_{test} = 1 - \frac{\sum_{b\in test}\|X_b - \widehat{X}_b\|_F^2}{\sum_{b\in test}\|X_b\|_F^2},$$

i.e., one minus the normalized squared error; it measures waveform energy, not task information. As a reference, the rank-K uncentered PCA/SVD projector $V_K V_K^T$ fitted on the same training windows is optimal for reconstruction on those windows. Band-resolved scores use the same operators in five bands between 0 and 5 kHz, with Parseval-weighted one-sided rFFT energies.

The diagnostic uses 32 training windows and 16 test windows from eight training and four held-out recordings. The learned projection retains 92.53% of the held-out waveform energy, compared with 92.88% for the training-SVD reference (training scores 92.60% and 92.95%), and the gap is similar in all bands (Table A9). The task-trained projection therefore retains almost as much waveform energy as the optimal rank-4 linear projection. The

diagnostic concerns the waveform projection only, not the nonlinear TF maps or the other information paths of the network.

Table A9. Held-out waveform reconstruction scores (%) of the frozen learned projection and the training-SVD reference ($K = 4$).

Projection	0–0.5 kHz	0.5–1 kHz	1–2 kHz	2–3 kHz	3–5 kHz	All
Learned P	92.49	92.52	92.53	92.49	92.57	92.53
Training SVD	92.84	92.87	92.87	92.84	92.92	92.88

Appendix C. Additional Methodological Details

This appendix records implementation-level details of Section 3.

Appendix C.1. ST-Vim Tokenization and Dual-Axis Scanning

Each prototype TF map I_k is split into anisotropic overlapping patches:

$$T_{in} = \text{Unfold}(I_k, \text{kernel} = (P_h, P_w), \text{stride} = (S_h, S_w))W_{proj} + E_{pos}, \quad \text{(A10)}$$

where $P_h \gg P_w$ favors tall, narrow patches, S_h and S_w denote overlap strides, W_{proj} is the linear projection, and E_{pos} is the positional embedding. The implementation specification in Table 2 fixes $(S_h, S_w) = (16, 2)$, zero padding and unit dilation. Thus $F' = \lfloor (224 - 32)/16 \rfloor + 1 = 13$, $T' = \lfloor (224 - 4)/2 \rfloor + 1 = 111$, and $n_{tok} = 1443$. Each 384-element patch is projected to 256 features. The final encoder output is mean-pooled over the 1443 patch positions to produce $z_k \in \mathbb{R}^{256}$. The design reduces truncation of vertical transient structures while maintaining frequency-local context.

The dual-axis scans and their re-alignment are defined in Section 3.2: both outputs are restored to the common $F' \times T'$ grid before feature concatenation, so the fusion is a per-patch $2d \to d$ projection rather than a transpose of the token–feature matrix.

Appendix C.2. Latent Temporal Routing

CG-TCM follows a channel-independent strategy for temporal routing. Each channel waveform is treated as a separate series so that local channel noise does not force all microphones through the same temporal expert. The router output $G(x_n)$ in Section 3.3 is learned end-to-end. It is not supervised by source height, playback intensity, source-region pair, or any other operating-condition metadata. This design prevents condition labels related to scene parsing from becoming privileged supervision for the diagnosis task.

The temporal expert bank contains $K_t = 4$ linear pattern extractors. Four is a configured expert count, not an inferred number of physical operating regimes, and no separate balancing loss is applied to this router.

Appendix C.3. Spectral-Coupling Matrix

With $R = A_q^T A_q$, $D_{ij} = \|A_q(X_i^{freq} - X_j^{freq})\|_2^2$. Its square root is a pseudometric (a metric only if A_q is injective); neither property implies physical proximity. With $A_q \in \mathbb{R}^{64\times 2049}$, $\text{rank}(R) \leq 64$, so distinct spectra can have zero dissimilarity. Writing σ for the logistic sigmoid, $M_{ij} = \sigma(\gamma - D_{ij})$ has derivatives $\partial M_{ij}/\partial\gamma = -\partial M_{ij}/\partial D_{ij} = M_{ij}(1 - M_{ij}) \leq 1/4$. Because $D_{ii} = 0$, the diagonal $\sigma(\gamma)$ bounds all entries, and $M_{ij} = 0.5$ is attainable only for $\gamma \geq 0$. Scaling R by $s > 0$ changes both the midpoint (γ/s on the original scale) and the slope of the transition, so R and γ jointly determine the weight distribution. For finite arguments, M is dense, with no binary sampling or pruning, during both training and inference.

Appendix C.4. Task-Aware Expert Routing

The decoder (Table A6) uses two always-active shared experts and the same bank of eight routed experts for both tasks. Each expert preserves the 16 × 256 channel-feature shape. A task router receives $\bar{Z} = 16^{-1} \sum_n Z_{n,:}$ and returns $p_t = \mathrm{Softmax}(\bar{Z} W_t + b_t)$, where $W_t \in \mathbb{R}^{256\times 8}$. Selection is per input window, shared across its 16 channel rows. Ties are resolved by ascending expert index. The task-specific sum is then channel-mean-pooled before its 256 → 16 classification head.

The selected weights are $w_{t,j} = p_{t,j} / \sum_{l\in\mathcal{K}_t} p_{t,l}$ on the three selected experts and zero elsewhere, and each shared expert keeps coefficient one.

Auxiliary-loss normalization. Let B be the number of input windows in a minibatch, $n_r = 8$, and $k = 3$. The specified probability–usage loss follows the balancing principle in [30]:

$$\bar{p}_{t,j} = \frac{1}{B}\sum_{b=1}^{B} p_{t,j}^{(b)}, \qquad f_{t,j} = \frac{1}{3B}\sum_{b=1}^{B} 1\{j \in \mathcal{K}_t^{(b)}\}, \qquad \mathcal{L}_{aux}^{spec} = \sum_{t\in\{scene,diagnosis\}} 8\sum_{j=1}^{8} f_{t,j}\bar{p}_{t,j}.$$

The probabilities precede Top-3 selection; the hard selection counts are treated as constants for differentiation. Both $\sum_j f_{t,j}$ and $\sum_j \bar{p}_{t,j}$ equal one. At uniform usage and probabilities, the two-task sum is 2, and its contribution with coefficient 0.1 is 0.2. The term applies only to the routed TA-MoE experts; observed utilization is reported in Section 4.4.

Appendix D. Additional Experimental Results and Analysis

This appendix reports additional results for Section 4.

Appendix D.1. Module-Level Ablation Values

Table A10 lists the configurations shown in Figure 7.

Table A10. Module-level configurations under the dual 16-class setting. Accuracy, JEM and total parameter counts are reported for each evaluated variant.

Variant	Branch	Fusion	Head	Scene (%)	Fault (%)	JEM (%)	Params Total (M)
V1 Temporal only	T	None	Shared FFN	74.5	58.7	43.7	4.73
V2 Image only	I	None	Shared FFN	55.6	66.8	37.1	6.60
V3 Dual, no fusion	T + I	Independent	Shared FFN	75.2	67.5	50.8	11.06
V4 Plain concat	T + I	Concat	Shared FFN	76.8	68.2	52.4	11.19
V5 Plain CA (internal)	T + I	Plain CA	Shared FFN	78.5	69.4	54.5	11.26
V6 Topology mask	T + I	TMCA	Shared FFN	85.3	72.8	62.1	11.39
V7 Full MM-MoE	T + I	**TMCA**	**TA-MoE**	**89.4 ± 0.5**	**76.5 ± 0.6**	**69.2 ± 0.6**	**13.76**

Bold marks the full MM-MoE (V7). T and I denote temporal and image branches. V1–V6 are point values; V7 is mean ± sample SD over five training runs. V5 is the internal Plain CA configuration, distinct from the standalone baseline; V6 introduces the coupling prior in both channel lifting and attention.

Appendix D.2. Masking-Specific Image-Branch Comparison

Table A11 compares image-branch encodings under the favorable and strong-masking conditions; Tables A12 and A13 give the four-class source-level results discussed in Section 4.3.2.

Table A11. Fault diagnosis accuracy under the favorable and strong-masking conditions for different image-branch encoders. Accuracies are mean ± sample standard deviation over five independent training runs on the same file-level split (Appendix A.6); drops are differences between the reported means.

Encoding Strategy	Favorable Target 60 dB Acc. (%)	Strong Masking Acc. (%)	Drop (pp)
PCA-CWT	75.0 ± 0.8	66.4 ± 0.9	8.6
LR-CWT, $K = 2$	76.7 ± 0.7	69.5 ± 0.8	7.2
TG-LR-CWT, $K = 4$	**79.2 ± 0.6**	**73.8 ± 0.6**	**5.4**
TG-LR-CWT, $K = 8$	79.3 ± 0.6	73.9 ± 0.6	5.4
Per-channel CWT	79.4 ± 0.6	74.0 ± 0.6	5.4

Bold marks the default configuration.

Table A12. Four-class source-level fault accuracy (%) under the favorable and strong-masking conditions, rounded to whole percent.

Model	State	Favorable %	Masked %	Drop pp
MC-ABDS	Normal	86	73	13
MC-ABDS	Pitting	80	57	23
MC-ABDS	Crack	82	65	17
MC-ABDS	Missing tooth	88	77	11
MM-MoE	Normal	90	84	6
MM-MoE	Pitting	85	73	12
MM-MoE	Crack	88	80	8
MM-MoE	Missing tooth	91	89	2

Table A13. Source-level confusion rows (%) under strong masking; columns are Normal/Pitting/Crack/Missing tooth.

True Class	MC Confusion Row %	MM Confusion Row %
Normal	73/12/9/6	84/7/5/4
Pitting	28/57/10/5	18/73/6/3
Crack	22/8/65/5	14/4/80/2
Missing tooth	15/5/3/77	7/2/2/89

Appendix D.3. Region-Holdout Results by Fold

Table A14 reports the three endpoints of Section 4.3.3 for each held-out region pair. For MM-MoE, the 7.2-point decrease in ordered accuracy consists of a 6.85-point decrease in order-invariant accuracy and 0.35 points of otherwise correct but source-swapped assignments. The standard deviations across the four held-out regions are 2.3, 1.9 and 2.3 points for the ordered, order-invariant and height–intensity endpoints.

Table A14. Region-holdout results by fold (%): mean over five training runs for seen and excluded (unseen) regions.

Model	Held Out	Ordered Seen/Unseen	Unordered Seen/Unseen	Height Seen/Unseen
MM-MoE	R1–2	75.1/66.4	79.8/72.0	90.0/80.4
MM-MoE	R3–4	75.8/70.0	80.5/74.7	90.7/84.1
MM-MoE	R5–6	75.9/70.6	80.8/75.2	90.9/84.5
MM-MoE	R7–8	75.2/66.2	79.7/71.5	90.4/80.2
MC-ABDS	R1–2	65.8/54.8	70.8/61.0	82.6/70.2
MC-ABDS	R3–4	66.4/58.6	71.3/64.1	83.3/74.4
MC-ABDS	R5–6	66.6/59.2	71.6/64.8	83.5/74.8
MC-ABDS	R7–8	66.0/54.6	71.1/60.5	82.6/69.8

Appendix D.4. Channel Lifting and Phase-Aware TF Construction

Table A15 shows that channel lifting is not only a dimensional adapter. A global pooled TF broadcast reduced JEM to 60.1%, whereas the full coupling-prior-guided lifting strategy reached 69.2% JEM. In Table A16, adding phase channels (D3) improves on the log-magnitude-only D2 by 1.8 scene, 2.4 fault and 3.2 JEM points and on the RGB-magnitude D1 by 2.6, 2.9 and 4.2 points.

Table A15. Ablation of channel lifting strategies. Entries with ± report mean ± sample standard deviation over five independent training runs on the same file-level split (Appendix A.6).

Variant	Lifting Strategy	Task 1 Acc. (%)/F1	Task 2 Acc. (%)/F1	JEM (%)
C1	Global pooled TF broadcast	82.8 ± 0.8 0.812	70.9 ± 0.8 0.693	60.1 ± 0.9
C2	Learnable lifting matrix A only	86.7 ± 0.6 0.852	73.5 ± 0.7 0.718	65.3 ± 0.7
C3	A initialized from P^T	87.8 ± 0.6 0.864	74.5 ± 0.7 0.730	66.7 ± 0.7
C4	A **refined by SC-CCM coupling prior** M	**89.4 ± 0.5** **0.882**	**76.5 ± 0.6** **0.751**	**69.2 ± 0.6**

Bold marks the default MM-MoE configuration (C4). F1 denotes the Macro-F1 point value on the 0–1 scale. Each task cell lists accuracy above Macro-F1.

Table A16. Ablation of prototype TF map construction. Classification metrics are mean ± sample standard deviation over five independent training runs on the same file-level split (Appendix A.6).

Variant	TF Components	Task 1 Acc. (%)	Task 2 Acc. (%)	JEM (%)
D1	RGB colormap of magnitude scalogram	86.8 ± 0.7	73.6 ± 0.7	65.0 ± 0.8
D2	$\log \lvert C_k \rvert$ only	87.6 ± 0.6	74.1 ± 0.7	66.0 ± 0.7
D3	$[\log \lvert C_k \rvert, \cos(\angle C_k), \sin(\angle C_k)]$	**89.4 ± 0.5**	**76.5 ± 0.6**	**69.2 ± 0.6**

Bold marks the default configuration (D3).

Appendix D.5. Image-Branch Computational Cost

Table A17 lists the relative CWT operation count (normalized to 16 transforms) separately from the training-time and GPU-memory ratios (normalized to $K = 4$). The per-channel ratios of 2.65× and 2.35× correspond to 8.27/3.12 min per epoch and 19.7/8.4 GB in Table A7; the corresponding inference-latency ratio is 78.6/31.8 = 2.47.

Table A17. Reported CWT operation counts, relative training time, and GPU memory for the encoding configurations.

Variant	TF Maps /Sample	CWT Ops /Sample	Relative CWT Count	Training Time Ratio	GPU Memory Ratio
PCA-CWT	1	1	6.25%	0.58×	0.72×
LR-CWT, $K = 2$	2	2	12.5%	0.78×	0.85×
TG-LR-CWT, $K = 4$	**4**	**4**	**25%**	**1.00×**	**1.00×**
TG-LR-CWT, $K = 8$	8	8	50%	1.45×	1.38×
Per-channel CWT	16	16	100%	2.65×	2.35×

Bold marks the default configuration. Relative CWT count counts transforms, not FLOPs or elapsed time. Per-epoch time is not total time to convergence.

Appendix E. Run-Level Uncertainty and Between-Model Comparisons

Tables A18–A20 give two-sided 95% confidence intervals for the mean metrics over five independent training runs (Appendix A.6). They describe training variability conditional on the fixed split and are marginal intervals for individual means, not intervals for differences between methods.

Table A18. Run-level 95% confidence intervals for the baseline comparison in Table 3.

Method	Scene (%)	Fault (%)	JEM (%)
SVM + Spectral	[30.0, 33.0]	[26.5, 30.3]	[8.8, 11.6]
RF + Wavelet	[34.1, 37.5]	[30.5, 33.7]	[11.3, 13.7]
1D-CNN (Time)	[67.0, 69.8]	[49.7, 52.7]	[35.2, 38.4]
DUET (Time)	[73.4, 75.6]	[57.5, 59.9]	[43.8, 46.6]
ResNet-50 (Image)	[50.7, 53.9]	[63.5, 65.5]	[33.9, 36.9]
Vanilla Vim (Image)	[54.4, 56.8]	[65.9, 67.7]	[37.7, 40.1]
Standalone Plain CA	[75.2, 77.2]	[67.4, 69.6]	[53.0, 55.2]
MC-ABDS (MTL)	[77.6, 79.4]	[66.4, 68.4]	[54.8, 56.8]
MM-MoE (Proposed)	[88.8, 90.0]	[75.8, 77.2]	[68.5, 69.9]

Table A19. Run-level 95% confidence intervals for encoding, lifting, and TF-construction comparisons.

Configuration	Scene (%)	Fault (%)	JEM (%)
Encoding configurations (Table 5)			
PCA-CWT	[83.2, 85.0]	[71.0, 73.0]	[61.7, 63.7]
LR-CWT, $K = 2$ (no topology refinement)	[86.2, 87.6]	[73.0, 74.8]	[65.0, 66.8]
TG-LR-CWT, $K = 4$	[88.8, 90.0]	[75.8, 77.2]	[68.5, 69.9]
TG-LR-CWT, $K = 8$	[89.0, 90.2]	[75.9, 77.3]	[68.6, 70.0]
Per-channel CWT	[89.1, 90.3]	[76.1, 77.5]	[68.7, 70.1]
Channel lifting (Table A15)			
C1: pooled TF broadcast	[81.8, 83.8]	[69.9, 71.9]	[59.0, 61.2]
C2: learnable A	[86.0, 87.4]	[72.6, 74.4]	[64.4, 66.2]
C3: A initialized from P transpose	[87.1, 88.5]	[73.6, 75.4]	[65.8, 67.6]
C4: SC-CCM-refined A	[88.8, 90.0]	[75.8, 77.2]	[68.5, 69.9]
TF construction (Table A16)			
D1: RGB magnitude scalogram	[85.9, 87.7]	[72.7, 74.5]	[64.0, 66.0]
D2: log magnitude	[86.9, 88.3]	[73.2, 75.0]	[65.1, 66.9]
D3: phase-aware TF maps	[88.8, 90.0]	[75.8, 77.2]	[68.5, 69.9]

Italic rows denote groups of configurations and the tables in which they are reported.

Table A20. Run-level 95% confidence intervals for the two reported masking-condition accuracies in Table A11.

Encoding	Favorable (%)	Masked (%)
PCA-CWT	[74.0, 76.0]	[65.3, 67.5]
LR-CWT, $K = 2$	[75.8, 77.6]	[68.5, 70.5]
TG-LR-CWT, $K = 4$	[78.5, 79.9]	[73.1, 74.5]
TG-LR-CWT, $K = 8$	[78.6, 80.0]	[73.2, 74.6]
Per-channel CWT	[78.7, 80.1]	[73.3, 74.7]

Between-Model Comparisons

Tables A21 and A22 report the Welch comparisons defined in Appendix A.6. Each cell gives the mean difference in percentage points (positive values favor MM-MoE or $K = 4$), its unadjusted 95% confidence interval, the Welch statistic t, the degrees of freedom ν and the Holm-adjusted p value p_H over all 30 comparisons.

Table A21. Welch tests of MM-MoE against all eight baselines. Differences and confidence limits are in percentage points; p_H is adjusted over the full 30-comparison family.

Comparator	Scene	Fault	JEM
SVM + Spectral	+57.9; [56.4, 59.4] $t = 99.59$; $\nu = 5.35$ $p_H = 1.62 \times 10^{-8}$	+48.1; [46.3, 49.9] $t = 66.57$; $\nu = 5.25$ $p_H = 1.45 \times 10^{-7}$	+59.0; [57.6, 60.4] $t = 105.29$; $\nu = 6.19$ $p_H = 7.71 \times 10^{-10}$
RF + Wavelet	+53.6; [51.9, 55.3] $t = 80.62$; $\nu = 5.00$ $p_H = 1.21 \times 10^{-7}$	+44.4; [42.8, 46.0] $t = 69.34$; $\nu = 5.63$ $p_H = 4.49 \times 10^{-8}$	+56.7; [55.4, 58.0] $t = 108.72$; $\nu = 6.55$ $p_H = 1.95 \times 10^{-10}$
1D-CNN (Time)	+21.0; [19.7, 22.3] $t = 38.86$; $\nu = 5.59$ $p_H = 4.65 \times 10^{-7}$	+25.3; [23.8, 26.8] $t = 42.17$; $\nu = 5.88$ $p_H = 2.16 \times 10^{-7}$	+32.4; [30.8, 34.0] $t = 50.60$; $\nu = 5.63$ $p_H = 1.90 \times 10^{-7}$
DUET (Time)	+14.9; [13.8, 16.0] $t = 32.36$; $\nu = 6.25$ $p_H = 3.35 \times 10^{-7}$	+17.8; [16.5, 19.1] $t = 34.13$; $\nu = 6.55$ $p_H = 1.90 \times 10^{-7}$	+24.0; [22.6, 25.4] $t = 42.83$; $\nu = 6.19$ $p_H = 1.45 \times 10^{-7}$
ResNet-50 (Image)	+37.1; [35.5, 38.7] $t = 59.56$; $\nu = 5.16$ $p_H = 2.16 \times 10^{-7}$	+12.0; [11.0, 13.0] $t = 26.83$; $\nu = 7.42$ $p_H = 1.90 \times 10^{-7}$	+33.8; [32.3, 35.3] $t = 56.33$; $\nu = 5.88$ $p_H = 6.95 \times 10^{-8}$
Vanilla Vim (Image)	+33.8; [32.6, 35.0] $t = 67.60$; $\nu = 5.88$ $p_H = 2.58 \times 10^{-8}$	+9.7; [8.7, 10.7] $t = 23.53$; $\nu = 7.82$ $p_H = 2.16 \times 10^{-7}$	+30.3; [29.0, 31.6] $t = 58.10$; $\nu = 6.55$ $p_H = 1.10 \times 10^{-8}$
Standalone Plain CA	+13.2; [12.2, 14.2] $t = 31.29$; $\nu = 6.71$ $p_H = 2.16 \times 10^{-7}$	+8.0; [6.9, 9.1] $t = 16.54$; $\nu = 6.97$ $p_H = 5.28 \times 10^{-6}$	+15.1; [14.0, 16.2] $t = 31.22$; $\nu = 6.97$ $p_H = 1.81 \times 10^{-7}$
MC-ABDS (MTL)	+10.9; [10.0, 11.8] $t = 28.33$; $\nu = 7.24$ $p_H = 1.90 \times 10^{-7}$	+9.1; [8.1, 10.1] $t = 20.35$; $\nu = 7.42$ $p_H = 7.09 \times 10^{-7}$	+13.4; [12.4, 14.4] $t = 29.96$; $\nu = 7.42$ $p_H = 1.19 \times 10^{-7}$

Table A22. Welch tests of the default $K = 4$ encoding against $K = 8$ and per-channel CWT, included in the same 30-comparison family.

Comparator	Scene	Fault	JEM
TG-LR-CWT, $K = 8$	−0.2; [−0.9, 0.5] $t = -0.63$; $\nu = 8.00$ $p_H = 1.000$	−0.1; [−1.0, 0.8] $t = -0.26$; $\nu = 8.00$ $p_H = 1.000$	−0.1; [−1.0, 0.8] $t = -0.26$; $\nu = 8.00$ $p_H = 1.000$
Per-channel CWT	−0.3; [−1.0, 0.4] $t = -0.95$; $\nu = 8.00$ $p_H = 1.000$	−0.3; [−1.2, 0.6] $t = -0.79$; $\nu = 8.00$ $p_H = 1.000$	−0.2; [−1.1, 0.7] $t = -0.53$; $\nu = 8.00$ $p_H = 1.000$

All 24 comparisons between MM-MoE and the eight baselines reject equal run-level means after Holm correction ($p_H \leq 5.28 \times 10^{-6}$). None of the six comparisons among $K = 4$, $K = 8$ and per-channel CWT does ($p_H = 1.000$); this does not establish equivalence, and $K = 4$ is retained as an accuracy–cost compromise.

Appendix F. Acoustic Rationale for Design Choices

Table A23 connects the fixed design choices of Section 3.1 to acoustic and signal-processing considerations. The nominal scales are calculated from the setup, using the 0.30 m array radius, a sound speed of $c = 343$ m/s and the ideal gear kinematics of Appendix A.1; none of them enters the model or the training objective.

Table A23. Acoustic rationale, calculated scales, and interpretation limits of the fixed design choices.

Design Choice	Acoustic Rationale	Calculated Scale or Relation	Interpretation Limit
Phase channels in prototype CWT maps	A delay changes spectral phase; sine and cosine retain phase information without an angular wrap discontinuity [37].	For a common source and direct paths, $\lvert\Delta d\rvert \leq 0.60$ m gives $\lvert\Delta\tau\rvert \leq 1.75$ ms, approximately 17.5 sample intervals at 10 kHz.	Prototype phase is not a measured microphone-pair delay. No delay estimator, direction-of-arrival model, or propagation equation is enforced.
Raw-waveform branch	The original waveform retains a finer sampling grid than the resized TF image, providing complementary temporal input.	A 409.6 ms window mapped to 224 columns gives a nominal 1.83 ms per column, versus 0.10 ms per waveform sample.	Column spacing is not CWT resolving power, which depends on wavelet scale and resampling. The branch performs no explicit beamforming or time alignment.
Amplitude spectra for SC-CCM	A pure delay changes phase without changing spectral magnitude [37]; multiple propagation paths can change magnitude through interference [38].	For two paths with a constant reflection coefficient, ripple spacing is $\Delta f = c/\lvert d_r - d_d\rvert$, where d_r and d_d are reflected and direct path lengths.	Finite-window effects and source mixtures can change magnitude. No setup-specific notch spacing is inferred; M is a learned similarity prior, not a transfer-function or source-strength estimate.
Anisotropic patches, $P_h = 32$, $P_w = 4$	Narrow temporal support is motivated by localized transients; the frequency extent provides context across spectral bands.	Four image columns span a nominal 7.31 ms. The ideal local sun-fault repetition period is 12.8 ms at 78.125 Hz.	The scale comparison does not guarantee separation of successive impulses; effective CWT support and patch boundaries also matter. Fault frequencies are not supplied to the model.
Dual-axis scanning	Time and frequency have different structure: Persistence within a band and distribution across bands. Separate traversal orders represent these structures.	The ideal mesh frequency is 546.875 Hz. Orders 1–9, including the fundamental, fall below the 5 kHz Nyquist frequency; order 9 is 4921.875 Hz.	No harmonic or sideband template is enforced. Frequencies below Nyquist are not necessarily detectable or resolved in the acquired signal.

The two-path relation is an idealized explanation of possible spectral coloration by the floor reflection, not a fitted room model, and the direct-path delay bound does not bound reverberant paths.

References

1. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [CrossRef]
2. Lei, Y.; Lin, J.; Zuo, M.J.; He, Z. Condition monitoring and fault diagnosis of planetary gearboxes: A review. *Measurement* **2014**, *48*, 292–305. [CrossRef]
3. Yin, H.; Xu, H.; Zhang, H.; Wang, J.; Fang, X. MC-ABDS: A system for low SNR fault diagnosis in industrial production with intense overlapping and interference. *Appl. Acoust.* **2025**, *227*, 110217. [CrossRef]
4. Li, C.; Chen, C.; Gu, X. Acoustic-based rolling bearing fault diagnosis using a co-prime circular microphone array. *Sensors* **2023**, *23*, 3050. [CrossRef]
5. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, 21–27 July 2024*; PMLR: Cambridge, MA, USA, 2024; Volume 235, pp. 62429–62442. Available online: https://proceedings.mlr.press/v235/zhu24f.html (accessed on 11 September 2026).
6. Sener, O.; Koltun, V. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems, Proceedings of the 32nd Conference on Neural Information Processing Systems, Montréal, QC, Canada, 2–8 December 2018*; Curran Associates Inc.: Red Hook, NY, USA, 2018; Volume 31, pp. 527–538. Available online: https://papers.neurips.cc/paper_files/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html (accessed on 11 September 2026).
7. Zhang, W.; Peng, G.; Li, C.; Chen, Y.; Zhang, Z. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors* **2017**, *17*, 425. [CrossRef]
8. Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A.K. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, 106587. [CrossRef]
9. Feng, Z.; Liang, M.; Chu, F. Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples. *Mech. Syst. Signal Process.* **2013**, *38*, 165–205. [CrossRef]
10. Bartelmus, W.; Zimroz, R. A new feature for monitoring the condition of gearboxes in non-stationary operating conditions. *Mech. Syst. Signal Process.* **2009**, *23*, 1528–1534. [CrossRef]
11. Chiariotti, P.; Martarelli, M.; Castellini, P. Acoustic beamforming for noise source localization – Reviews, methodology and applications. *Mech. Syst. Signal Process.* **2019**, *120*, 422–448. [CrossRef]
12. Li, M.; Hu, D.; Liao, A.; Shi, W.; He, S. A multi-beam focused data fusion approach for gearbox fault diagnosis based on microphone array measurement. *SSRN* **2026**, *preprint*. [CrossRef]
13. Deng, J. Gearbox multimodal fault diagnosis based on microphone array and cross-attention. In *Proceedings of the Sixth International Conference on Testing Technology and Automation Engineering (TTAE 2026), Zhuhai, China, 12–14 June 2026*; SPIE: Bellingham, WA, USA, 2026; Volume 14383. [CrossRef]
14. Purohit, H.; Tanabe, R.; Ichige, K.; Endo, T.; Nikaido, Y.; Suefusa, K.; Kawaguchi, Y. MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop, New York, NY, USA, 25–26 October 2019; pp. 209–213. [CrossRef]
15. Lee, J.-G.; Kim, K.S.; Lee, J.H. Sound-based unsupervised fault diagnosis of industrial equipment considering environmental noise. *Sensors* **2024**, *24*, 7319. [CrossRef]
16. Mo, C.; Huang, K.; Li, W.; Xu, K. A lightweight and efficient multimodal feature fusion network for bearing fault diagnosis in industrial applications. *Sensors* **2024**, *24*, 7139. [CrossRef]
17. Chu, Z.; Xing, S.; Han, B.; Wang, J. A novel cross-domain mechanical fault diagnosis method fusing acoustic and vibration signals by vision transformer. *Sensors* **2024**, *24*, 5120. [CrossRef]
18. Sun, D.; Li, Y.; Jia, S.; Feng, K.; Liu, Z. Non-contact diagnosis for gearbox based on the fusion of multi-sensor heterogeneous data. *Inf. Fusion* **2023**, *94*, 112–125. [CrossRef]
19. Zhang, Y.; Ding, J.; Li, Y.; Ren, Z.; Feng, K. Multi-modal data cross-domain fusion network for gearbox fault diagnosis under variable operating conditions. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108236. [CrossRef]
20. Qiu, Z.; Fan, S.; Liang, H.; Liu, J. Multimodal fusion fault diagnosis method under noise interference. *Appl. Acoust.* **2025**, *228*, 110301. [CrossRef]
21. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems, Proceedings of the 35th Conference on Neural Information Processing Systems, Virtual Event, 6–14 December 2021*; Curran Associates Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 14200–14213. Available online: https://proceedings.neurips.cc/paper/2021/hash/76ba9f564ebbc35b1014ac498fafadd0-Abstract.html (accessed on 18 October 2026).
22. Li, C.; Mo, L.; Kwoh, C.K.; Li, X.; Chen, Z.; Wu, M.; Yan, R. Noise-robust multi-view graph neural network for fault diagnosis of rotating machinery. *Mech. Syst. Signal Process.* **2025**, *224*, 112025. [CrossRef]

23. Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems, Proceedings of the 35th Conference on Neural Information Processing Systems, Virtual Event, 6–14 December 2021*; Curran Associates Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 28877–28888. Available online: https://proceedings.neurips.cc/paper_files/paper/2021/file/f1c1592588411002af340cbaedd6fc33-Paper.pdf (accessed on 18 October 2026).
24. Qiu, X.; Wu, X.; Lin, Y.; Guo, C.; Hu, J.; Yang, B. DUET: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1, Toronto, ON, Canada, 3–7 August 2025*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 1185–1196. [CrossRef]
25. Xie, Z.; Chen, J.; Feng, Y.; Zhang, K.; Zhou, Z. End to end multi-task learning with attention for multi-objective fault diagnosis under small sample. *J. Manuf. Syst.* **2022**, *62*, 301–316. [CrossRef]
26. Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; Chi, E.H. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1930–1939. [CrossRef]
27. Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; Houlsby, N. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems, Proceedings of the 35th Conference on Neural Information Processing Systems, Virtual Event, 6–14 December 2021*; Curran Associates Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 8583–8595. Available online: https://papers.neurips.cc/paper_files/paper/2021/hash/48237d9f2dea8c74c2a72126cf63d933-Abstract.html (accessed on 11 September 2026).
28. Tang, H.; Liu, J.; Zhao, M.; Gong, X. Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. In *Proceedings of the Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, 22–26 September 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 269–278. [CrossRef]
29. Dai, D.; Deng, C.; Zhao, C.; Xu, R.X.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; Xie, Z.; Li, Y.K.; Huang, P.; Luo, F.; Ruan, C.; Sui, Z.; Liang, W. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: New York, NY, USA, 2024; Volume 1, pp. 1280–1297. [CrossRef]
30. Fedus, W.; Zoph, B.; Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **2022**, *23*, 1–39. Available online: https://jmlr.org/papers/v23/21-0998.html (accessed on 18 October 2026).
31. Shi, P.; Liu, H.; Xu, X.; Zhao, D.; Hua, C. M4oE: A multitask multi-input multiscale mixture-of-experts method for multisensor fusion diagnosis. *IEEE Internet Things J.* **2026**, *13*, 22103–22118. [CrossRef]
32. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtual Event, 25–29 April 2022. Available online: https://openreview.net/forum?id=uYLFoz1vlAC (accessed on 11 September 2026).
33. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In Proceedings of the First Conference on Language Modeling (COLM 2024), Philadelphia, PA, USA, 7–9 October 2024. Available online: https://openreview.net/forum?id=tEYskw1VY2 (accessed on 11 September 2026).
34. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems, Proceedings of the 38th Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024*; Curran Associates Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 103031–103063. [CrossRef]
35. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [CrossRef]
36. Wu, Y.; Sicard, B.; Gadsden, S.A. Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring. *Expert Syst. Appl.* **2024**, *255*, 124678. [CrossRef]
37. Smith, J.O., III. Shift theorem. In *Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications*; W3K Publishing: Stanford, CA, USA, 2007. Available online: https://dsprelated.com/freebooks/mdft/Shift_Theorem.html (accessed on 18 October 2026).
38. Smith, J.O., III. Acoustic modeling with digital delay. In *Physical Audio Signal Processing*; W3K Publishing: Stanford, CA, USA, 2010. Available online: https://dsprelated.com/freebooks/pasp/Acoustic_Modeling_Digital_Delay.html (accessed on 18 October 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.