University of St. Gallen

School of Computer Science

to obtain the title of

Master of Science in Computer Science

Master Thesis on

Multimodal Sensing for Acoustic–Vibration Based Fault Localization in Hydropower

---

Submitted by:

Zakaria OMARAR

20-889-010

Approved on the application by:

Prof. Dr. Bruno Rodrigues

Date of Submission:

June 30, 2026

Acknowledgements

This thesis is the result of a long road, and I did not walk it alone.

I am deeply grateful to my supervisor, Prof. Dr. Bruno Rodrigues, for trusting me with this project and for the guidance that shaped it at every turn. The critical questions, the steady encouragement, and the freedom to follow the evidence even when it led to unexpected and uncomfortable answers taught me more about doing honest research than any single result ever could.

I owe equal thanks to my assistant, Jinyuan Zhang, who was part of this work from beginning to end. He produced the data the thesis rests on, followed every idea closely as it took shape, and challenged the ones that needed challenging so that we could push the work as far as it would go. That mix of hands-on help and honest, demanding feedback shaped the result more than any other single contribution, and I am sincerely grateful for it.

Above all, I thank my mother and my sister. Through every year of my studies, in the good weeks and the difficult ones, they never once stopped believing in me. Their faith, their patience, and their quiet, unconditional support are the foundation everything here stands on. This work is as much theirs as it is mine, and I dedicate it to them with all my gratitude and love.

St. Gallen, June 30, 2026 Zakaria OMARAR

Abstract

Pumped-storage hydropower underpins grid-scale flexibility for renewable generation, but its frequent mode changes expose a weakness of acoustic fault diagnostics: when the operating regime shifts, the baseline signature moves with it and unimodal detectors raise synchronized false alarms, a domain shift confirmed on the industrial Rodundwerk II (ROW II) machine. This thesis asks a sharper question than the usual fuse-and-condition remedy: at realistic data scale, when does jointly trained multimodal fusion earn its complexity over simpler unimodal, late-fusion, or classical alternatives? It answers with a leakage-controlled comparison across anomaly detection and source localization.

Because ROW II recordings could not be released in time, the study uses a 3D-printed, reduced-scale turbine prototype recorded over five campaigns. A single chained system combines per-modality convolutional encoders, a cross-attention stage that learns a self-supervised context vector, a context-conditioned normalizing-flow anomaly head, and a localization head fusing a classical steered-power map with structure-borne time-difference features, all under recording-level splits. The available anomalies are induced casing impacts, which bounds the claims.

The verdict is consistent and partly unexpected: no fusion paradigm dominates, and parameter-free or single-pathway combinations repeatedly match or beat joint training. Conditioning the anomaly head on recovered context helps only slightly; robustness comes instead from a parameter-free late-fusion AND rule whose healthy alert rate stays near-zero and calibrated as conditions change. Localization is geometry-bound, with the accelerometer time-difference pathway transferring best to roughly 0.11 to 0.13 m, while offline mining of the real ROW II supervisory-control archive recovers the operating mode at up to 0.96 nats of mutual information on real-turbine data. At this scale, disciplined parameter-free use of the learned components beats end-to-end joint training, leaving whether fusion eventually overtakes them a falsifiable, data-scale-dependent prediction.

Table of Contents

List of Abbreviations

3D	Three-Dimensional
ADC	Analog-to-Digital Converter
ADU	Analog-to-Digital Units
AE	Acoustic Emission
AG	Aktiengesellschaft (German public limited company)
AI	Artificial Intelligence
AUC	Area Under the Curve
BEP	Best Efficiency Point
BiLSTM	Bidirectional Long Short-Term Memory
BPF	Blade Passing Frequency
CANDE-CP	Context-Aware Novelty Detection Autoencoder with Context Prediction
CBM	Condition-Based Maintenance
CI	Confidence Interval
CMA	Cross-Modal Alignment
CNF	Conditional Normalizing Flow
CNN	Convolutional Neural Network
CRNN	Convolutional Recurrent Neural Network
CVAE	Conditional Variational Autoencoder
CWT	Continuous Wavelet Transform
D1–D5	Recording Campaigns One to Five
DCASE	Detection and Classification of Acoustic Scenes and Events

DMA	Direct Memory Access
F1	F1 Score (harmonic mean of precision and recall)
FDR	False Discovery Rate
FiLM	Feature-wise Linear Modulation
FPR	False Positive Rate
GCC	Generalized Cross-Correlation
GCC-PHAT	Generalized Cross-Correlation with Phase Transform
GELU	Gaussian Error Linear Unit
KDE	Kernel Density Estimator
LN	Layer Normalization
LOPO	Leave-One-Position-Out (cross-validation)
LORO	Leave-One-Recording-Out (cross-validation)
LSTM	Long Short-Term Memory
LSTM-AE	Long Short-Term Memory Autoencoder
MAAN	Multi-Adversarial Attention Network
MAB	Multi-head Attention Block
MAE	Mean Absolute Error
MFCC	Mel-Frequency Cepstral Coefficients
MI	Mutual Information
MLP	Multilayer Perceptron
MMT-FD	Multi-Attention Meta Transformer for Fault Diagnosis
NLL	Negative Log-Likelihood
NMI	Normalized Mutual Information

NZE	Net Zero Emissions
OC-SVM	One-Class Support Vector Machine
PCM	Pulse-Code Modulation
PdM	Predictive Maintenance
PH	Phase-Shifting (operating mode)
PHAT	Phase Transform
PMA	Pooling by Multi-head Attention
PSH	Pumped-Storage Hydropower
PU	Pump (operating mode)
PV	Photovoltaic
RealNVP	Real-Valued Non-Volume-Preserving (flow)
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
ROC-AUC	Area Under the ROC Curve
ROW II	Rodundwerk II (reference pumped-storage facility)
rpm	Revolutions Per Minute
RQ1–RQ4	Research Questions One to Four
RSI	Rotor-Stator Interaction
SCADA	Supervisory Control and Data Acquisition
SNR	Signal-to-Noise Ratio
SR	Sample Rate
SRP	Steered-Response Power
SRP-PHAT	Steered-Response Power with Phase Transform

ST Standstill (operating mode)

STFT Short-Time Fourier Transform

SVM Support Vector Machine

TBM Time-Based Maintenance

TDOA Time-Difference-of-Arrival

TU Turbine (operating mode)

V0–V4 Pipeline Stages: V0 baselines, V1 encoders, V2 fusion, V3 anomaly head, V4 localization head

VAE Variational Autoencoder

VHC Vibrational Hill Chart

VRE Variable Renewable Energy

WAV Waveform Audio File Format

List of Figures

List of Tables

Chapter 1

Introduction

The global energy system is undergoing a structural transformation driven by binding climate commitments and national decarbonization strategies, as centralized fossil-fuel plants give way to decentralized, weather-dependent renewable generation. The International Energy Agency projects global renewable power capacity to grow by nearly 4,600 GW between 2025 and 2030, double the pace of the preceding five years, with variable renewable energy (VRE) sources, chiefly solar photovoltaics and wind, accounting for roughly 95% of all new additions [50]. This shift is indispensable for Net Zero Emissions (NZE) targets, but it introduces a structural vulnerability: as inverter-based generation displaces synchronous thermal machines, power grids lose the rotational inertia that once resisted frequency deviations and stabilized active power balance [43, 103, 112]. Compensating for this erosion demands responsive, dispatchable storage at grid scale, and among the available technologies pumped-storage hydropower (PSH) has consolidated its position as the primary stabilizing backbone of clean power systems [52, 130]. Hydropower remains the world's largest source of renewable electricity, supplying about 14.3% of global electricity demand, and global PSH capacity stood at roughly 189 GW by end-2024, forecast to add 16.5 GW annually by 2030 as the need for flexibility accelerates [50, 51].

These dynamics are acutely apparent in Austria, where hydropower accounted for up to 67% of domestically generated electricity in 2024, from a total installed capacity of 14,130 MW with a substantial pumped-storage share [51, 52]. The national energy strategy targets 100% renewable electricity generation by 2030,[4, 28] which places an unprecedented premium on the flexibility, efficiency, and continuous availability of the existing pumped-storage fleet [4, 113].

Continuous flexible operation, however, carries profound mechanical consequences. Pumped-storage plants were historically engineered for steady-state baseload generation

or predictable peak-shaving schedules. Today, their Francis pump-turbines must execute frequent start-stop cycles, sustain extended operation in deep part-load regimes, and function in phase-shifting mode to supply reactive power on demand [76, 97]. These regimes subject the hydraulic machinery to severe transient phenomena including vortex rope formation, high-frequency fluid-structure interactions, intense pressure pulsations, and acoustic cavitation [61, 76, 100]. Over time, these stresses accelerate structural fatigue: micro-fractures propagate in runner blades,[76, 100] guide vane linkages wear progressively,[20] and critical thrust and journal bearings degrade prematurely [9, 20]. Unplanned outages in high-capacity PSH units carry economic losses that extend well beyond repair costs to the disruption of grid balancing services [8, 9].

These pressures have catalyzed a paradigm shift in asset management. The hydroelectric industry is actively transitioning from reactive and time-based preventive maintenance to data-driven Predictive Maintenance (PdM) [9, 39, 61]. PdM seeks to optimize component lifetime by continuously estimating machine health from sensor data, identifying degradation trends before the operational impact becomes irreversible [9, 39]. Conventional approaches, however, rely on macroscopic indicators such as elevated bearing temperatures or vibration amplitudes that only trigger alerts when deterioration is already well-advanced and potentially irrecoverable [9, 28]. There is therefore an urgent scientific and industrial imperative to develop diagnostic systems capable of detecting and localizing incipient faults at their earliest emergence, long before conventional alarm thresholds are reached [39, 61].

Several sensing modalities have been proposed to meet this need, including thermal imaging, electrical signature analysis, and oil-debris monitoring [9, 61]. Recent studies show that the combination of acoustic and vibration signals yields higher fault-detection accuracy than either modality alone [25, 57]. Acoustic emissions respond to micro-fractures, cavitation, and blade-wake interactions that precede most mechanical failures. In isolation, however, acoustic signals suffer from severe ambiguity in the subterranean machine room of a PSH plant, where massive concrete walls, steel spiral casings, and complex networks of penstocks and draft tubes aggressively reflect and scatter sound waves, producing a complex superposition of direct and multiply reflected paths at every microphone [56, 77]. When multiple auxiliary systems such as cooling fans, governor oil pumps, and compressor units operate concurrently, their broadband emissions mask the subtle, high-frequency signatures of incipient faults [56, 61].

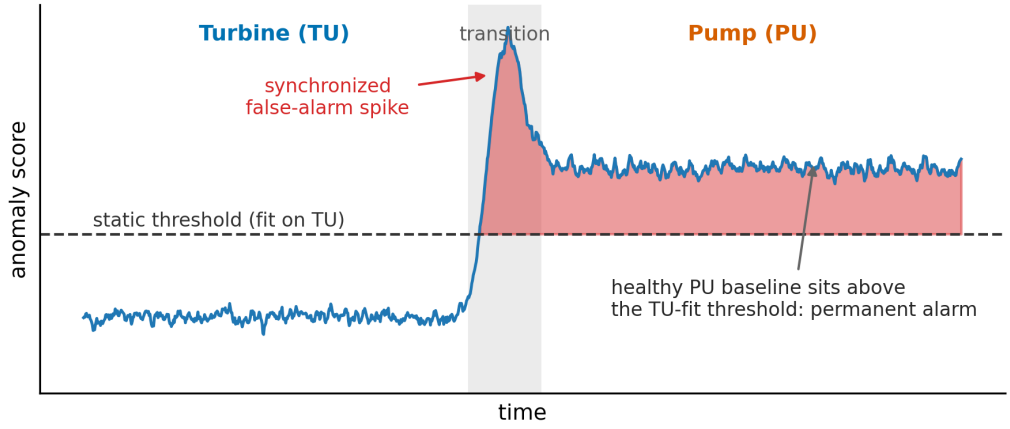


Figure 1.1: Schematic of the domain-shift failure mode. A detector calibrated on TURBINE-mode data fires a synchronized burst of false alarms across a healthy TU→PU transition, because the healthy PUMP baseline sits above the TURBINE-fit threshold. Illustrative, not measured.

This acoustic ambiguity makes fault localization highly unreliable when attempted from sound alone. An anomaly detected in the audio stream may originate from the upper generator assembly, the lower hydraulic turbine housing, or an entirely separate auxiliary system. The challenge is compounded by the inherent non-stationarity of reversible pump-turbines: the baseline acoustic and vibratory signature in Turbine (TU) mode is fundamentally different from those in Pump (PU) or Phase-shifting (PH) mode [47]. A static anomaly detection model applied uniformly across all modes will inevitably misclassify the routine acoustic shift accompanying a mode transition, such as TU $\rightarrow$ PU, as a catastrophic fault, producing an unmanageable rate of false-positive alarms. This phenomenon, known as *domain shift*, is the central failure mode of unimodal acoustic diagnostics and fundamentally undermines their operational reliability [47]. It was observed directly at industrial scale by Khamaisi et al., whose otherwise near-perfect acoustic anomaly detectors generated a synchronized spike of false alarms during a five-second generation-to-pumping transition that plant personnel confirmed as entirely normal machine behavior [65]. Current literature offers no systematic investigation of synchronized multimodal sensing combined with context-aware deep learning models capable of adapting to such domain shifts in the setting of large-scale pumped-storage hydropower.

Industrial Context and the Move to a Prototype

This thesis was initiated in direct cooperation with illwerke vkw AG, the operator of the Rodundwerk II (ROW II) pumped-storage facility in Vorarlberg, Austria [49]. ROW II was the intended object of study throughout. Commissioned in 1976 and comprehensively rehabilitated between 2009 and 2011, it houses a single reversible Francis pump-turbine that delivers 295 MW in turbine mode and consumes up to 286 MW in pump mode at a nominal rotational speed of 375 rpm [48]. The machine operates across four steady regimes, Standstill (ST), Turbine (TU), Pump (PU), and Phase-shifting (PH), each imposing a fundamentally different physical loading on the machinery [49]. The brief intervals between these regimes are mode transitions rather than separate operating modes, and they are precisely where unimodal detectors fail most often. The original intention was to develop and validate the proposed diagnostic architecture directly on operational data recorded at the ROW II machine.

That plan could not be realized, because the operational recordings from ROW II could not be released within the timeframe of this work. A low-rate supervisory-control archive of the machine was eventually released, but too late in the project, and in too coarse a form (one-hertz process channels rather than the high-rate acoustic and vibration recordings) to serve as the object on which the diagnostic architecture is trained and validated. Rather than abandon the scientific objectives, the investigation was carried out instead on a 3D-printed, reduced-scale Francis turbine prototype that mimics the ROW II machine at laboratory scale. This fallback carries three consequences that must be stated plainly. First, the prototype is much smaller than the 295 MW industrial machine and does not preserve its absolute scale, power, or hydraulics. Second, neither the rig nor its instrumentation matches the hardware a production plant would use: the microphones and accelerometers available for this work are less capable than industrial-grade sensors, and the housing is 3D-printed plastic rather than steel and reinforced concrete. Third, and most consequentially for the detection results, the anomalies recorded on the rig are deliberately induced casing impacts (“knocks”) rather than the naturally occurring incipient faults, such as cavitation, crack propagation, or bearing wear, that motivate the work. A knock is a broadband impulsive event that excites both modalities at once, which makes it a convenient but comparatively easy anomaly target; the detection numbers therefore establish a method and a relative ordering between paradigms rather than a sensitivity to real incipient damage, and this bound is revisited in Chapter 7.

The methodology nonetheless transfers, because the problems it targets are invariant to scale. The questions of how to fuse acoustic and vibration streams, how to infer the machine's operating context without manual labels, how to attribute an anomaly to a physical location inside a reverberant housing, and how to remain calibrated across abrupt shifts in operating condition do not depend on the absolute size of the turbine; they arise from the structure of the signals and the geometry of the sensor array, both of which the prototype reproduces. Where prototype scale materially constrains a result, this is flagged explicitly, and the limitations of the laboratory setting are revisited in the discussion of Chapter 7.

Two differences from the field machine must nevertheless be made explicit. First, the prototype reproduces three of the four operating modes, ST, PU, and TU, but not PH, and each mode was recorded in isolation rather than capturing the mode-to-mode transition events themselves. These three modes constitute the operational context that the model is required to recover without supervision: in the later campaigns the operating mode is genuinely present in the signal but was never annotated, and is instead discovered post hoc by clustering the learned context representation. Second, the later campaigns vary an added background-noise fan across three levels, denoted speed1 through speed3; these are acquisition settings rather than operating modes, and are pooled when the encoder learns its context representation (Chapter 3).

Because the corpus contains no transition events, robustness to a change of operating condition is measured not across a real transition but by a threshold-transfer protocol on the steady recordings, detailed in Chapter 5, while the transition event itself is approximated separately by synthetic crossfades. This is a partial proxy: it tests whether a calibrated decision boundary survives a change of steady operating and acquisition condition, not the abrupt regime change of a field TU $\rightarrow$ PU transition, and that gap is revisited in Chapter 7.

The prototype data was not collected as a single fixed dataset. It grew over five recording campaigns, denoted D1 through D5, each one extending the sensor array or refining the acquisition protocol of the one before. The array combines airborne acoustics and structural vibration: between four and nine microphones across the campaigns, together with four to five accelerometers mounted on the casing wall. Operating mode is labeled explicitly only in the first two campaigns (D1 and D2); the later three (D3, D4, and D5) hold the mode fixed and instead vary the speed1–speed3 noise conditions. The acoustic and vibration streams share an acquisition trigger, but hardware-level

alignment between them is not guaranteed, particularly for the raw-vibration stream, so a custom cross-modal synchronization step verifies and, where necessary, corrects the offset between the two streams so that they can be fused reliably both for this corpus and for future field data. The prototype, its sensor array, the five campaigns, and the full preprocessing pipeline are described in Chapter 3.

Research Questions

The central research question addressed in this thesis is:

> To what extent can latent representations of operational context be extracted from multi-modal sensor streams (acoustic and vibration), and does conditioning anomaly detection and source localization on these representations improve performance under domain shifts?

This question is pursued in a deliberately adversarial form. The prevailing assumption in the multimodal fault-diagnosis literature is that fusing modalities and jointly training a shared representation improves on simpler alternatives. Rather than assume that gain, this thesis tests it directly, treating each research question as a leakage-controlled, head-to-head contest between the unimodal, late-fusion, and intermediate (jointly trained) paradigms, and asking at what data scale, and for which task, a jointly trained fusion model earns its added complexity over a parameter-free or single-pathway alternative. The deliverable of this framing is a map of which paradigm wins which operating point, and the chained architecture developed in Chapters 3 and 4 is the testbed through which each paradigm is tried rather than a system assumed to win. A measured result in which a parameter-free alternative matches or beats the jointly trained model is therefore a designed outcome of the comparison, not a shortfall of the architecture.

More specifically, this thesis addresses four questions:

- Which multimodal fusion strategy best captures cross-modal dependencies between acoustic and vibration signals for the purpose of operational context learning?
- To what extent does conditioning anomaly detection on the inferred operational regime reduce false positive rates caused by domain shifts during mode transitions?
- How effectively can a source localization model exploit the fused representation and TDOA features to spatially attribute detected anomalies within the turbine housing?

- Does incorporating complementary supervisory signals such as temperature and pressure reduce residual uncertainty and further improve detection and localization performance? This fourth question is posed from the outset as a deployment-feasibility analysis, answered by offline mining of the real plant archive and a wiring check on the prototype, rather than as a hypothesis the prototype corpus can settle, for the reason given immediately below.

The fourth question cannot be settled in full with the data on which the system is trained, because the prototype corpus contains only acoustic and vibration channels and carries no temperature, pressure, or SCADA signals. It is therefore retained as an exploratory constraint that scopes the present work rather than as a question the prototype alone can answer. It is nonetheless addressed in two partial ways. First, the conditioning pathway is exercised on the prototype with the recorded operating-condition token as a stand-in supervisory variable to test the end-to-end mechanism. More substantively, the real ROW II data, delivered by the company after the prototype experiments were complete, is mined offline: an information-theoretic ranking of its process channels is conducted to investigate whether the operating context is recoverable, and to explore which specific channels a field deployment might route into the supervisory vector. What remains for future instrumented campaigns is the one part neither source can supply, a per-fault validation of supervisory conditioning on a machine that carries these channels live. This offline analysis, and the caveats that bound it, are reported in Chapter 6 and discussed in Chapter 7.

Contributions

The contributions of this thesis follow directly from this adversarial framing, and the chained system it builds (the preprocessing pipeline, the label-free cross-attention context encoder, the context-conditioned normalizing-flow anomaly head, and the fusion localization head, all detailed in Chapters 3 and 4) is the vehicle through which each paradigm is tested rather than an end in itself.

The first and central contribution is methodological: a leakage-controlled, head-to-head comparison of the unimodal, late-fusion, and intermediate-fusion paradigms across both anomaly detection and source localization, conducted under recording-level splits, held-out calibration, multi-seed stability reporting, and baselines drawn from the prior ROW II study the system is meant to improve on. This turns the usual question of whether

fusion helps into the more useful question of which paradigm to reach for under which operational demand.

The second contribution is a set of measured negative and null results, each reported with its mechanism rather than buried: that the fused operating-context representation is acoustic-dominated and that fusion adds no seed-robust mode-discovery gain at this scale; that conditioning the density head improves the healthy model only modestly; that adding the acoustic pathway harms position-held-out localization through soft-argmax overfitting; and that no supervisory channel survives a global anomaly-association test on the real archive. A negative result with a measured cause is usable guidance for a field deployment, and reporting it as such is treated here as a first-class outcome.

The third contribution is a concrete, deployable recipe that follows from those results: a parameter-free late-fusion AND rule as the high-specificity anomaly operating point, an accelerometer time-difference (TDOA) pathway for spatial transfer with sensor placement as the governing design variable, and a ranked set of supervisory channels to route into the detector's context.

The fourth contribution carries the operating-context premise off the prototype and onto real-turbine data: an offline information-theoretic mining of the real ROW II supervisory-control archive that confirms the operating mode is almost fully determined by its process channels, and that yields a bounded recommendation for the channels a field deployment should use.

In support of these contributions, the complete source code that produced every result and figure in this thesis is released publicly; Appendix C gives the repository and, together with the hyperparameter grids and run manifests of Appendix A, is intended to make the entire study reproducible.

System Overview

Before the individual components are developed in the chapters that follow, this section presents the complete system in a single view, so that each later chapter reads as the elaboration of one part of a whole that is already familiar. End to end, the system takes the two synchronized sensor streams of the prototype, airborne acoustics and structural vibration, and turns them into two operator-facing products: a calibrated anomaly alert and, for the windows that raise one, an estimate of the three-dimensional source position inside the casing. The same chain is also the apparatus through which the central comparison of this thesis is run, because each learned stage can be exchanged for a

unimodal, late-fusion, or intermediate-fusion variant without disturbing the stages around it. Figure 1.2 shows this flow, organized into the three subsystems that the remaining chapters expand in turn.

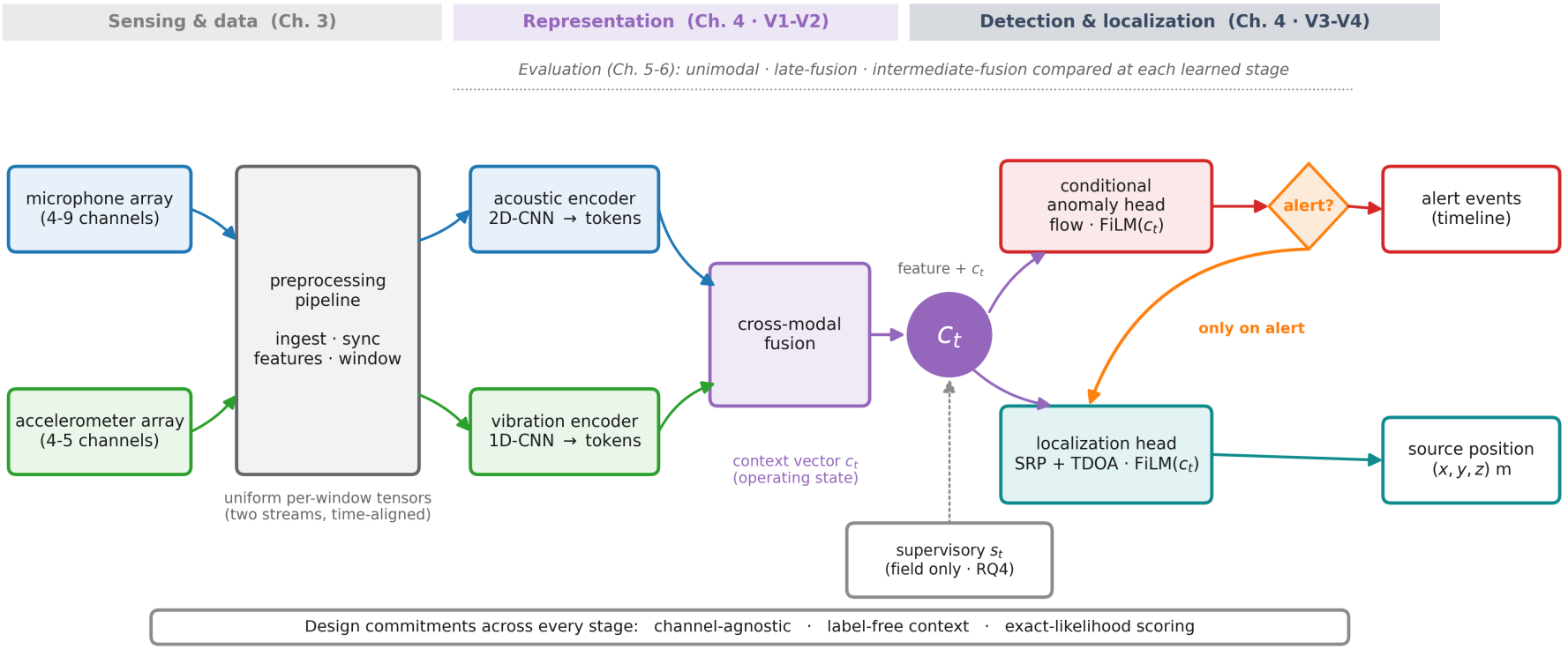


Figure 1.2: High-level overview of the complete system: synchronized acoustic and vibration streams are reduced to per-window tensors, encoded into a label-free context vector $\boldsymbol{c}_t$, and passed to two $\boldsymbol{c}_t$-conditioned heads that emit alerts and, only on alert, a source position. The model core is expanded block by block in §4.1.

The first subsystem (Chapter 3) is responsible for turning a heterogeneous, incrementally collected corpus into one uniform input format. The five recording campaigns differ in array size, in the rate and format of their vibration stream, and in what they annotate, and the preprocessing pipeline absorbs all of this variation: it ingests each recording, verifies the cross-modal time alignment, extracts a time-frequency image per microphone and a three-channel series per accelerometer, and cuts both into time-aligned windows. Its output is a per-window tensor pair in which the two modalities are registered in time but never mixed in value, so the downstream model is left free to decide how, and how much, to combine them. Keeping the streams separate at this boundary is precisely what makes the fusion question an empirical one rather than a choice baked into the data.

The second subsystem (Chapter 4, stages V1 and V2) learns a representation of the machine's operating state. A pair of per-modality convolutional encoders reduce each sensor channel to a token, a single cross-attention stage lets the acoustic and vibration tokens read one another, and a pooling step compresses the result into one context vector $\boldsymbol{c}_t$ that summarizes the current operating condition. This representation is learned without operating-mode labels, by self-supervision, because the mode is present in the signal but annotated on only two of the five campaigns; and because every operation in the subsystem is permutation-invariant, a single set of weights serves the differently-sized arrays of the corpus. The context vector is the pivot of the whole design: it is the quantity examined to test whether operating mode is recoverable without supervision, and the signal on which both downstream heads are conditioned.

The third subsystem (Chapter 4, stages V3 and V4) turns that representation into decisions. A conditional density model scores each window by how improbable it is under a learned model of healthy operation that is modulated by $\boldsymbol{c}_t$, and a per-context threshold converts the score into an alert; conditioning on $\boldsymbol{c}_t$ is what holds the alert rate steady as the operating regime changes, which is the domain-shift failure the system is built to resist. Only windows that raise an alert reach the localization head, which fuses a classical acoustic power map with structure-borne time-difference cues, again conditioned on $\boldsymbol{c}_t$, to place the source inside the casing. The architecture additionally exposes an optional supervisory pathway for slow process channels such as temperature and pressure; these are absent from the prototype and are therefore carried as a field-only provision, the subject of the fourth research question, and drawn dashed in Figure 1.2.

Three commitments run through every stage and account for the shape of the design as a whole. The system is *channel-agnostic*, so the same trained weights process any

array geometry; its operating context is *label-free*, so it leans on no annotation a field deployment might lack; and its anomaly score is an *exact likelihood* rather than a reconstruction error, so the score is directly interpretable and can be streamed. Because the stages are chained on frozen weights and each learned block has a parameter-free or single-pathway counterpart, this one system doubles as the testbed for the paradigm comparison of Chapters 5 and 6. With this map in place, the remaining chapters develop each subsystem in detail.

Thesis Outline

Mapping the subsystems of Figure 1.2 onto the document, the remainder of this thesis is organized as follows. Chapter 2 reviews the relevant literature in predictive maintenance, acoustic and vibration sensing, multimodal deep learning, anomaly detection, and source localization. Chapter 3 presents the design of the data and preprocessing pipeline: the ROW II reference machine, the bench-top prototype and its sensor array, the five recording campaigns (D1 through D5), and the steps that turn the raw recordings into the per-window tensors used for learning. Chapter 4 describes the proposed multimodal architecture, covering the per-modality feature-extraction branches, the cross-attention fusion mechanism, the conditional anomaly detection head, and the source localization component. Chapter 5 details the experimental setup, including the training protocols, baseline configurations, and evaluation metrics. Chapter 6 reports the experimental results, which are discussed in Chapter 7 alongside the limitations of the prototype setting and directions for future work. Chapter 8 concludes the thesis.

Chapter 2

Related Work

2.1 Predictive Maintenance in Hydropower

Hydropower maintenance was historically dominated by corrective (run-to-failure) and time-based preventive maintenance (TBM) frameworks, both calibrated to the steady-state baseload operations that characterized the twentieth-century hydroelectric fleet [6]. The accelerating integration of VRE has restructured this context. PSH plants now execute frequent start-stop cycles, sustain prolonged operation in deep part-load regimes, and provide rapid frequency regulation and reactive power support on demand,[61] driving the industry toward condition-based maintenance (CBM) and predictive maintenance (PdM) under the broader Maintenance 4.0 paradigm [10].

The efficacy of any predictive model depends on how faithfully it captures the underlying hydro-mechanical degradation physics [86]. In Francis pump-turbines, operation below the Best Efficiency Point (BEP) generates a highly energetic precessing vortex rope in the draft tube, with a sub-synchronous precession frequency of approximately $0.25f$ to $0.35f$ (where f denotes the rotational frequency) [135]. The resulting low-frequency pressure pulsations can trigger acoustic resonance with the hydraulic circuit, imposing severe cyclic fatigue on the runner blades and shaft. In parallel, Rotor-Stator Interaction (RSI) arises from the periodic pressure disturbance encountered by each runner blade as it traverses the wake of the stationary guide vanes. The associated blade passing frequency is defined as $f_b = Z_R \cdot f$, where Z_R is the number of runner blades,[31] and prolonged exposure to RSI-induced pressure pulsations leads to high-cycle fatigue cracking at the blade roots [31]. Transient phase transitions between pumping and generating modes introduce water hammer effects alongside rapid reversals of hydraulic axial thrust. These reversals can destabilize the hydrodynamic oil film in the journal and thrust bearings, driving them into boundary lubrication regimes [110]. Each of these degradation mechanisms

produces a distinct acoustic and vibratory signature. The shifts in those signatures across operational states define the domain-shift problem that any production-grade diagnostic system must solve.

To capture these degradation mechanisms, diagnostic systems must combine synchronized data from multiple types of sensors [47]. Vibration is the most foundational measurement. Non-contact proximity probes measure the relative radial displacement of the rotating shaft, while accelerometers mounted on the bearings measure the structural acceleration of the surrounding assembly [61].

In modern pumped-storage hydropower (PSH), operating conditions vary continuously, so fixed warning thresholds produce an excessive rate of false alarms. To address this, advanced diagnostic frameworks employ Vibrational Hill Charts (VHCs) [135]. VHCs use artificial neural networks to derive adaptive warning thresholds conditioned on two physical variables: water pressure (net head H) and valve position (wicket gate opening α). This state-aware approach measurably improves the model's classification accuracy (ROC AUC) [135].

This state-aware principle, adjusting the diagnostic boundary to the machine's current operating point, underpins context-conditioned diagnostics. A natural extension, rather than relying on manually selected physical variables such as H and α, is to learn the operating context directly from the raw acoustic and vibration data.

While standard vibration sensors capture structural motion, they cannot detect early-stage fluid-borne phenomena. Acoustic Emission (AE) sensors address this gap by responding to high-frequency elastic waves emitted as micro-cracks form and cavitation bubbles collapse [110]. Under normal conditions, acoustic energy remains concentrated in lower frequency bands. At the onset of cavitation, however, this energy rises sharply across a broad frequency range, and its intensity increases with the physical size of the cavitation event [110].

To complete this sensory profile, dynamic pressure transducers are installed throughout the turbine's water passages, specifically in the spiral casing, the inter-vane space, and the draft tube cone. By measuring the amplitude of water pressure pulses, these sensors help distinguish whether a measured vibration originates from the water flow itself or from a structural mechanical fault [61].

In addition to high-rate acoustic and vibration data, slowly varying operational variables, such as oil temperature and water pressure, provide essential supervisory signals. As lubricating oil degrades, for instance, bearing temperatures rise steadily,

providing a long-term indicator of mechanical health that fast-acting vibration sensors do not capture [61]. Similarly, water pressure measurements in the turbine housing reveal the actual hydraulic load on the machine, which again helps determine whether an unusual vibration is driven by the flow or by a mechanical fault [110].

While prior predictive maintenance research has used these operational variables as static inputs for fault classification,[135] their potential as dynamic conditioning signals for deep anomaly detection models remains largely unexplored [61]. This gap motivates the fourth research question of this thesis: whether fusing real-time temperature and pressure data into the model's latent context can reduce false positive rates beyond what acoustic and vibration data achieve alone.

Because these multimodal sensor streams are high-dimensional and non-stationary, predictive models rely on time-frequency representations such as the Short-Time Fourier Transform (STFT) and the Continuous Wavelet Transform (CWT), which track how spectral content evolves during rapid operational changes [92]. Furthermore, because machine failures in hydropower are rare, labeled fault data are scarce, and the standard approach is therefore unsupervised anomaly detection [120]. Under this paradigm, a model learns the characteristics of healthy operation from normal data alone and flags departures from that learned baseline at inference time. Section 2.4 reviews the architectures used for this task, including reconstruction-based autoencoders and density-based methods.

Two studies conducted with illwerke vkw AG at the Rodundwerk II facility are the immediate predecessors of this work. Khamaisi et al. benchmarked three unsupervised acoustic models, a Long Short-Term Memory Autoencoder, K-Means clustering, and a One-Class SVM, on real-world data from ROW II, reaching near-perfect steady-state accuracy (ROC AUC up to 0.998) while exposing the transition-induced false-positive failure introduced in Chapter 1 [65]. From that failure they identified the need for a multimodal framework integrating acoustics, vibration, and process variables such as rotational speed and bearing temperature.

Finally, Haller and Länzlinger developed the *pysoundlocalization* library, which uses time-difference-of-arrival (TDOA) techniques to locate sound sources in noisy environments [40]. They validated this pipeline in simulations based on the ROW II sensor layout. Their work establishes the classical acoustic front-end that serves as one component of the learning-based localization framework proposed here.

The remainder of this chapter reviews the sensing, fusion, anomaly-detection, and localization literature on which these challenges bear.

2.2 Acoustic and Vibration Sensing for Fault Detection

Among the sensing methods used to monitor hydraulic machinery, microphones (airborne acoustics) and accelerometers (structural vibrations) provide the richest diagnostic information [59]. Although both capture the dynamic behaviour of the machine, they do so through different physical channels: microphones register pressure fluctuations propagating through the air, whereas accelerometers measure motion directly on the machine casing [101]. Their respective strengths, limitations, and data requirements are reviewed below. Section 2.3 then examines the fusion architectures that combine the two modalities to exploit their complementary nature.

2.2.1 Airborne Acoustic Sensing

Airborne acoustic sensing is non-invasive and covers a wide spatial extent, which makes it well suited to monitoring large-scale infrastructure where instrumenting every component is either too costly or geometrically infeasible [59]. A distributed network of microphones captures sound from multiple directions and, through time-difference-of-arrival (TDOA) techniques, can attribute specific acoustic events to their physical origin on the machine [17]. The role of these TDOA features in fault localization is examined in Section 2.5.

TDOA across multiple microphones is, however, acutely sensitive to timing: a synchronization error of one millisecond corresponds to roughly 0.34 m of location error for sound traveling through air, and several times more along the faster structure-borne path through a plastic casing [40]. Maintaining channel alignment well within that tolerance is therefore a prerequisite for any TDOA-based localization.

The acoustic spectrum of a hydraulic turbine contains several distinct fault signatures. One of the most critical is cavitation, which occurs when the local water pressure drops below the fluid's vapor pressure, causing vapor-filled cavities to form and collapse abruptly and generating shockwaves and micro-jets against the turbine blades [29]. The onset of cavitation appears acoustically as an abrupt rise in broadband noise that extends into the ultrasonic range [26]. During partial-load operation, rotating vortices in the water channel further modulate this noise, producing pulsations tied to the rotational speed [38]. Importantly, microphones can detect early-stage vaporous cavitation before it

produces measurable structural vibration or triggers standard SCADA-level performance alarms [114].

Beyond cavitation, mechanical and structural faults also produce distinct acoustic signatures. As water flows past the trailing edges of the guide vanes, vortex shedding generates tonal noise [99]. Under abnormal operating conditions these frequencies can lock in with the acoustic resonance of the draft tube, inducing severe vibration and accelerated metal fatigue [99]. Similarly, propagating micro-cracks and bearing-surface spallation emit short, high-energy acoustic bursts that propagate through the machine casing and into the surrounding air [12].

Despite this sensitivity, relying on acoustic sensing alone in a pumped-storage plant is difficult owing to the physical environment [98]. Enclosed concrete machine halls are strongly reverberant: reflections from the walls, floor, and ceiling create multipath propagation that smears the sharp transients of early-stage faults and distorts the frequency spectrum [18]. Moreover, the target fault sounds are often masked by background noise from auxiliary equipment, cooling fans, and adjacent operating turbines [98]. Together, these environmental constraints indicate why airborne acoustic sensing alone is insufficient and why it must be fused with structural vibration data.

2.2.2 Vibration Sensing with Accelerometers

Accelerometers provide a direct, physical measure of machine health. They transduce the motion of the casing, bearing pedestals, and guide vanes into electrical signals, allowing diagnostic systems to identify internal mechanical forces, mass unbalance, and structural resonances [2, 109]. In hydraulic turbines, these sensors are typically installed in triaxial configurations (measuring all three spatial axes) across the guide bearing, thrust bearing, and turbine head-cover.

Microphones and accelerometers have complementary sensitivities. Low-frequency mechanical faults couple poorly from thick steel casings into the surrounding air, so microphones largely miss them whereas accelerometers detect them readily [131]. Conversely, the heavy steel casing attenuates the high-frequency vibrations produced by early-stage cavitation or incipient bearing cracks; these signatures attenuate before reaching the external accelerometers, yet radiate into the air where microphones register them [105]. In short, acoustic sensors excel at high-frequency, early-stage fluid-borne phenomena, while vibration sensors provide direct evidence of low-frequency, established mechanical wear [30]. This complementarity is precisely what a fused acoustic-vibration localizer

can exploit, using structure-borne timing to resolve location ambiguities that acoustics cannot resolve alone.

Once the vibration data are acquired, diagnostic systems analyze them against established frequency patterns. A strong component at the rotational speed (1X) typically indicates mass unbalance. Components at twice or three times that speed (2X and 3X), particularly in combination with axial motion, indicate shaft bend or misalignment. Fractional components (such as 0.5X or 1.5X) indicate mechanical looseness in the assembly [3].

Beyond mechanical rotation, the Blade Passing Frequency (BPF) is a key indicator of flow health. The BPF is the product of the rotational speed and the number of turbine blades; an abnormal BPF amplitude indicates flow restrictions, blade damage, or friction between rotating and stationary components [111]. When the BPF component begins to pulse together with lower-frequency instabilities during partial-load operation, it is an early indicator of escalating flow-induced problems [88].

Finally, sub-synchronous components below the running speed serve as warnings of severe faults. Lubricating-oil instabilities (oil whirl and whip), for instance, occur between 0.4X and 0.48X [82]. The draft tube vortex rope discussed in Section 2.1 likewise exhibits strong activity between 0.2X and 0.4X, imposing cyclic loading that can fracture the turbine shaft through metal fatigue [127].

2.2.3 Limitations of Each Modality in Isolation

Systems that rely on acoustic sensing alone are limited primarily by background noise and changing acoustic environments. Because standard microphones are sensitive to sound from all directions, they readily capture interfering noise from nearby machines and auxiliary equipment [128]. In addition, enclosed concrete halls are strongly reverberant, and this reverberation smears the sharp transients of early faults into the continuous background [17]. Finally, acoustic propagation in a room varies with temperature, humidity, and equipment placement, so an acoustic-only model trained on historical data often fails on deployment under altered environmental conditions [41].

Systems that rely on vibration sensing alone are, conversely, limited in detecting small or early-stage faults. The high-frequency, low-energy vibrations produced by early mechanical wear are attenuated by the heavy casing and bolted joints before reaching the external accelerometers [105]. This produces a high false-negative rate, in which the system misses the earliest signs of degradation [108]. Accelerometers are, moreover,

largely insensitive to purely fluid-borne phenomena (such as early cavitation) until they grow severe enough to excite the metal structure mechanically [131].

Both limitations are aggravated when the machine changes its operating state. When a reversible pump-turbine switches from generating to pumping, for example, the baseline acoustic and vibratory signature shifts substantially [69]. Comparable shifts occur even in laboratory prototypes through changing ambient noise or minor rig reconfigurations. A diagnostic system that relies on a single modality and lacks broader operational context will systematically misclassify these benign transitions as faults, producing a false-alarm rate that renders the monitor unusable for plant operators [118]. This failure was confirmed directly at ROW II by Khamaisi et al. [65].

2.2.4 Empirical Evidence for Multimodal Fusion

The physical differences between the two sensor types translate into measurable improvements when their data are combined. In a study on wind turbine bearings, models using vibration data alone achieved about 86.5% accuracy and acoustic-only models about 79.1%; fusing the two streams raised the accuracy to 99.2% in clean conditions and maintained above 95% under severe background noise, a 16-percentage-point improvement over the acoustic-only model [22]. Studies on other rotating machinery report that sensor fusion routinely pushes overall accuracy past 99.5% [42]. Notably, one study on electric motors found that adding microphone data to a vibration network reduced the false-positive rate to zero, eliminating an entire class of spurious alarms [68].

Fused models are also more robust under difficult real-world conditions. They have been reported to maintain classification accuracy above 94% even with very limited training data (as few as 20 samples per fault) and when the background noise exceeds the target signal (signal-to-noise ratios as low as -10 dB) [37].

In aggregate, these results indicate that fusing acoustic and vibration data yields a diagnostic system whose performance exceeds that of either modality alone. The deep learning architectures used to achieve this fusion are reviewed in Section 2.3.

2.2.4 Feature Extraction

Raw acoustic and vibration signals are high-rate, non-stationary data streams that must be reduced to informative features before a learning model can use them. Because hydropower machines change operating state frequently, these signals shift substantially over time, so static, single-resolution analysis is insufficient [74].

For audio, the standard approach is the Short-Time Fourier Transform (STFT), which produces a time-frequency representation, the spectrogram [122]. The STFT imposes a fixed trade-off between time and frequency resolution. The Continuous Wavelet Transform (CWT) mitigates this through a multi-resolution analysis, resolving both brief high-frequency transients and slower low-frequency components,[73] which makes CWT scalograms informative inputs for modern neural networks [35].

Mel-spectrograms and CWT scalograms are the representations most commonly supplied to neural fault detectors,[73, 81] and Mel-Frequency Cepstral Coefficients (MFCCs) remain a compact perceptual baseline [123]. A relevant caveat from the ROW II benchmarks is that MFCC and Mel-spectrogram inputs, while effective during steady operation, degrade systematically during mode transitions [65].

For vibration data, diagnostic systems rely on established statistical descriptors. The Root Mean Square (RMS) quantifies the overall energy of the motion, whereas Kurtosis is far more sensitive to early-stage damage, rising sharply when incipient cracks or pits produce impulsive shocks well before the overall vibration amplitude increases [11]. Envelope analysis, based on the Hilbert transform, removes the dominant low-frequency flow-induced vibration to expose the high-frequency periodic signature of damaged mechanical components [12].

Because the appropriate impulsiveness statistic depends on the available sample rate, rolling kurtosis[58] is the standard choice on high-rate streams, while a simpler crest factor remains well-defined when only a slow, decimated signal is available [35].

In summary, neither sensor type is adequate in isolation: microphones are degraded by reverberation and background noise, while accelerometers tend to miss early-stage fluid-borne phenomena. Both are, moreover, vulnerable to the distribution shifts induced when the turbine changes operating state. Existing multimodal models address the sensor limitations but assume a single, steady operating state. This gap defines the central requirement of this thesis: the diagnostic architecture must adjust its decision boundaries according to a learned representation of the machine's current operating context. The foundations for this approach are reviewed in the following section.

2.3 Multimodal Deep Learning and Sensor Fusion

Sensor data can be combined in three standard ways: early fusion (combining raw data), late fusion (combining final decisions), and intermediate fusion (combining extracted features) [14]. Early fusion is ill-suited to streams with markedly different sampling

rates: directly merging high-rate audio (needed to capture rapid cavitation transients) with much lower-rate vibration either discards the high-frequency acoustic content or incurs prohibitive computational overhead [14]. Late fusion, by contrast, processes each sensor independently and combines only the final decisions; while this is robust to the failure of a single sensor, it discards the cross-modal timing between events, for example the coincidence of an acoustic transient with a simultaneous mechanical event [128].

Intermediate fusion addresses this limitation. It first compresses each sensor stream into a compact representation and then fuses the representations at the feature level [83]. Operating in this shared space bridges the gap between airborne pressure and structural motion, allowing the model to learn cross-modal interactions and to weight whichever modality is more reliable for the current operating state [128].

To compress this data, intermediate fusion uses Convolutional Neural Networks (CNNs) adapted to the physics of each sensor. For vibration, one-dimensional CNNs operate directly on the raw time series. Acting as learned filters, their shallow layers capture brief impulsive shocks from bearing defects while deeper layers track slower flow-induced instabilities,[85] which makes 1D CNNs effective at capturing sudden mechanical impacts [85]. Acoustic signals in industrial halls are, by contrast, too corrupted by reverberation for a one-dimensional time-series model, and are therefore converted into a two-dimensional spectrogram and processed by a 2D CNN [83].

Modern diagnostic systems often pair these CNNs with attention mechanisms that selectively emphasize informative fault frequencies and suppress background noise [83]. Cross-attention extends this idea across modalities rather than within a single sensor.

A limitation of standard CNNs is their restricted receptive field, which makes it difficult to capture long-range dependencies or cross-sensor relationships [16]. Cross-attention addresses this with a Transformer mechanism: one stream forms the queries while the other provides the keys and values, yielding a scaled dot-product alignment score defined as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{2.1}$$

where d_k is the dimensionality of the key vectors. The operation quantifies the relevance of every vibration feature to every acoustic feature, coupling the two physical domains.

A particularly useful family of building blocks for this kind of fusion is the Set Transformer of Lee et al.,[70] whose Multi-head Attention Block (MAB) lets every token of one modality attend to every token of the other, and whose Pooling-by-Multi-head-Attention (PMA) module compresses a variable-size set of fused tokens into a fixed-length context vector using learned seed queries.

These building blocks are permutation-invariant and agnostic to the number of input tokens, so a model built on them can in principle process a varying number of sensors without architectural changes, a property that ordinary dual-branch designs lack.

Other researchers have proposed more elaborate multi-stage transformers for fault diagnosis, including the Twins Transformer with parallel intra-modal and cross-modal branches,[71] the Domain-Collaborative Multimodal Transformer that projects deep CNN-BiLSTM features into a shared embedding space,[66] and multi-scale spatiotemporal attention pyramids [72]. These designs add considerable capacity, but most cannot accommodate a changing number of sensors without retraining, a limitation that permutation-invariant pooling avoids.

2.3.1 Unsupervised Operational Context Learning

A fused representation alone is insufficient when the same fault produces different sensor signatures under different operational states. The network must additionally learn a representation of the current operational context without requiring explicit state labels, which are frequently unavailable or misaligned with high-frequency sensor streams during rapid transitions [79].

Self-supervised contrastive learning addresses this by training encoders to maximize representational similarity between augmented views of the same temporal window while minimizing similarity between temporally separated windows, forcing the model to learn the invariant physics of each operating regime rather than superficial noise [84]. Masked segment reconstruction provides a complementary self-supervised signal by forcing the network to predict masked portions of a feature sequence from the surrounding context, capturing the baseline periodicity and harmonic structure of the current machine state.

2.3.2 Context Conditioning Mechanisms

Once the network has inferred the machine's current operating state (the context vector), this information must be propagated to the anomaly detection head. Feature-wise Linear Modulation (FiLM) is the most widely used mechanism for this purpose [95]. It applies

a context-dependent affine transformation ($\tilde{h}_i = \gamma_i h_i + \beta_i$) parameterized by the context vector $\mathbf{c}$: the predicted scale γ_i rescales each feature and the shift β_i offsets it, enabling the network to amplify regime-relevant components and attenuate irrelevant ones according to the current operating state.

Other context-aware mechanisms exist alongside FiLM. Conditional normalization replaces fixed normalization layers with layers whose statistics shift to match the machine's current load [27]. Conditional Variational Autoencoders (CVAEs) feed the context into both the encoding and decoding stages of the network [62]. Conditional normalizing flows, reviewed in Section 2.4.2, offer a further alternative that yields exact probability scores rather than the approximations used by CVAEs.

Finally, the CANDE-CP architecture illustrates the practical value of context conditioning. The system infers the machine's operating state from unlabeled data, conditions a decoder on it through FiLM layers, and smooths the resulting predictions over time. In doing so it forms an adaptive decision boundary that tightens and relaxes with the operating state, substantially reducing false alarms in multi-state environments and demonstrating the operational value of context conditioning [102].

The final major challenge identified in the literature is domain shift, which occurs when a machine changes its operating state and alters its baseline sensor data. To address it, models such as Multi-Adversarial Attention Networks (MAAN) use adversarial training, pitting network components against one another so that the learned fault representations remain invariant to rotational speed, valve angle, and water pressure [104]. Other approaches, such as the Multi-Attention Meta Transformer (MMT-FD), use meta-learning: through aggressive augmentation of unlabeled data during training, these models learn representations that generalize to operating conditions unseen at training time [124].

The literature reviewed above establishes that cross-attention transformers, self-supervised context learning, and FiLM conditioning each address a distinct, critical challenge in industrial fault detection. These three techniques have not, however, previously been combined for multi-state pumped-storage hydropower.

Specifically, the field lacks a unified, label-free architecture capable of handling the operational shifts of reversible pump-turbines. No existing system uses a cross-attention transformer to fuse audio and vibration data into a single context vector that simultaneously conditions both an anomaly detector and a source localizer, and no existing framework incorporates slowly varying supervisory signals (such as bearing temperature and hydraulic pressure) into that context pathway. This leaves an open

question: can these low-bandwidth variables be used to further reduce false alarms within a deep detection boundary? Closing these gaps constitutes the primary architectural contribution of this thesis.

2.4 Anomaly Detection Methods

In large-scale hydropower, severe machine failures are rare, so historical sensor databases consist almost entirely of healthy operational data and lack the balanced fault examples needed to train standard supervised models [46]. Supervised models can, moreover, only recognize fault types present in their training data, which leaves them insensitive to new or unanticipated forms of damage [64].

For these reasons, the standard approach in this setting is unsupervised anomaly detection. A model is trained exclusively on healthy data to learn the boundaries of normal machine behavior, and any sensor reading that falls outside these boundaries at inference time is flagged as anomalous [91].

2.4.1 Reconstruction-Based Methods

Reconstruction-based anomaly detection rests on the principle of compression. A network is trained to compress and reconstruct healthy sensor data; when it encounters data unlike anything seen in training, the reconstruction is inaccurate, and the magnitude of this reconstruction error serves as the anomaly score [91]. Standard autoencoders (CNN-AEs) operate on spectrogram representations. While effective in steady conditions, they tend to memorize the training sequences and adapt poorly to the benign baseline shifts that accompany turbine aging [19].

Variational Autoencoders (VAEs) address this rigidity by learning a smooth, continuous latent distribution rather than memorizing individual data points [78]. This flexibility suppresses false alarms caused by benign background variation, but it introduces a known side effect, over-smoothing: because VAEs effectively average over the data to construct this latent space, they behave as a low-pass operation, attenuating sharp transients and thus missing the high-frequency indicators of early-stage structural damage [78]. For sequential data, LSTM Autoencoders offer a complementary advantage: through recurrent state, they model long-range temporal dependencies, which makes them sensitive to rhythmic phenomena such as a localized acoustic burst recurring at a fixed phase of each rotation, or the slow pulsation of a draft tube vortex rope [65].

In a pumped-storage setting, however, all three architectures share a critical vulnerability: domain shift. A model trained primarily while the machine is generating (Turbine mode) is optimized for that specific operating state. When the machine reverses into Pump mode, the flow and mechanical loads change, the sensor data diverge from the training distribution, the reconstruction error inflates, and the system produces a burst of false alarms [134].

For this reason, standard reconstruction models are largely unsuitable for variable-state environments unless they are explicitly conditioned on the operating context [78]. This limitation was observed directly in the ROW II benchmarks introduced in Chapter 1 [65].

Context-aware architectures address this limitation directly. The CANDE-CP system (Context-Aware Novelty Detection AutoEncoder with Context Prediction), for example, infers the operating state from unlabeled data and conditions the decoder on it through FiLM layers, allowing the decision boundary to shift with the operating state [102]. To accommodate the transient sensor noise that accompanies a mode change, the system applies temporal smoothing (a rolling exponential-decay queue), yielding a boundary that tightens during steady operation and relaxes during transitions, without manual intervention [102]. This capability directly motivates the anomaly detection design proposed in this thesis.

The DCASE Task 2 benchmark (a major competition for unsupervised machine sound detection) provides extensive empirical evidence of the domain-shift problem. Across the 2021–2025 editions, standard reconstruction algorithms consistently exceed 90% accuracy on their source domain but degrade sharply when evaluated on data from different operating conditions [63]. Indeed, a persistent negative correlation has been observed between the two: the more closely a model fits the source domain, the worse it generalizes to shifted conditions [87]. Consequently, the most successful competition entries have moved away from static reconstruction toward context-aware adjustment and metadata conditioning [19].

It should be noted that the DCASE competition concerns small industrial components such as fans and valves, which operate at a power scale far below that of a 295 MW Francis pump-turbine. While the absolute accuracy figures are therefore not directly comparable, the underlying domain-shift problem is identical, and the need for context-aware architectures carries over to the ROW II machines, as confirmed by the Khamaisi experiments [65].

2.4.2 Density-Based and Normalizing Flow Methods

Density-based anomaly detection takes a fundamentally different approach. Rather than measuring how poorly a network reconstructs an input, it estimates the probability that a given sensor reading belongs to normal operation [96]. This yields an interpretable probability score that measures statistical rarity directly, removing the need to select an arbitrary error threshold [94].

The One-Class SVM (OC-SVM) is the most established algorithm of this type in industrial monitoring. It projects the data into a high-dimensional space and constructs a boundary (a hyperplane) enclosing the healthy data; anything outside that boundary is classified as anomalous [121]. As noted earlier, the Khamaisi study reported near-perfect accuracy (0.998 ROC AUC) for the OC-SVM on steady-state ROW II data, rivaling deep models such as LSTM-AEs [65]. The OC-SVM nonetheless has two significant drawbacks for this setting. First, its computational cost scales quadratically with dataset size, which is prohibitive for continuous high-rate audio and vibration streams. Second, its static boundary cannot adapt to mode transitions (domain shift) without extensive manual feature engineering [121].

Normalizing flows overcome these limitations through a sequence of invertible transformations that map a simple base distribution to the complex distribution of the observed sensor data [106]. Because every step is invertible, the network can compute the exact log-likelihood of any incoming reading, a key advantage over autoencoders [106]. Whereas an autoencoder may reconstruct an out-of-distribution noise pattern accurately and thus score it as normal, a normalizing flow assigns such a pattern the low likelihood it warrants during healthy operation [106].

To make likelihood evaluation efficient enough for real-time monitoring, several architectures exploit structural simplifications. The RealNVP architecture partitions the variables and uses one partition to transform the other, which keeps the Jacobian tractable and the likelihood inexpensive to evaluate [80]. The Glow architecture extends this with learnable invertible convolutions, enabling channel mixing across audio and vibration features at every layer [116]. In both, the diagnostic principle is the same: the onset of a fault produces an immediate drop in the assigned likelihood [60].

In environments where the machine changes operating mode frequently, Conditional Normalizing Flows (CNFs) provide an effective solution. Rather than a single fixed density, CNFs incorporate the operating-context vector directly into the transformation, allowing the model to maintain a state-specific density [62]. Studies on complex time-

series data confirm that CNFs adapt as the operating state changes, treating a given signal as normal under one regime while flagging the identical signal as anomalous under another [36].

One practical difficulty nonetheless remains: training flows directly on raw, high-rate sensor streams (such as high-speed audio and vibration) is numerically unstable. The standard remedy is to apply a lightweight flow to the compressed latent representations produced earlier in the network rather than to the raw signal itself [129].

Synthesizing this literature yields a clear set of design rules for the anomaly detection system built in this thesis. Reconstruction methods model temporal structure well but trigger an intolerable number of false alarms during mode transitions unless explicitly conditioned on the operating context. Normalizing flows, by contrast, provide exact anomaly scores and naturally support such conditioning; because they are unstable on raw data, they must be applied to compressed features.

Crucially, neither approach has been evaluated with context-conditioned, multimodal sensors in a pumped-storage hydropower plant, and neither has examined whether adding simple temperature and pressure signals improves the context guidance.

The anomaly detection head in this thesis must therefore satisfy a strict set of requirements. It must remain calibrated across distinct operating modes without relying on manual labels, and it must accommodate slowly varying supervisory process variables (such as temperature and pressure) to suppress residual false alarms that the audio and vibration sensors cannot resolve alone.

2.5 Sound Source Localization

Once a diagnostic system has confirmed an anomaly, engineers must determine its physical location in order to act on it. The physical environment of the ROW II machine hall makes this attribution problem particularly difficult, for two reasons.

First, the enclosed concrete turbine housing is strongly reverberant. Each transient event, such as a mechanical crack, produces a cascade of reflections that reach the microphones from spurious directions, corrupting the timing differences on which classical localization relies to triangulate the source.

Second, the continuous operation of nearby auxiliary equipment introduces substantial background noise whose spectral content often overlaps that of genuine turbine faults. This produces a continually shifting balance between the target signal and the interference, degrading the system's ability to localize the source across operating states.

Together, these reflections and overlapping noise create conditions in which a dominant direct path rarely exists. Because classical localization assumes free-field propagation, these estimators fail systematically in this setting. The following subsections review classical time-difference (TDOA) methods and their learning-based successors, and define the open localization questions this thesis addresses.

2.5.1 Time-Difference-of-Arrival Methods

By measuring the difference in arrival time of a sound at two microphones, a system can constrain the source to a curved surface in three dimensions. Intersecting these surfaces across multiple microphone pairs triangulates the source location [34].

The standard tool for estimating this delay is the Generalized Cross-Correlation with Phase Transform (GCC-PHAT). The phase transform whitens the cross-spectrum so that stationary background noise does not dominate the target signal, sharpening the correlation peak that marks the arrival-time difference [34]. Steered-Response Power with Phase Transform (SRP-PHAT) extends this to entire microphone networks: it evaluates a discrete three-dimensional grid of candidate locations against all microphone pairs, producing a steered-power map whose maximum is taken as the source location. Because it aggregates evidence across all sensors, it is more robust to individual corrupted measurements caused by reflections [33].

In highly reverberant environments such as a concrete turbine hall, however, both methods share a fundamental limitation. Reverberation and interference populate the steered-power map with spurious maxima, and without a strong prior on the source location the true peak cannot be reliably distinguished from reflection artifacts [33]. Reliably localizing multiple concurrent sources in a cluttered, reverberant industrial environment therefore remains an open research problem [132].

Haller and Länzlinger recently implemented this classical pipeline in a software toolbox. In simulation they showed that for a single source the method is highly accurate, achieving a mean localization error below one centimeter (0.009 m) with four microphones in a noisy environment [40].

When multiple overlapping sources (polyphonic conditions) were introduced, however, the classical method degraded sharply: lacking a source-separation stage, the mean error rose to 9.46 m, with worst-case errors exceeding 136 m in the same simulated room [40].

This degradation makes a key point: classical TDOA methods alone are inadequate for the multi-source environment of the real ROW II machine hall, where the main turbine, auxiliary fans, governor pumps, and compressors emit sound concurrently, producing overlapping interference that these methods cannot resolve.

In structural health monitoring, acoustic emission TDOA methods mount piezoelectric sensors directly on a structure to record impulsive elastic waves from internal damage events, combining pairwise TDOAs with the known wave velocity to triangulate the damage origin [136]. Closed-form velocity-free formulations using the complete set of pairwise TDOAs have demonstrated robustness to TDOA noise,[136] and jointly estimating both source coordinates and propagation velocity from TDOA measurements has been shown to approach the Cramér-Rao lower bound [137]. Accelerometers mounted on the ROW II turbine housing record structure-borne elastic waves from fault-induced impulsive excitations in a manner directly analogous to AE sensors, making these methods relevant to the vibration-sensor component of the localization problem.

The nine microphones and four accelerometers installed at ROW II have complementary timing characteristics, which makes them well suited to joint fusion. Sound travels through air at roughly $c \approx 343$ m/s, whereas elastic waves travel through solid steel at roughly $c \approx 5{,}000$ m/s. Because steel transmits elastic energy about fourteen times faster than air, accelerometers resolve the onset of a mechanical impact with high temporal precision.

Airborne microphones, by contrast, benefit from the lower speed of sound in air, which produces larger and more easily measured time delays between sensors; as established above, however, these microphones are highly susceptible to reverberation. Fusing both sets of timing measurements allows the precise structure-borne arrivals of the accelerometers to disambiguate the reverberant acoustic measurements, localizing faults that neither modality resolves alone.

Ultimately, classical estimators are fundamentally limited in this setting. They presuppose a dominant direct path; because the ROW II environment violates this assumption, classical estimators fail. This motivates the shift toward the learning-based approaches discussed next.

2.5.2 Learning-Based Localization

A major survey by Grumiaux et al. reviewed over two hundred learning-based localization systems. It found that a network trained on recordings from a given room implicitly

learns that room's reverberation and noise characteristics, compensating for the very distortions that defeat classical TDOA estimators [34]. CNNs operating on raw correlation features, for instance, roughly doubled the accuracy of classical methods in highly noisy environments [34]. Convolutional Recurrent Neural Networks (CRNNs) improved on this further: their recurrent state regularizes the per-frame estimates over time, halving the mean localization error relative to earlier algorithms [1].

The Cross3D architecture demonstrates the value of combining classical physics with learned models. Rather than discarding the SRP-PHAT steered-power maps, Cross3D feeds them into a three-dimensional network, retaining the geometric prior of the classical estimator while gaining the robustness of a learned model; under heavy reverberation it continues to track the source and consistently outperforms models trained on raw audio alone [21]. This hybrid design, combining classical spatial maps with deep networks, is the direct inspiration for the localization system proposed in this thesis.

More recently, Transformer-based architectures have set new accuracy records for localization in reverberant rooms. Rather than relying on recurrent state, Transformers use self-attention to model the entire temporal context jointly, capturing the long-range acoustic patterns that carry essential location information [55]. Networks equipped with channel attention have likewise proven effective, learning to suppress the frequency bands most corrupted by reverberation [45]. A 2025 review identified multimodal sensor fusion as the most promising route for localization problems that microphones alone cannot resolve, concluding that combining attention mechanisms with multiple sensor types is key to accurate localization in loud, highly reverberant environments [126].

Throughout the machine diagnostics literature, networks combining microphones and accelerometers consistently outperform single-sensor systems [125]. A substantial gap nonetheless remains in application: existing work largely reduces localization to classifying the faulty component, rather than estimating the three-dimensional coordinates of the fault within a large turbine housing.

To date, no published study has built a system that uses fast-traveling structure-borne timing (accelerometer TDOA) to correct the localization errors of an acoustic-only baseline. Nor has any study run an ablation in a real machine hall to quantify the spatial-accuracy gain from adding vibration sensors. Finally, no learning-based localization method has been evaluated in a real pumped-storage hydropower plant, an environment characterized by multiple operating modes, sensors distributed across different elevations, severe reverberation, and heavy background interference.

Discussion

The preceding sections have identified four research gaps, corresponding directly to the four research questions stated in Chapter 1.

First, domain shift in pumped-storage plants is well documented but remains unsolved. Current diagnostic systems assume a single, steady operating state, and their accuracy degrades sharply when the machine transitions between operating modes, a failure confirmed directly at the ROW II plant by Khamaisi et al. [65].

Second, while cross-attention transformers and self-supervised learning each solve part of the sensor-fusion problem, they have not been combined. The literature reviewed in Section 2.3 lacks a unified, label-free architecture capable of handling the extreme environment of a multi-mode pumped-storage plant.

Third, the literature reviewed in Section 2.4 shows that anomaly detection must adapt to the machine's current context, and that normalizing flows applied to compressed features provide an exact, principled means of doing so. No existing anomaly detection method, however, incorporates slowly varying supervisory signals such as temperature and pressure to further reduce false alarms, an integration the prior literature explicitly calls for.

Finally, the localization literature reviewed in Section 2.5 shows that learning-based approaches can overcome the reverberation that defeats classical estimators. No study, however, has combined the fast structure-borne timing of accelerometers with the acoustic timing of microphones to estimate the three-dimensional location of a fault inside an active turbine housing.

The unified architecture proposed in the following chapter is designed to close all four of these gaps simultaneously.

Chapter 3

Design

This chapter describes the prototype data this thesis is built on and the preprocessing pipeline used to prepare it. The data's provenance, the move from the ROW II machine to the bench-top prototype, and the five recording campaigns (D1 to D5) are introduced in Chapter 1; this chapter takes that corpus as given and documents how it is prepared.

The focus is on the data as received and on the steps that turn it into the tensors the model reads: how a mixed, incrementally collected prototype corpus is ingested, normalized, synchronized, and windowed into a single uniform format. The model itself, how it is trained, and how it is evaluated are covered in later chapters.

The operating mode matters throughout preprocessing because the machine's healthy signature changes with it (Chapter 1). The pipeline is therefore built so that healthy operation is sampled at several operating points, and the operating mode is carried through to each window rather than averaged away. The chapter moves from the prototype and its sensors (§3.1, §3.2), through the five campaigns that make up the corpus (§3.3), to the pipeline that produces the per-window tensors (§3.4).

3.1 The Prototype

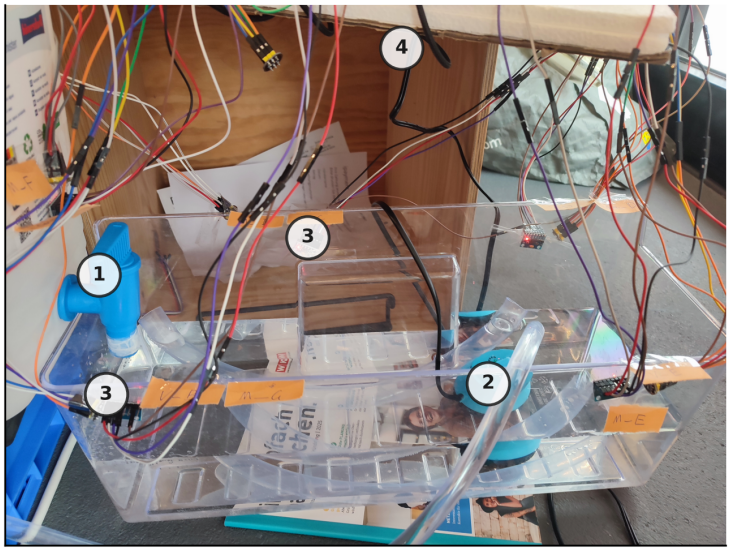
The prototype replicates the ROW II casing geometry at a reduced scale. The acoustic analysis grid (§3.4.3) is selected empirically on the prototype recordings and is, in addition, fine enough to resolve the characteristic tones of the full-scale machine, so that the same preprocessing would transfer to a field deployment without re-tuning; the reduced-scale prototype is not claimed to reproduce those tones itself.

The corpus was recorded on two different bench-top rigs, which are kept separate because they use different coordinate frames. The first two campaigns (D1, D2) used a rectangular rig whose sensors span a $15.5 \times 41 \times 24\,\text{cm}$ corner-origin frame. The last three (D3, D4, D5) used a circular 3D-printed casing about 10 cm across, with the sensors

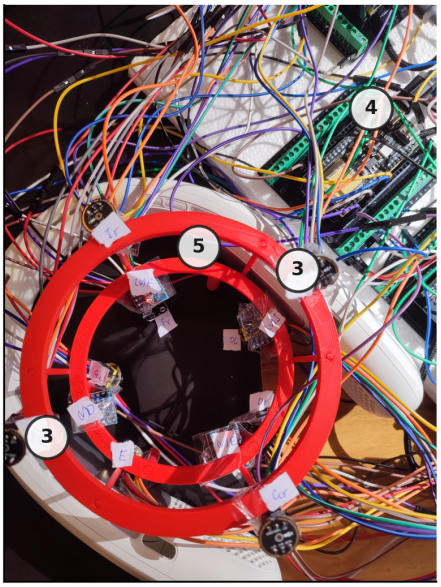
inside an $11 \times 10 \times 7\,\mathrm{cm}$ envelope and the deliberately-evoked knock positions placed around it, spanning $x \in [-20, 22]$ and $y \in [-25, 4]\,\mathrm{cm}$. Several knock positions therefore sit outside the sensor footprint.

The rig is roughly 1:110 of the full machine's casing radius ($R \approx 5.5\,\mathrm{m}$). The speed of sound is taken as $343\,\mathrm{m\,s^{-1}}$ at about $20\,^{\circ}\mathrm{C}$, and the structure-borne wave speed in the plastic casing is treated as a single fixed constant, so every distance and delay in the pipeline is expressed in physical units. All spatial labels in this thesis are reported in meters on this prototype scale. The single-constant model is a deliberate first-order approximation. A fused-deposition print is anisotropic along its layer direction and its thin, curved shell is geometrically complex, so no single bulk speed is exact; moreover, flexural waves in such a shell are dispersive, with group and phase velocity that vary with frequency, which is a known confound for time-difference-of-arrival estimation in plates. The constant is therefore used only as a first-order propagation model; per-part and dispersion-aware calibration is left to future work, and the consequence of this choice for localization, namely that an over-estimated wave speed compresses the time-difference geometry and suppresses the vibration timing channel, is reported in Chapter 6.

(a) rectangular rig (D1, D2)



(b) circular 3D-printed rig (D3-D5)



1 inlet valve 2 pump / drive 3 sensor breakouts on the casing (mics + accelerometers)
4 acquisition boards (shared trigger) 5 circular casing (~10 cm)

Figure 3.1: The two bench-top prototypes. (a) The rectangular rig of D1/D2; (b) the circular 3D-printed rig of D3–D5 (∼10 cm across) with its nine microphones and four accelerometers. Parts are keyed below the panels; the handwritten tape labels are the sensor IDs of the `position.json` registry.

How the rig was driven changed as the corpus grew, and that evolution runs through §3.3. The first two campaigns ran the rig through the three steady regimes and labeled each recording PUMP, STANDSTILL, or TURBINE. The last three held the runner in a single, unrecorded regime and instead varied an added background-noise fan, trading mode labels for denser anomaly coverage and, from D4 on, full raw-waveform vibration. As a result, operating mode is labeled on only two of the five campaigns, which constrains how the data can be used and shapes several of the choices in the pipeline below.

3.2 Sensor Array and Data Acquisition

Every recording carries two sensor modalities placed around the casing: airborne acoustics and structural vibration.

Acoustic stream. The acoustic channels are wide-band single-channel microphones recording mono WAV at 16 kHz, 16-bit per channel. The number of microphones grew

from four to nine over the campaigns as the rig was instrumented, so the channel count is not constant across the corpus.

Vibration stream. The vibration channels are accelerometers on the casing wall, recorded in one of two ways.

- **Peak-amplitude mode** (D1, D2, D3). The acquisition board reports, per channel, only a timestamp plus the amplitude and dominant frequency of the strongest spectral peak it found in a short on-board window. This limits the effective vibration rate to about 3 Hz to 16 Hz, so these campaigns give only a slow, envelope-like signal and cannot recover structure-borne detail at the 43.75 Hz runner-blade-passing tone or above.
- **Raw-waveform mode** (D4, D5). The same hardware was reconfigured to stream the raw ADC samples in direct-memory-access (DMA) batches of 128 slots, of which a median of about 109 are real samples and the rest zero-padding, arriving roughly every 290 ms. This gives an effective rate of about 376 Hz on D4 and about 446 Hz on D5, with one D5 sensor running at about 471 Hz. These rates are high enough to capture every characteristic frequency in Table 3.3.

The move from peak-amplitude to raw-waveform vibration is the biggest change across the corpus, and it makes the per-dataset vibration rate a value the pipeline has to track explicitly. It feeds directly into the choice of impulsiveness statistic in §3.4.4.

File layout and timing. Each session is one directory of microphone WAVs and per-channel vibration CSVs. The metadata, namely which sensors were active, which operating mode the unit was in, and where a knock was applied, is encoded in the folder names rather than in a separate file, and a per-campaign parser reads it (§3.3). The two streams are started together by the acquisition firmware, so they share a common start time. That alignment is still checked on every recording (§3.4.2), because the bench rig has no way to enforce it.

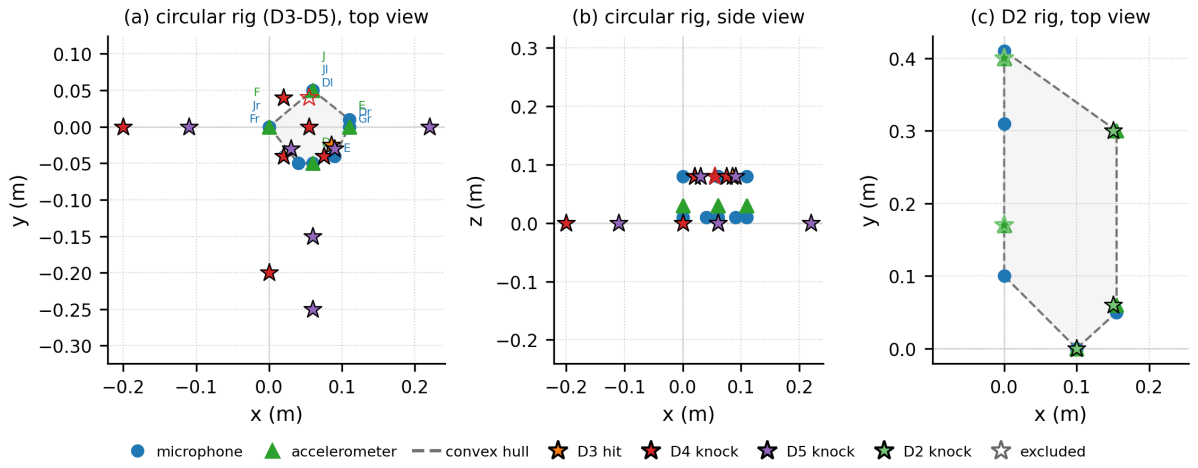


Figure 3.2: Sensor arrays and the 16 labeled knock positions (prototype frame, m). (a, b) The circular rig shared by D3–D5, top and side views: several knock positions fall well outside the dashed sensor convex hull, the geometric limit behind the per-position failures of Figure 6.10. (c) The earlier rectangular D2 rig with its three single-mode positions; open stars mark excluded recordings (dual-mode context or five-accelerometer folder).

3.3 The Five Campaigns

The corpus is the union of five campaigns, each with its own array geometry, vibration acquisition mode, label scheme, and purpose. They are summarized in Table 3.1. The microphone rate is fixed at 16 kHz across all of them and kept without resampling; the effective vibration rate differs from campaign to campaign and is resampled onto a per-dataset regular grid at ingestion (§3.4.1) so the feature extractors always see a known, uniform rate.

Table 3.1: The five bench-top prototype campaigns. *Vib mode* distinguishes the embedded peak-detection format from the raw-waveform DMA format of §3.2; *Vib eff. SR* is the effective vibration rate after ingestion has resampled the stream onto a regular grid. *Spatial* is the number of folder-encoded knock positions. The *Operating points* column separates the annotated modes of D1 and D2 from the fan-noise levels (`speed{1,2,3}`) that D3 and D4 carry instead of a mode label (§3.3.1).

Dataset	n_{mic}	n_{vib}	Mic SR	Vib mode	Vib eff. SR	Spatial	Operating points
D1	4	4	16 kHz	peak	$\approx$ 4 Hz	0	PUMP, STANDSTILL, TURBINE, anomaly
D2	5	5	16 kHz	peak	4 Hz	5	PUMP, STANDSTILL, TURBINE, anomaly
D3	9	4	16 kHz	peak	16 Hz	1	speed{1,2,3}, knock
D4	9	4	16 kHz	raw	$\approx$ 376 Hz	7	speed{1,2,3}, knock
D5	9	4	16 kHz	raw	$\approx$ 446 Hz	6	healthy, knock

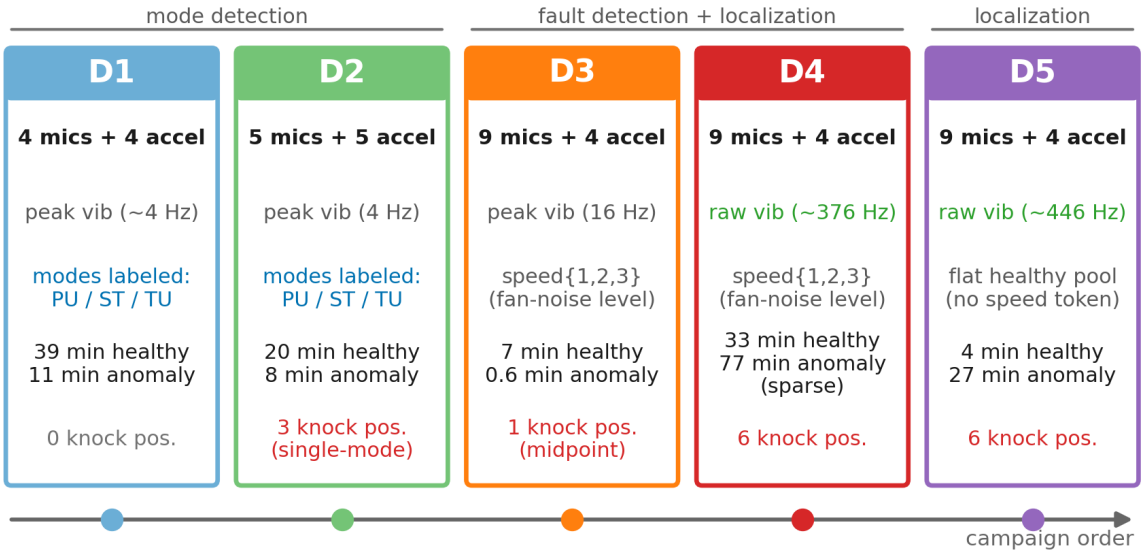


Figure 3.3: The five campaigns at a glance (Table 3.1): the array grows from a 4+4 to a 9+4 rig, and the vibration stream jumps two orders of magnitude in effective rate (peak-amplitude $\sim$4 Hz $\rightarrow$ raw-waveform $\sim$446 Hz) between D3 and D4. Operating mode is labeled only on D1/D2; D3/D4 carry the `speed{N}` fan-noise token instead (§3.3.1).

The campaigns fall into three groups: a mode-detection pair (D1, D2), a fault-detection-and-localization pair (D3, D4), and a focused localization campaign (D5). D1 and D2 sampled the three operating regimes and labeled every recording at the folder level. D3 and D4 traded mode labels for denser anomaly coverage and the first raw-waveform

vibration. D5 was recorded for localization alone, with a flat healthy pool and six knocks at known positions due to its late reception.

3.3.1 Handling of Background Noise Conditions

One naming detail is easy to misread. The `speed1`, `speed2`, and `speed3` tokens in the D3 and D4 folder names are three levels of an added background-noise fan, used at collection time to vary the signal-to-noise ratio. They are not three settings of the runner. The runner's actual mode during D3 and D4 was held fixed but never written into the folder names, and the D5 healthy recordings carry no `speed` token at all. All three campaigns are therefore treated as unlabeled for operating mode, and their healthy recordings are pooled together. Reading `speed{N}` as a mode would split the healthy data along an axis that has nothing to do with the machine's regime. This pooling governs both representation learning and the fitting of the healthy density model, and the speed levels are not otherwise used to partition the data. The anomaly-detection transfer test of §5.2.2 fits a threshold on one set of healthy recordings and evaluates it on a disjoint set drawn from the pooled corpus, so the fan level is only one of several conditions that may differ between the two sides, alongside operating mode and recording session, rather than the axis along which the split is made. Its effect on robustness is reported separately as a noise-level ablation.

3.3.2 Labeling Schemes Across Campaigns

The campaigns differ along two axes that matter for the rest of the pipeline: whether a recording carries an operating-mode label, and how densely anomalies fill an anomaly recording.

Anomaly mechanism. Across every campaign the anomalies are of one physical kind: deliberately-induced impulsive impacts, produced by striking the casing with an impactor (a "knock"), rather than naturally occurring faults. The campaigns differ only in whether a knock carries a spatial label, not in the underlying mechanism.

D1 and D2. Each recording is labeled PUMP, STANDSTILL, or TURBINE at the folder level. D1's anomaly is a single bulk `RandomFault` recording of these knocks with no per-event spatial annotation, so it contributes no knock position (the zero in Table 3.2) and is used only for global, non-localized anomaly detection. D2 also splits its anomaly recordings by knock position and operating context. Three of D2's five anomaly folders are single-mode (TURBINE); the other two are dual-mode, with pump and turbine engaged

at once, an artifact of how the test stand was wired at the time. The dual-mode recordings mix two regimes under one label, so they are kept out of the labeled spatial set. The single-mode anomaly recordings are anomalous throughout, so their knock position applies to the whole recording.

D3 and D4. Each recording carries a `speed{N}` token but no mode. D3 has a single instrumented-hit recording, whose position is the midpoint of the two named accelerometers, at about $(0.085, -0.025, 0.080)$ m on the prototype scale. D4 has seven knock folders, each named after the casing position where an impactor was applied. Unlike the earlier campaigns, anomalies inside a D3 or D4 recording are sparse rather than continuous: each D4 recording is about 11 minutes of mostly healthy background with a few impulsive knocks inserted, so the folder's knock position does not describe most of its windows. One of D4's seven folders carries a fifth accelerometer that the strict ingestion step rejects (§3.4.1), leaving six usable D4 positions.

D5. D5 reuses the circular rig's nine-microphone, four-accelerometer layout but is organized around localization: a flat healthy pool with no speed split, plus six knock folders at known positions, all recorded as raw waveform.

The durations behind these campaigns are summarized in Table 3.2, read directly from the recording headers. D4 and D5 contribute most of the labeled knock positions; the earlier campaigns add the D3 hit and the three D2 single-mode recordings, for sixteen labeled positions in all.

Table 3.2: Durations and the number of labeled knock positions per campaign. The "Anomaly" column is the total length of every anomaly recording in the campaign.

Campaign	Healthy	Anomaly	Knock positions
D1	$\sim$ 39 min	$\sim$ 11 min	0
D2	$\sim$ 20 min	$\sim$ 8 min	3 (single-mode)
D3	$\sim$ 7 min	$\sim$ 0.6 min	1 (midpoint)
D4	$\sim$ 33 min	$\sim$ 77 min (sparse)	6
D5	$\sim$ 4 min	$\sim$ 27 min	6
Total			**16**

3.4 Preprocessing Pipeline

This section describes the pipeline that turns a recording directory into the per-window tensors. It has four stages: ingestion and normalization (§3.4.1), acoustic feature ex-

traction (§3.4.3), vibration feature extraction (§3.4.4), and windowing (§3.4.5), with a synchronization check inside the first stage (§3.4.2). Each stage is configured per campaign, so the differences between campaigns (vibration rate, channel count, label scheme) all end up in one uniform output format.

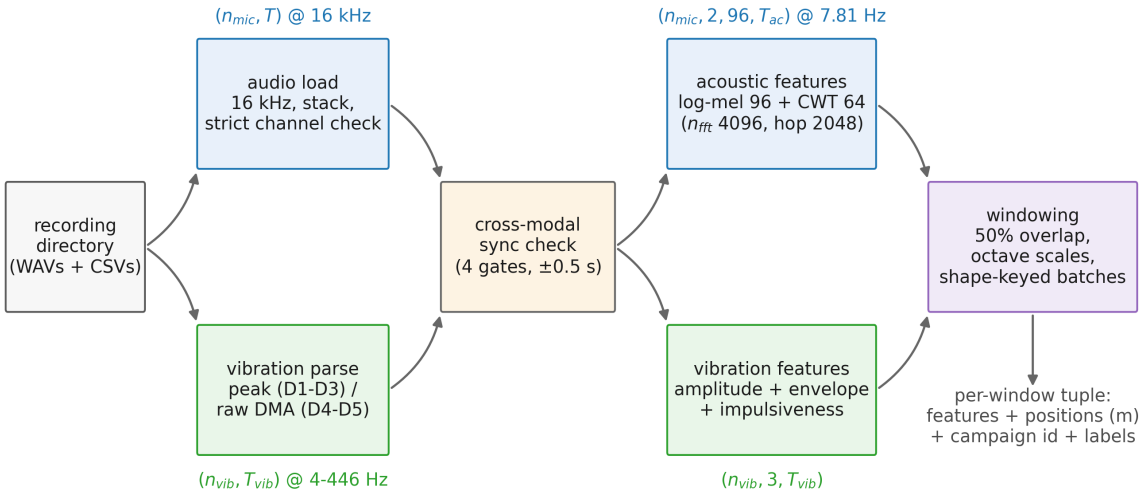


Figure 3.4: The four-stage preprocessing pipeline. Per-campaign differences (vibration format, channel count, label scheme) are absorbed at ingestion; the output is one uniform per-window tuple pairing an $(n_{\text{mic}}, 2, 96, T_{\text{ac}})$ acoustic tensor at $\sim$7.81 Hz with an $(n_{\text{vib}}, 3, T_{\text{vib}})$ vibration tensor at the campaign rate, aligned in time but never mixed in value.

3.4.1 Ingestion and Normalization

A single ingestion step reads each recording directory into a paired data segment and hides the per-campaign details from the rest of the pipeline. It does three things in order: loads the audio, parses the vibration, and attaches the campaign metadata.

Audio load. Each microphone WAV is read into a floating-point array and stacked into an $(n_{\text{mics}}, n_{\text{samples}})$ tensor at the native 16 kHz rate. Integer PCM is scaled to $[-1, 1]$, and a float WAV whose peak sits far outside that range is rescaled with a warning. If the microphone count is not one the campaign expects, the step raises an error instead of silently truncating, so a miscount surfaces immediately rather than as an unexplained problem later. This is also what rejects the five-accelerometer D4 folder mentioned in §3.3.2.

Vibration parse. Depending on the acquisition mode, the pipeline routes the data through one of two parsers. The peak-amplitude parser (D1, D2, D3) standardizes the various on-disk column names into a single internal schema, using a sample-and-hold step to resample the irregular stream of peaks to the campaign's target rate (4 Hz for D1 and D2, 16 Hz for D3). For the raw waveforms (D4, D5), the parser reconstructs the DMA batches by using the median real-sample count to define the batch size, sequentially concatenating the samples from each row, and deriving the sample interval from the median timestamp spacing between those rows. Since each accelerometer operates on its own independent clock, every channel is mapped to a common target grid based on its individually measured rate to maintain precise time alignment. To ensure data integrity, the parser will halt and throw an error if these per-channel rates diverge by more than 10%. Regardless of the parser used, the final output is always a uniform $(n_{\text{vib}}, T_{\text{vib}})$ array at a known, consistent rate.

Positions and units. The last step attaches the sensor positions, the parsed mode label, the operating-condition token, and the knock position where one exists. The position files store centimeters, so every coordinate is divided by 100 and reported in meters. This conversion matters: without it the circular-rig microphones would land at 0–11 meters instead of 0–0.11 meters, which silently inflates every time-difference. A regression test guards the published D3 position against this kind of unit drift.

3.4.2 Cross-Modal Time Synchronization

The ingestion step lines up the audio and vibration by sample index, taking sample 0 of every WAV to match sample 0 of every vibration channel and trimming both to the shorter length. Because the firmware starts the two streams together this is usually right, but the bench rig cannot prove it, so the alignment is checked on every recording. The two streams share information only at the slow envelope scale, where a common physical event appears in both, so the check cross-correlates the mean microphone envelope, decimated to the vibration rate, against the mean vibration envelope over a ± 0.5 s window:

$$r(k) = \frac{\sum_t \tilde{m}(t)\,\tilde{v}(t-k)}{\sqrt{\sum_t \tilde{m}(t)^2}\,\sqrt{\sum_t \tilde{v}(t)^2}}, \qquad k \in [-K, +K], \quad K = \lfloor 0.5\, f_v \rfloor, \tag{3.1}$$

where $\tilde{m}$ and $\tilde{v}$ are the zero-mean audio and vibration envelopes and f_v is the vibration rate. A confidence score, the ratio of the dominant peak to the next-largest local maximum, separates a real shift from noise.

A correction is applied only when the recording clears four gates: a peaked enough audio envelope (excess kurtosis at least 1.0), a confidence of at least 1.5, an offset stable across five sub-segments (spread and drift both within 10 ms), and an offset of at least 1 ms. When all four hold the offset is removed by dropping whole samples from the lagging stream and then applying a sub-sample frequency-domain shift refined with a three-point parabola [54], which avoids inventing samples and leaves a residual below $1/\max(f_{\text{mic}}, f_{\text{accel}}) \approx 62.5\,\mu\text{s}$. On this corpus the gates almost never fire: the healthy recordings are too steady to give the correlation a peak and the knock peaks do not line up across the two streams, so most recordings confirm the firmware-level alignment rather than override it. The step is therefore off by default and is retained for a future deployment whose streams share no common trigger.

3.4.3 Acoustic Feature Extraction

Each microphone channel becomes a two-channel time-frequency image: channel 0 is a log-mel spectrogram over the full audio band, and channel 1 is a continuous-wavelet scalogram over the narrow band that holds the machine tones. The analysis grid is set by the characteristic frequencies in Table 3.3.

Table 3.3: Characteristic mechanical and electrical frequencies of the ROW II pump-turbine at 375 rpm. These are the reference tones of the full-scale machine; the prototype is not claimed to reproduce them, and the rationale for sizing the analysis grid to resolve them is given in §3.4.3. The shaft fundamental and the guide-vane-passing tone are given at their loading-adjusted values (5.87 Hz and 117.3 Hz) rather than the no-load nominals (6.25 Hz and 125 Hz); see footnote[1].

Quantity	Frequency (Hz)
Shaft fundamental (nominal, 375 rpm)	6.25
Shaft fundamental (loading-adjusted)	5.87
Runner-blade-passing (7×)	43.75
Electrical-network	50.00
Rotor-pole-passing (16×)	100.00
Guide-vane-passing (loading-adjusted, 20×)	117.30

[1]The nominal shaft fundamental of 6.25 Hz corresponds to a synchronous 375 rpm runner under no load. Under load, synchronous slip and guide-vane back-pressure pull the effective fundamental down to about 5.87 Hz, and that shift carries through to every harmonic, moving the guide-vane-passing tone from 125 Hz to 117.3 Hz. The loading-adjusted values are used so the analysis grid matches the content actually present in the recordings.

Analysis grid. The grid was selected empirically, with a sweep over n_{fft}, hop length, and the number of mel bands across the five campaigns, settling on $n_{\text{fft}} = 4096$, hop $= 2048$, and $n_{\text{mels}} = 96$. At 16 kHz this is a 256 ms window, a 128 ms hop, and a frame rate of about 7.81 Hz. This choice is also forward-compatible with the field machine: the tightest frequency-resolution demand a ROW II deployment would place on the grid is the 17 Hz gap between the rotor-pole-passing tone at 100 Hz and the guide-vane-passing tone at 117.3 Hz, and $n_{\text{fft}} = 4096$ resolves it with room to spare (3.9 Hz bins), so the same preprocessing would transfer to the field without re-tuning. On the prototype data the fine resolution is not harmful: it is the empirical optimum of the sweep, the short impulsive events are captured by the multi-scale windowing and the wavelet channel rather than by frequency resolution, and the hop length barely affected the result once window and band count were fixed. The same grid is therefore used for all five campaigns.

Channel 0: log-mel spectrogram. The log-mel channel is computed with librosa [81] over 20 Hz to the 8 kHz Nyquist limit. The mel scale puts most of its resolution at low frequencies, where the machine tones live, which is the usual reason to prefer it over a linear scale in this setting [15]. An absolute decibel reference is kept rather than normalizing each recording to its own peak, so that real loudness differences between regimes (PUMP is louder at the casing wall than STANDSTILL) survive into the features, and the noise floor is clipped 80 dB below the peak so silent bins do not go to $-\infty$.

Channel 1: CWT scalogram. The second channel is a continuous-wavelet scalogram with a complex Morlet wavelet on 32 log-spaced scales from 20 Hz to 250 Hz. That spacing puts about 8 Hz between bins near 100 Hz, so the 100 Hz and 117.3 Hz tones, separated by 17.3 Hz, still land about two bins apart, in separate bins where the mel filterbank cannot tell them apart. The 250 Hz cap covers every machine tone with an octave to spare. To keep the transform affordable, the signal is anti-aliased and decimated to 1 kHz first; the resulting 500 Hz Nyquist still covers the whole band. The log-mel and CWT outputs are stacked into one $(n_{\text{mels}}, T_{\text{frames}})$ image, so the same row index means a different physical frequency in each channel: the log-mel channel covers the full audio band, and the CWT channel zooms into the machine-tone band. The two are brought to a common shape by max-pooling the CWT along time (which keeps the peak of a short knock rather than averaging it away) and linear interpolation along frequency. The acoustic stage outputs an $(n_{\text{mics}}, 2, 96, T_{\text{frames}})$ tensor at about 7.81 Hz.

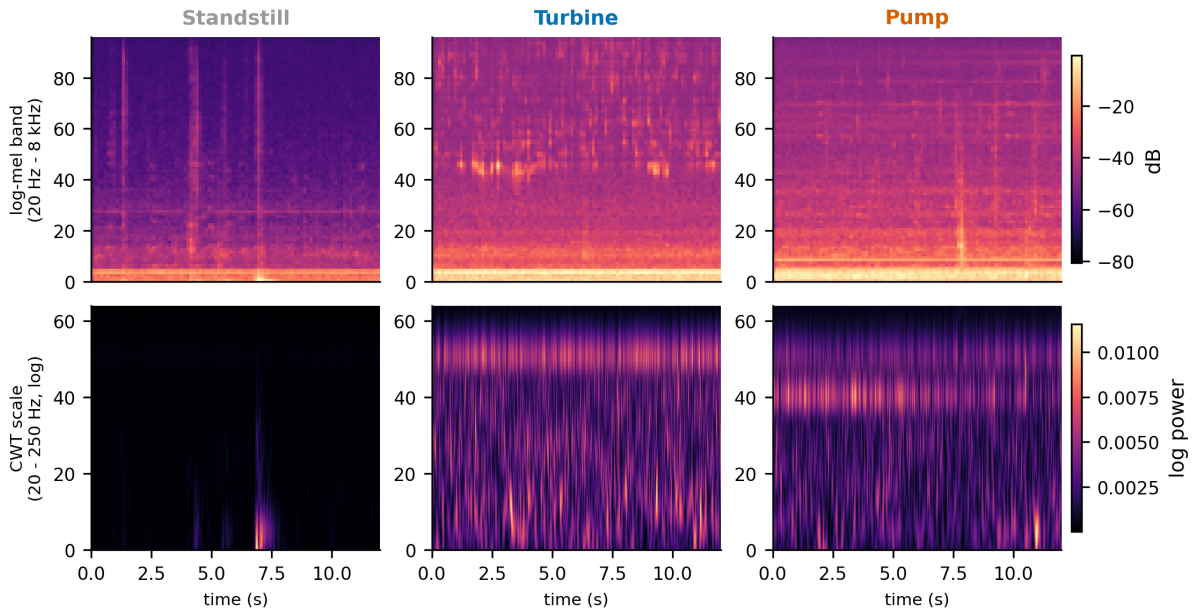


Figure 3.5: The premise of RQ1, visible by eye: the two acoustic input channels (top: log-mel spectrogram; bottom: CWT scalogram of the machine-tone band) differ systematically across STANDSTILL, TURBINE and PUMP recordings of the same rig. Absolute dB references are retained, so the loudness difference between regimes survives into the features.

3.4.4 Vibration Feature Extraction

Each accelerometer channel becomes a three-channel time series: amplitude, envelope, and impulsiveness.

Channel 0: amplitude. The raw amplitude is zero-meaned and z-score normalized. This matters because the on-disk scales differ by two orders of magnitude across campaigns, with the peak-amplitude streams near 0–2000 ADU and the D4 raw values around 18 000. The mean and standard deviation are computed once over each dataset's healthy pool and reused for every window of that dataset, which keeps real loudness differences inside a recording while still bringing the campaigns onto a common scale.

Channel 1: Hilbert envelope. The second channel is the magnitude of the analytic signal, z-score normalized. The envelope is the standard tool for picking out bearing-type defects in vibration data [5, 11]: on the raw waveforms it shows amplitude modulation at the blade-passing frequency, and on the peak streams it keeps the slow modulation that separates a developing condition from a steady one.

Channel 2: rolling impulsiveness. The third channel flags departures from a steady background, the signature of an impulsive event. Which statistic to use depends on the vibration rate, because no single sample-count window works across all five campaigns. The sampling error of the kurtosis estimator on Gaussian data is about $\sqrt{24/N}$, so at $N = 5$ samples it is roughly ± 2.2, too noisy to trust, while at 4 Hz a window long enough to reach $N = 31$ would span more than 7 s and smear a single knock into its surroundings. The window is fixed in time instead, at 100 ms (the rough ring-down of a knock on the plastic casing), and the statistic is chosen from the resulting sample count. When the window holds at least 31 samples, the bias-corrected excess kurtosis of Joanes and Gill is used; otherwise a rolling crest factor $\max|x|/\mathrm{RMS}(x)$ over a 1 s window is used, which stays well-defined down to a few samples [5, 24]. Both say the same thing to the model, that a high value means an impulsive event. The per-dataset choice is shown in Table 3.4. The vibration stage outputs an $(n_{\mathrm{vib}}, 3, T_{\mathrm{vib}})$ tensor at the campaign's target rate.

Table 3.4: Per-dataset choice of the channel-2 impulsiveness statistic. The window is fixed in time (100 ms for kurtosis, 1 s for crest factor) and rounded to the nearest odd sample count; kurtosis is used only when that count reaches 31. Only the two raw-waveform campaigns have a high enough rate to support it.

Dataset	Accel. SR	Kurtosis-window samples	Channel-2 statistic
D1 (peak)	4 Hz	1 (< 31)	Crest factor ($N = 5$ over 1 s)
D2 (peak)	4 Hz	1 (< 31)	Crest factor ($N = 5$ over 1 s)
D3 (peak)	16 Hz	3 (< 31)	Crest factor ($N = 17$ over 1 s)
D4 (raw)	≈ 376 Hz	39 (≥ 31)	Joanes–Gill excess kurtosis
D5 (raw)	≈ 446 Hz	45 (≥ 31)	Joanes–Gill excess kurtosis

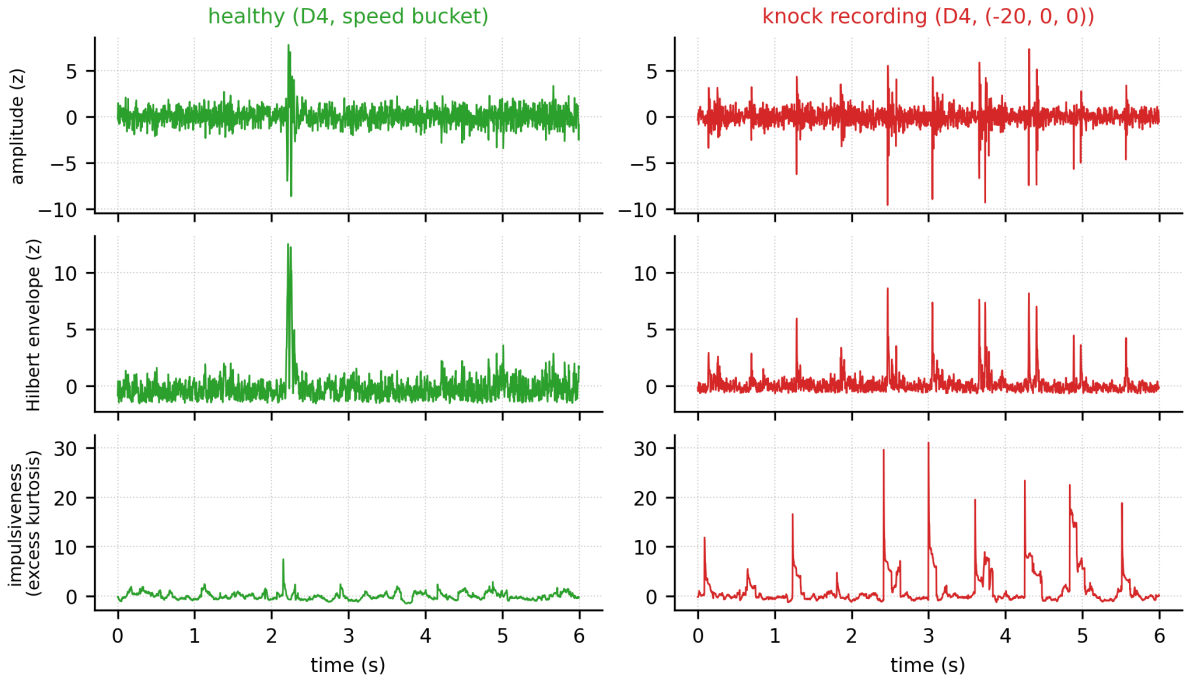


Figure 3.6: The three vibration input channels on raw-waveform D4 data, healthy (left) versus a knock recording (right) on a shared y-scale. Amplitude and Hilbert envelope carry the event; the rolling Joanes–Gill excess kurtosis (channel 2) isolates its impulsiveness against the steady background, the signature the AND rule's vibration branch fires on.

3.4.5 Windowing

The two modalities now sit at different rates: about 7.81 Hz on the audio side and 4 Hz to 446 Hz on the vibration side. The last stage cuts them into the paired windows the model reads, and it is the only place the two modalities are brought together at the data level, and only in time, with no values mixed. A window is set by a duration W in seconds, and the number of frames on each side is computed from that side's own rate:

$$\text{frames}_{ac} = \max\big(2, \text{round}(W\, f_{ac}^{seg})\big), \tag{3.2}$$

$$\text{frames}_{vib} = \max\big(2, \text{round}(W\, f_{vib}^{seg})\big). \tag{3.3}$$

Using each side's own rate keeps a high-rate D4 window from being cut down to the slowest stream in the corpus. The vibration crop's start is taken from the audio window's start time, $\text{start}_{\text{vib}} = \text{round}(\text{start}_{\text{ac}} / f_{\text{ac}}^{\text{seg}} \cdot f_{\text{vib}}^{\text{seg}})$, so the two crops begin at the same instant. Windows are grouped into batches by their shape, keyed by $(n_{\text{mics}}, n_{\text{vib}}, \text{frames}_{\text{ac}}, \text{frames}_{\text{vib}})$, so windows from arrays with different channel counts (the 4+4, 5+5, and 9+4 rigs) are never stacked together and no padding is needed.

A single window length does not suit every task: longer windows give steadier estimates of the operating mode, while shorter ones capture a short knock better. The shortest usable window is also set by the vibration rate. Each dataset is therefore given a small tuple of window lengths, spaced by octaves, with windows drawn at 50% overlap (Table 3.5).

Table 3.5: Per-dataset window lengths. Lengths are octave-spaced; the peak-amplitude campaigns cannot resolve windows shorter than $5/f_{\text{vib}}$ (one length-five convolution kernel on the vibration side), while the raw-waveform campaigns are effectively unconstrained.

Dataset	Accel. SR	Min window	Window lengths (s)
D1	4 Hz	1.25 s	1.5, 3.0, 6.0
D2	4 Hz	1.25 s	1.5, 3.0, 6.0
D3	16 Hz	0.31 s	0.5, 1.0, 2.0, 4.0
D4	$\approx$ 376 Hz	$\sim$ 0.013 s	0.5, 1.0, 2.0, 4.0
D5	$\approx$ 446 Hz	$\sim$ 0.011 s	0.5, 1.0, 2.0, 4.0

The per-window output. The pipeline produces one uniform tuple per window: an $(n_{\text{mics}}, 2, 96, T_{\text{ac}})$ acoustic tensor, an $(n_{\text{vib}}, 3, T_{\text{vib}})$ vibration tensor, the per-channel sensor positions in meters, the campaign identifier, the operating-mode label where it exists, and the knock position where it exists. The two tensors are aligned in time but kept separate, of separate shapes, so the varying per-campaign array geometry is preserved all the way through. These per-window tuples are the output of the preprocessing pipeline and the input to the model described in the following chapters.

Chapter 4

Architecture

This chapter describes the model that consumes the per-window tensors produced by the preprocessing pipeline of Chapter 3. The design follows directly from the research questions of Chapter 1: it must fuse acoustic and vibration streams, recover the operating context without labels, score windows for anomaly in a way that adapts to that context, and localize a detected anomaly inside the casing. The chapter presents the architecture and its training objectives only; the evaluation protocol, baselines, and metrics are set out in Chapter 5, and the results follow in Chapter 6.

4.1 Architecture Overview

The complete system, from the sensor streams through the data pipeline to the operator-facing outputs, was set out at a high level in §1 (Figure 1.2); this section expands its learned core into the four stages that the rest of the chapter details. The architecture is a single chained system, built and trained in four stages that are referred to throughout as V1 to V4. Each stage adds one capability and reuses the frozen output of the stage before it.

1. **Per-modality encoders (V1).** A two-dimensional CNN over the acoustic representation and a one-dimensional CNN over the vibration representation turn each sensor channel into a feature vector, which is then enriched with position and dataset information and pooled into a per-channel token sequence (§4.2).

2. **Cross-modal fusion and context (V2).** A bidirectional cross-attention block lets the two token sequences attend to each other, and a pooling-by-attention module compresses the fused tokens into a single *context vector* $\boldsymbol{c}_t$ that represents the machine's current operating state (§4.3). Both stages are trained without labels by self-supervision (§4.4).

3. **Conditional anomaly detection (V3).** A conditional normalizing flow scores each window by its exact negative log-likelihood under a density that is modulated by $\boldsymbol{c}_t$, and a per-cluster percentile threshold turns that score into an alert (§4.5).

4. **Source localization (V4).** For windows that V3 flags as anomalous, a conditional localization head combines a classical acoustic power map with structure-borne time-difference features to estimate a three-dimensional source position (§4.6).

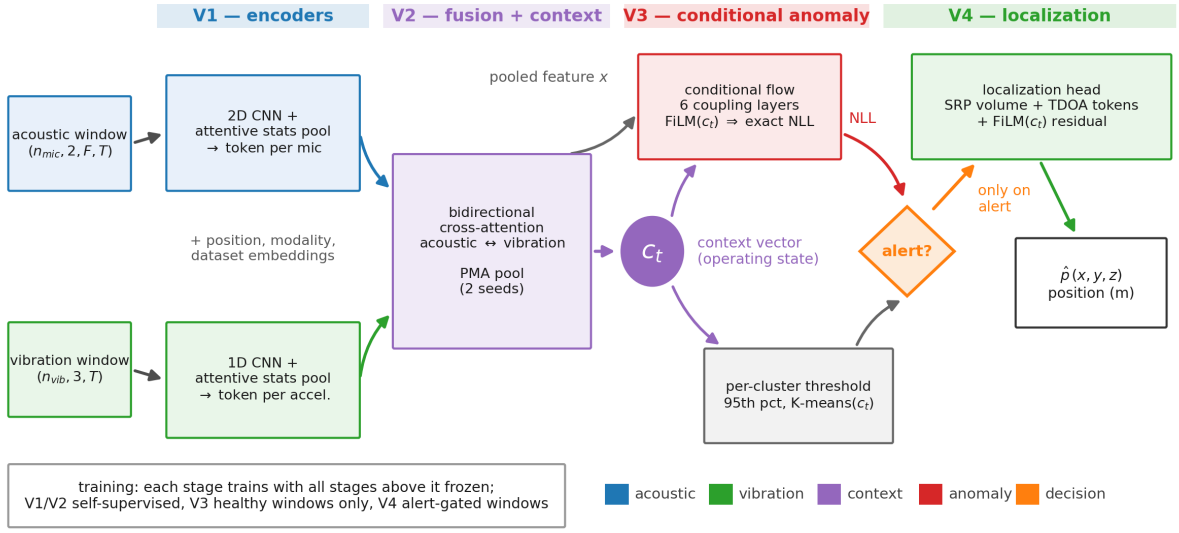


Figure 4.1: The chained V1→V4 system. Per-modality encoders (V1) turn each sensor channel into a token; bidirectional cross-attention and a PMA pool (V2) compress the token sequences into the context vector $\boldsymbol{c}_t$; a FiLM-conditioned normalizing flow (V3) scores each window by exact negative log-likelihood against a per-cluster threshold; and only alert windows reach the localization head (V4). Each stage trains with all stages above it frozen.

Three design commitments run through every stage. First, the model is *channel-agnostic*: it is built on permutation-invariant set operations so that the same trained weights process the 4+4, 5+5, and 9+4 arrays of the five campaigns without retraining [70, 133]. Second, the operating context is *label-free*: the context vector is learned by self-supervision and the anomaly threshold is fitted on healthy data alone, so the operating-mode labels that exist for only two of the five campaigns (Chapter 3) are never used at training time. Third, scoring is *exact*: the anomaly head reports a true log-likelihood rather than a reconstruction error, which gives a directly interpretable and

streamable anomaly score. The way the four stages connect at inference time, and how the system runs as a stream, is set out in §4.8.

4.2 Per-Modality Encoders

Each modality is first encoded independently. The two backbones are deliberately plain convolutional networks; the modeling effort is concentrated in the fusion and conditioning stages rather than in ever-deeper feature extractors. Both backbones share a feature width of 64 and use GELU activations [44] and batch normalization [53]. Batch normalization is retained deliberately: its cross-sample coupling supplies the implicit anti-collapse pressure that the contrastive objective of §4.4 depends on.

4.2.1 Acoustic Encoder

The acoustic encoder is a three-block two-dimensional CNN applied to each microphone independently. Its input is the $(2, F, T)$ per-microphone image of Chapter 3, with the log-mel spectrogram on one channel and the wavelet scalogram on the other. The blocks widen the channel count $2 \to 32 \to 64 \to 128$ with 3×3 convolutions and 2×2 max-pooling, after which the 128-channel feature map is collapsed to a single vector per microphone by the attentive statistics pool and projected to the 64-dimensional feature width.

The pooling step is an Attentive Statistics Pool rather than a plain global average. Attentive statistics pooling, a technique originally developed for text-independent speaker verification [89], replaces the first-moment summary with an attention-weighted concatenation of the mean and the standard deviation, $[\boldsymbol{\mu} \,\|\, \boldsymbol{\sigma}]$, computed over the time-frequency positions:

$$\alpha = \mathrm{softmax}\big(\phi(\boldsymbol{h})\big), \qquad \boldsymbol{\mu} = \sum_m \alpha_m \boldsymbol{h}_m, \qquad \sigma_c = \sqrt{\textstyle\sum_m \alpha_m h_{m,c}^2 - \mu_c^2}, \tag{4.1}$$

where $\boldsymbol{h}_m$ are the per-position feature vectors and ϕ is a small attention network with a bottleneck reduction of 8. Retaining the standard deviation is the point: a short knock occupies only a small fraction of the frames in a multi-second window, so a mean-only summary dilutes it, whereas the second-moment term preserves the spread that a transient produces regardless of how much steady background surrounds it.

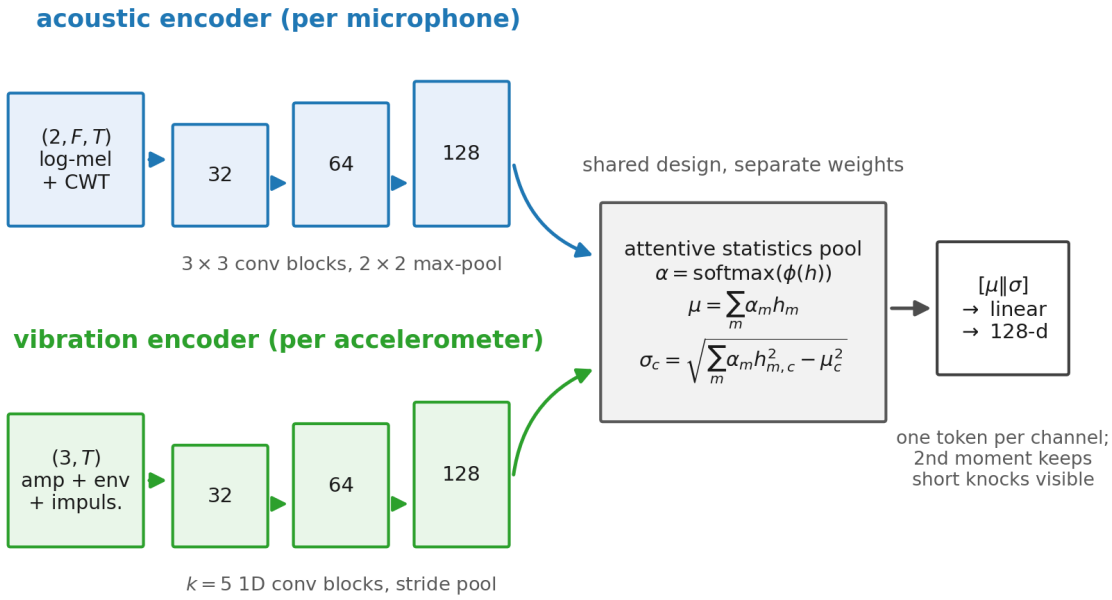


Figure 4.2: The two deliberately plain encoder backbones. Both widen to a 128-channel convolutional map, pool it to a shared 64-dimensional feature width, and meet in the same attentive statistics pool (Eq. 4.1); retaining the attention-weighted standard deviation $[\boldsymbol{\mu}\|\boldsymbol{\sigma}]$ keeps a short knock visible inside a multi-second window.

4.2.2 Vibration Encoder

The vibration encoder is the one-dimensional analogue, applied to each accelerometer channel independently. Its input is the $(3, T)$ per-channel series of Chapter 3 (amplitude, Hilbert envelope, and rolling impulsiveness). Three one-dimensional convolution blocks with kernel size 5 widen the channels $3 \to 32 \to 64 \to 128$, and the same attentive statistics pool (now over the time axis) and linear projection produce one 64-dimensional feature vector per accelerometer. Keeping the two backbones structurally parallel is what allows the per-modality weights learned in V1 to be transferred directly into the fusion encoder of V2.

4.2.3 Channel Tokens and the Set-Transformer Pool

The per-channel feature vectors are turned into a token sequence that the attention stages can operate on. Each token concatenates its feature vector with the sensor's three-dimensional position in meters, a learned modality embedding, and a learned dataset embedding, and projects the result to a shared 64-dimensional embedding space. Carrying

the position and the campaign identity in the token is what lets a single model reason about different array geometries.

The token sequence is then passed through one Multi-head Attention Block (MAB) for intra-modality self-attention, followed by a Pooling-by-Multi-head-Attention (PMA) module with a single learned seed query that compresses the sequence into one summary vector. Both blocks are the permutation-invariant building blocks of the Set Transformer [70]: the MAB applies the standard scaled dot-product attention of the Transformer [119],

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{4.2}$$

inside a pre-norm transformer block, and the PMA uses learned seed queries in place of a fixed average so the network can decide which channels carry the most information [133]. Because both operations are invariant to the ordering and the number of input tokens, the encoder accepts any channel count without change. The per-modality encoder returns both the token sequence, which the fusion stage consumes, and the single summary vector, which the V1 contrastive objective and the operating-mode sanity check use.

4.3 Cross-Modal Fusion and the Context Vector

Fusion is a single bidirectional cross-attention block. It runs two MAB passes that share no weights: the acoustic tokens form the queries and attend over the vibration tokens, and the vibration tokens symmetrically attend over the acoustic tokens, by (4.2). Each modality therefore emerges having read the other, while the two streams stay distinct and keep their own channel counts. A single bidirectional block is used instead of deeper stacks. One symmetric hop suffices to expose every acoustic token to every vibration token while preserving channel-count invariance.

The fused tokens are pooled into the context vector $\boldsymbol{c}_t \in \mathbb{R}^{64}$ by a PMA module with two learned seed queries whose outputs are averaged. The two seeds give the pool enough capacity to summarize the joint sequence from more than one aspect while keeping the context dimension fixed. This vector is the operating-state representation that the rest of the system conditions on: it is the quantity clustered to recover operating mode (§4.4) and the conditioner fed to the anomaly and localization heads (§4.5, §4.6).

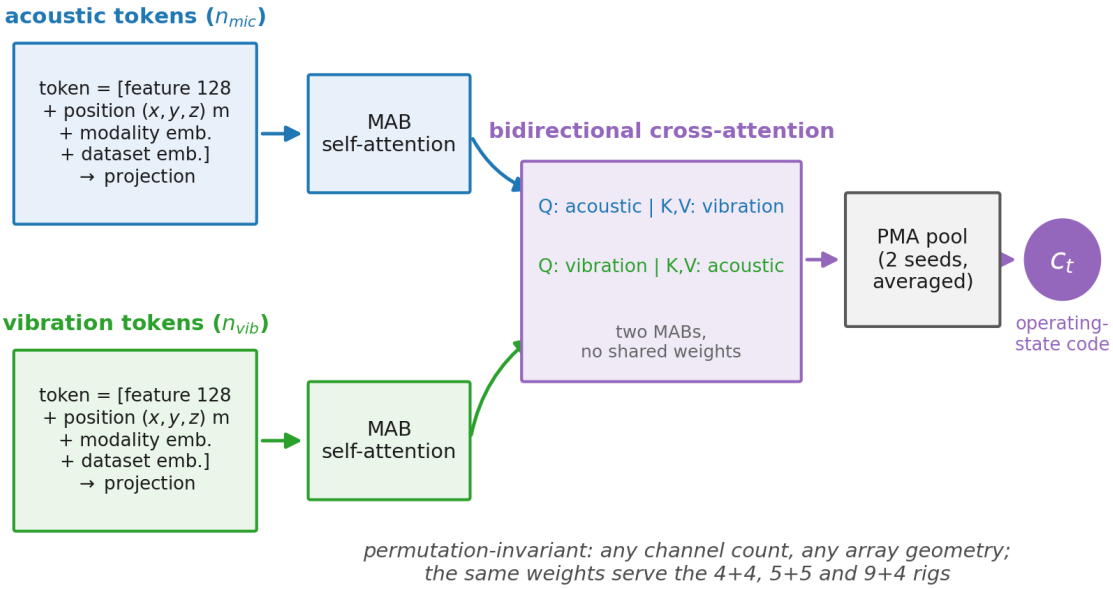


Figure 4.3: Token construction and fusion. Each channel token carries its feature vector, metric sensor position, and learned modality and dataset embeddings; one bidirectional cross-attention hop couples every acoustic and vibration token, and a two-seed PMA pool produces the context vector $\boldsymbol{c}_t$. Every block is permutation-invariant, so the same weights serve the 4+4, 5+5 and 9+4 arrays without retraining.

4.4 Self-Supervised Context Learning

The encoders and the fusion block are trained without any operating-mode labels, in two self-supervised stages. The premise is that the operating mode is present in the signal but unannotated on most campaigns (Chapter 3), so the representation must be learned from structure in the data rather than from supervision.

V1: per-modality contrastive pretraining. Each backbone is pretrained on its own modality with a contrastive objective. Two augmented views of the same window are pulled together in embedding space while views from different windows are pushed apart, using the normalized temperature-scaled cross-entropy loss standard in contrastive representation learning [13, 90]. The augmentations are mild and physically motivated: spectrogram masking [93], channel dropout, and gain jitter. The effect is to make the encoder invariant to nuisance variation while remaining sensitive to the operating regime.

V2: joint fusion and context. The fusion block and the context pool are then trained on paired windows, with the per-modality encoder weights inherited from V1. Three

objectives are combined. A contrastive loss on the context vector $\boldsymbol{c}_t$ pulls together two augmented views of the same window. A latent masked-modeling loss [7, 32] replaces a random fraction of per-modality tokens with a learned mask token before fusion and regresses the fused output at the masked positions back to the pre-mask embeddings under a cosine objective; this operates in the latent space rather than on raw pixels or samples, so the context vector is not forced to reconstruct the irrelevant high-frequency detail of either modality. A cross-modal alignment term encourages the two per-modality summaries of the same window to agree. Together these objectives shape $\boldsymbol{c}_t$ into a compact code of the operating state.

To answer the first research question, the learned context vectors of the labeled campaigns are clustered with K-means [75] and compared against the recorded operating modes. Crucially, only the label-free cluster identities are used by anything downstream; the operating-mode labels enter solely in the evaluation of Chapter 6, never at fit time.

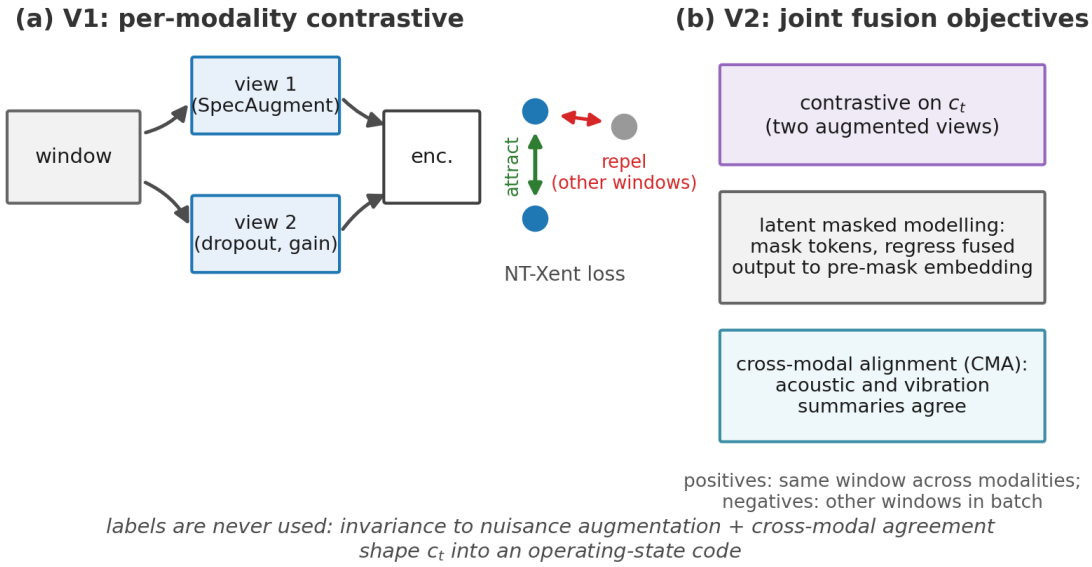


Figure 4.4: The label-free training objectives. (a) V1 pulls two augmented views of the same window together and pushes other windows apart (NT-Xent). (b) V2 combines a contrastive loss on $\boldsymbol{c}_t$, a latent masked-modeling loss, and a cross-modal alignment (CMA) term. On this corpus CMA yields no measured mode-discovery gain over an un-aligned variant (Table 6.4, §7.2).

4.5 Conditional Anomaly Detection Head

The anomaly head scores a window by how probable it is under a learned density of healthy operation, conditioned on the operating context. Following the normalizing-flow approach to density estimation [106, 116], it is a conditional normalizing flow built from affine coupling layers [23], conditioned on $\boldsymbol{c}_t$ by Feature-wise Linear Modulation (FiLM) [95]. Working with an exact likelihood, rather than a reconstruction error, is what makes the score directly interpretable as statistical rarity and lets the same model adapt its notion of normal to the current regime [36, 62].

The head consumes a fixed-dimensional feature $\boldsymbol{x}$, obtained by pooling the fused per-channel tokens of a window, together with the context vector $\boldsymbol{c}_t$. Each coupling layer splits $\boldsymbol{x}$ by a binary mask $\boldsymbol{m}$ that alternates between layers, passes the masked half through unchanged, and applies a conditioned affine transform to the other half:

$$\boldsymbol{z} = \boldsymbol{m} \odot \boldsymbol{x} + (1 - \boldsymbol{m}) \odot \big(\boldsymbol{x} \odot \exp(\boldsymbol{s}(\boldsymbol{x}_a, \boldsymbol{c}_t)) + \boldsymbol{t}(\boldsymbol{x}_a, \boldsymbol{c}_t)\big), \qquad \boldsymbol{x}_a = \boldsymbol{m} \odot \boldsymbol{x}, \tag{4.3}$$

where the scale and translation networks $\boldsymbol{s}$ and $\boldsymbol{t}$ are FiLM-modulated MLPs and the scale is bounded as $\boldsymbol{s} = s_{\max} \tanh(\cdot)$ with $s_{\max} = 2$ for stable, invertible layers. The FiLM modulation normalizes the conditioner first and applies a learned scale and shift between hidden layers,

$$\boldsymbol{h} \leftarrow \big(1 + \gamma(\tilde{\boldsymbol{c}})\big) \odot \boldsymbol{h} + \beta(\tilde{\boldsymbol{c}}), \qquad \tilde{\boldsymbol{c}} = \mathrm{LN}(\boldsymbol{c}_t), \tag{4.4}$$

with γ and β zero-initialized so the flow starts as an unconditional density and only learns to use the context once that demonstrably helps. Stacking six such layers gives an invertible map to a standard-normal base distribution, and the conditional log-likelihood follows in closed form from the change of variables,

$$\log p(\boldsymbol{x} \mid \boldsymbol{c}_t) = \log \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I}) + \sum_{\text{layers } i} \sum_{: m_i = 0} s_i, \tag{4.5}$$

so the anomaly score is simply $-\log p(\boldsymbol{x} \mid \boldsymbol{c}_t)$. Training maximizes this likelihood on healthy windows only. The flow's width and depth, and the training settings, selection grids, and parameter counts of all four stages, are tabulated in Appendix A.

Per-cluster thresholds. A raw likelihood is turned into an alert by a context-dependent threshold, which is what keeps the false-positive rate stable across operating modes. The healthy context vectors are clustered into $K = 3$ groups and the 95th percentile of the

healthy anomaly score is taken within each cluster. At inference a window is assigned to its nearest healthy centroid in $\boldsymbol{c}_t$ space and flagged when its score exceeds that cluster's threshold. Here K is a design parameter: the number of operating regimes the per-cluster calibration is meant to distinguish. It is set to $K = 3$ to match the three steady modes the prototype was operated in (Pump, Standstill, and Turbine), so that each steady regime receives its own healthy threshold and a change of regime reads as a context shift rather than a fault. The choice enters only through the cluster count, never through labels: the thresholds are fitted on healthy data alone, and the cluster identities are arbitrary integers, matched to mode names only for the reporting of Chapter 6. The same scheme extends to a field machine by setting K to its number of operating regimes. It could be extended further to treat the transition phases between regimes as their own clusters, giving K equal to the number of steady modes plus the number of transition types, which might let the per-cluster threshold stay calibrated through a transition as well as within each steady mode. Whether such transition clusters help is an empirical question left to future work, since the prototype corpus contains no transition recordings on which to fit or test them.

Event extraction. For streaming use, the per-window score timeline is reduced to discrete events by a hysteresis (Schmitt-trigger) rule: an event opens when the score crosses the cluster threshold and closes only when it falls back below a lower bound, which prevents a single event from fragmenting as the score fluctuates around one boundary [107]. This event view is the form an operator would act on and the gate that selects which windows reach the localization head.

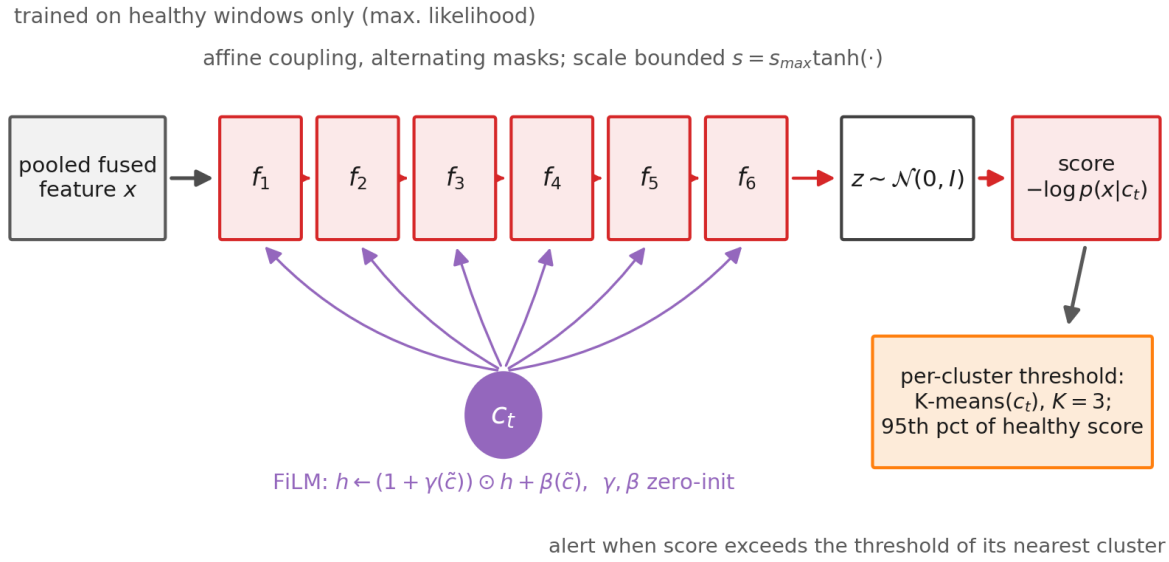


Figure 4.5: The conditional anomaly head. Six affine coupling layers with FiLM-modulated scale/translation networks (Eq. 4.3, 4.4) map the pooled window feature to a standard normal, giving the exact score $-\log p(\boldsymbol{x} \mid \boldsymbol{c}_t)$ (Eq. 4.5); zero-initialized FiLM makes the unconditional ablation exact. Alerts are raised against the 95th percentile of healthy scores within the window's nearest healthy $\boldsymbol{c}_t$-cluster.

4.6 Source Localization Head

For a window the anomaly head has flagged, the localization head estimates the three-dimensional source position. It is conditional and channel-agnostic, and it combines a classical acoustic prior with a learned, structure-borne correction, following the hybrid principle that feeding a classical steered-power map into a neural network outperforms learning from raw audio alone [21].

Acoustic prior. The front-end computes a Steered-Response-Power map with phase transform (SRP-PHAT) over a fixed three-dimensional candidate grid in meters [33]. Because the grid is fixed, the resulting power volume has the same shape regardless of how many microphones contributed, which preserves channel agnosticism. When a window contains a short transient inside a long span, the map is computed on the highest-energy sub-window so the cross-correlations concentrate on samples that actually carry source information.

Heatmap soft-argmax. A small three-dimensional CNN turns the power volume into a per-voxel logit map and a global summary feature. A differentiable soft-argmax over the

logit volume yields a continuous initial estimate in meters,

$$\hat{\boldsymbol{p}}_0 = \sum_v \operatorname{softmax}\big(\boldsymbol{\ell}/\tau\big)_v \boldsymbol{g}_v, \tag{4.6}$$

where $\boldsymbol{\ell}$ are the voxel logits, $\boldsymbol{g}_v$ are the voxel-center coordinates, and τ is a temperature. This integral-regression formulation [115] keeps the spatial structure of the map that a global pool would discard and gives sub-voxel precision on a fixed grid.

Structure-borne correction. Accelerometer pairs supply a complementary timing cue. For each pair the front-end computes a generalized cross-correlation peak delay [67], converts it to a path difference in meters through the assumed structure-borne wave speed, and forms an eight-dimensional token of the path difference, the two sensor positions, and the pair distance. A Set-Transformer pool over the variable number of pairs produces one channel-agnostic vibration feature. A FiLM-conditioned residual MLP then reads the global acoustic feature, the vibration feature, and the initial estimate, and predicts a bounded additive correction,

$$\hat{\boldsymbol{p}} = \hat{\boldsymbol{p}}_0 + r \tanh\big(\mathrm{MLP}_{\mathrm{FiLM}(\boldsymbol{c}_t)}(\cdot)\big), \tag{4.7}$$

with the correction capped at $r = 0.20$ m. The bound is tied to the physical scale of the prototype: a fault originates inside the instrumented rig, so the learned correction is restricted to roughly the size of the machine rather than left free to move the estimate further. A larger cap could only carry the prediction beyond the physical extent of the casing, which is never the correct location for a source inside it. At $r = 0.20$ m the correction stays on the order of the prototype's spatial extent and is also enough to repair the inward bias that soft-argmax exhibits on grid-edge positions. This bound is the architecture default and produces every localization number reported in this thesis; the residual range is swept in §6.7, and the relationship between that sweep and this default is reconciled there. As in the anomaly head, the FiLM projections are zero-initialized, so at the start of training, and under the unconditional ablation (Chapter 5), the head reduces exactly to soft-argmax plus an unconditional residual. The acoustic speed of sound is taken as $343\,\mathrm{m\,s^{-1}}$ and the structure-borne wave speed in the plastic casing as a fixed assumed value of approximately $2000\,\mathrm{m\,s^{-1}}$; calibrating it per casing material is left to future work.

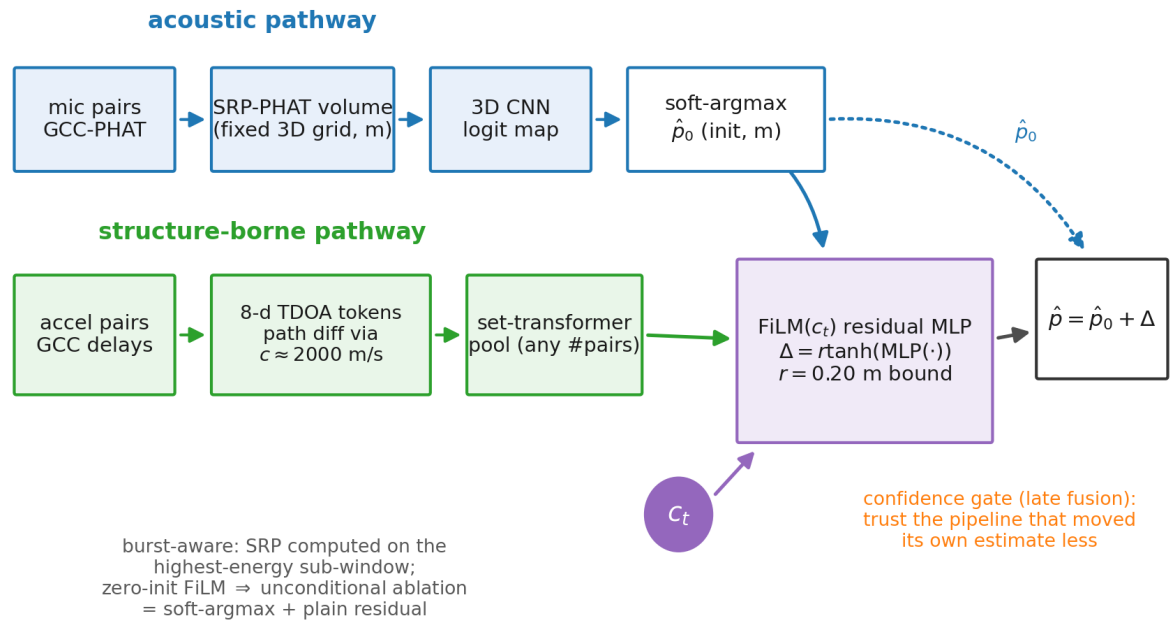


Figure 4.6: The localization head. A classical SRP-PHAT volume on a fixed metric grid seeds a soft-argmax initial estimate (Eq. 4.6); accelerometer-pair GCC delays become channel-agnostic TDOA tokens through the assumed structure-borne wave speed; and a FiLM($\boldsymbol{c}_t$)-conditioned MLP predicts a bounded residual correction (Eq. 4.7, $r = 0.20\,\text{m}$). The confidence-gated late-fusion variant trusts whichever pipeline moved its own initial estimate less.

4.7 Integration of Supervisory Signals

The fourth research question asks whether slow supervisory signals such as temperature and pressure can sharpen detection and localization. The architecture provides the pathway directly: both heads accept an optional supervisory vector $\boldsymbol{s}_t$ that is concatenated with the context vector before the FiLM projection, so a process variable conditions the same modulation mechanism the operating context already drives.

The prototype corpus carries no native temperature, pressure, or SCADA channels (Chapter 1), so this question cannot be answered on the prototype data and is treated as exploratory. Two limited checks are possible: the mechanism is exercised with the recorded operating-condition token as a stand-in supervisory variable, and the real ROW II SCADA archive is mined offline to rank its channels by the mutual information they carry with historical anomaly events, identifying the pressure, thermal, and hydraulic

channels a field deployment would wire into $\boldsymbol{s}_t$. Their setup is given in Chapter 5 and their outcomes in Chapter 6; full validation is left to that future deployment.

4.8 Chained Inference and Streaming

At inference the stages run as a chain on the frozen weights of the stages above them. The encoder and fusion block of V1 and V2 map a window to its context vector $\boldsymbol{c}_t$ and its pooled feature $\boldsymbol{x}$. The anomaly head scores the window by (4.5) and compares it to the per-cluster threshold. Only windows that raise an alert are passed to the localization head, which returns a position. This gating is also how the supervised localization cohort is assembled during training (Chapter 3): on campaigns where anomalies are sparse, a recording-level position label is applied only to the windows the anomaly head surfaces, rather than to every window of the recording.

The same trained system runs as a stream. Following standard practice in acoustic event detection [117], the inference-time window stride is decoupled from the training-time stride, so the score timeline can be produced at a finer cadence than training used, and the hysteresis rule of §4.5 converts that timeline into the discrete events an operator would see. Because every component is permutation-invariant and channel-count agnostic, the same weights serve the four-, five-, and nine-microphone arrays of the corpus, which is the property a single deployable model on a varying sensor array requires.

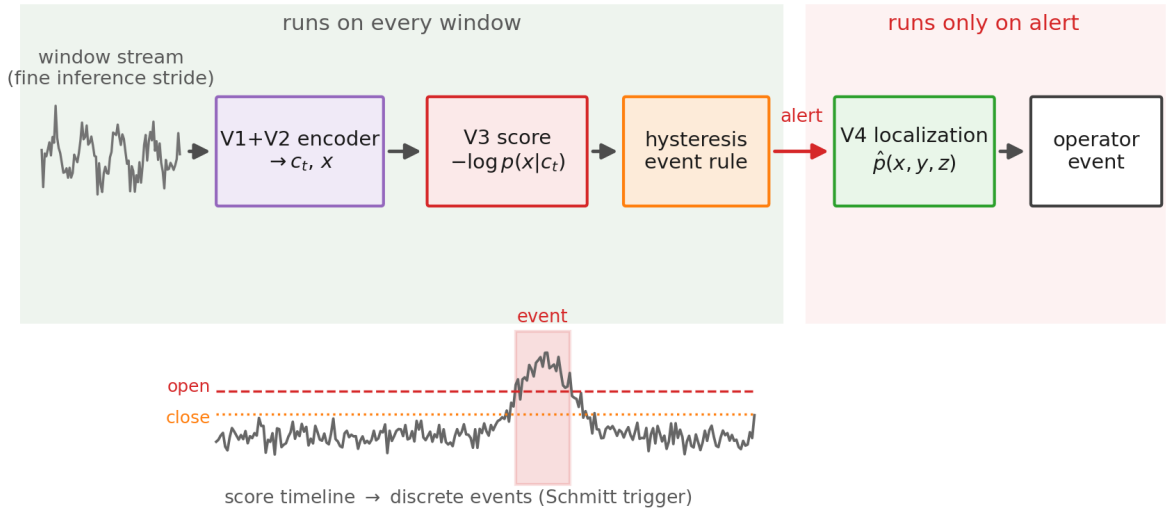


Figure 4.7: Streaming inference. The encoder, flow score, and hysteresis event rule run on every window at a fine inference stride; the localization head runs only on windows the anomaly gate flags. The Schmitt-trigger hysteresis (open above the cluster threshold, close at a lower bound) prevents one event from fragmenting as the score fluctuates around the boundary.

This completes the description of the architecture and the way it is trained and run. Each of its design commitments is, in the end, an empirical claim: that the context vector recovers the operating mode without labels, that conditioning the anomaly head on that context holds the false-positive rate down at mode boundaries, that the structure-borne path sharpens localization over an acoustic-only baseline, and that the supervisory pathway would add value if the signals were available. Chapter 5 sets out the data splits, metrics, baselines, and ablations that put each of these claims to the test, and Chapter 6 reports what they show.

Chapter 5

Experimental Evaluation

This chapter defines how the architecture of Chapter 4 is evaluated against the four research questions of Chapter 1, on the corpus of Chapter 3. It specifies the evaluation protocol, the metric attached to each question, the baselines the model is measured against, and the ablations that isolate the contribution of each design choice. The numerical outcomes belong to Chapter 6; the purpose here is to fix the rules under which those numbers are produced, so that every claim in the next chapter traces to a stated procedure rather than to a choice made after the result was seen.

Four properties of the data shape the design throughout. Operating-mode labels exist on only two of the five campaigns. Anomalies are annotated at the level of a recording rather than a window, so per-window anomaly ground truth does not exist for the field data. The labeled localization cohort is small, 16 positions in total. And because each labeled cohort is small, several of the metrics couple to quantities that are themselves stochastic, most visibly the K-means assignment used to score the operating context and to set the per-cluster thresholds, so a number read from a single training run can move under a different random seed. Each of the first three constraints rules out a conventional supervised metric; the fourth rules out single-run reporting. The protocol below answers all four.

The chapter proceeds in the order of the four research questions. §5.1 fixes the evaluation protocol and data splits that all four share, and §5.2 then attaches a metric to each question in turn: the operating context (RQ1), anomaly detection under domain shift (RQ2), source localization (RQ3), and supervisory signals (RQ4). The baselines against which each stage is measured follow in §5.3, and the ablations that isolate each design choice close the chapter in §5.4.

5.1 Evaluation Protocol and Data Splits

Label discipline. The central rule is that no label is used before the stage that is permitted to use it. The self-supervised encoder (§4.4) and the per-cluster anomaly threshold (§4.5) are fitted on healthy data alone; operating-mode labels enter only at evaluation time, where they are compared against the already-fixed K-means clusters of the context vector. Spatial labels are used only to supervise and to score the localization head. This keeps every label-conditioned claim free of leakage by construction, and it mirrors the intended deployment, in which mode and position annotations are not available at run time.

Splits. Train and validation partitions are drawn at the level of *recordings*, stratified by operating mode, so that the windows of one recording are never split across a fold and the leakage a window-level partition would invite is avoided. The healthy-versus-anomaly distinction is likewise a recording-level label. Within an anomaly-bearing recording the fault is sparse, and only a subset of windows actually contains it, so the detector is scored per window and the localization head is run only on the windows the detector gates as alerts (the gating rule of Chapter 4). The recording is the atomic unit of the partition; the window is the operative unit of detection and localization. For anomaly detection a held-out pool of healthy recordings is reserved both for fitting the per-cluster thresholds and for measuring the healthy alert rate.

Localization protocols. For localization, three complementary generalization protocols are used, in increasing order of strictness.

1. **Leave-one-recording-out.** Cross-validation folds over the labeled localization recordings, holding out one recording at a time. It is the loosest test, because two recordings at the same physical position can still fall on opposite sides of a fold.

2. **Leave-one-position-out.** Cross-validation closes that gap by folding over every labeled casing position in the D2, D3, D4, and D5 cohorts, 16 in total: for each fold the head is trained from scratch on all other positions and evaluated on the held-out one. The error is reported as a mean and standard deviation across folds, the standard small-sample generalization estimator, with the per-fold minimum and maximum given alongside as a supplementary range.

3. **Cross-dataset transfer.** Training is on one set of campaigns and evaluation on a disjoint one, for example training on D2, D3, and D4 and testing on the later D5

session, since D1 carries no spatial labels. This tests generalization across recording sessions and rig states rather than only across positions within a session.

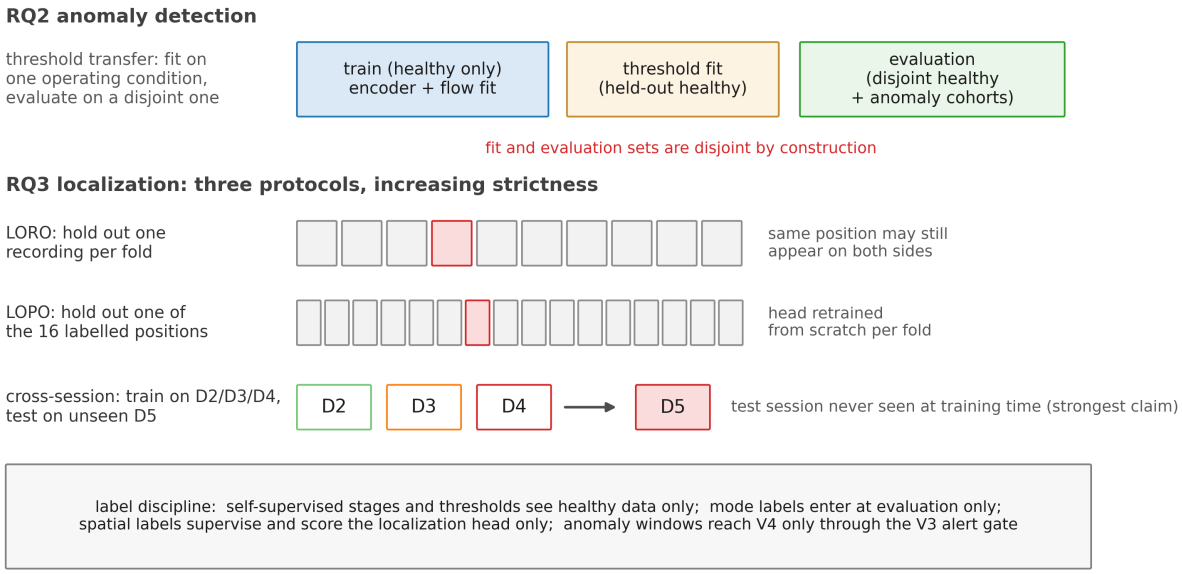


Figure 5.1: The evaluation protocol at a glance. All partitions are drawn at recording level; the anomaly threshold is fitted and evaluated on disjoint healthy subsets; and localization is tested under three protocols of increasing strictness: leave-one-recording-out, leave-one-position-out over all 16 labeled positions, and cross-session transfer to the never-seen D5 session. Every stage that depends on a stochastic fit is repeated over five seeds.

Seeds and significance. Two kinds of uncertainty are reported, and they are kept distinct. The first is sampling uncertainty at fixed training: a single-sample percentile bootstrap gives a confidence interval on a mean error or an area under the curve, a paired bootstrap on the per-window difference between a conditioned model and its unconditioned ablation gives both an interval on the difference and a two-sided p-value, and alert rates, being proportions, carry a Wilson score interval. The second is training uncertainty across random initializations, which at this cohort size is not negligible. Every stage that depends on a stochastic fit is therefore trained over a fixed set of five seeds, $\{42, 1337, 2024, 7, 99\}$, and the headline quantity is the distribution over those seeds, summarized as a median with the observed range, rather than a single point. This matters because the scoring of the operating context and the per-cluster thresholds both rest on a K-means assignment whose initialization varies with the seed, so a metric that couples to that assignment, the

mode-recovery score and the per-cohort alert rates in particular, can shift between runs that are otherwise identical. Reporting the spread separates a result that recurs across seeds from one that holds only on a single fortunate run. Where an ordering between two model variants is stable across the five seeds it is reported as a finding, and where it is not, the instability is itself the finding.

5.2 Metrics

The metrics are organized by research question. Each is chosen to be meaningful under the data constraints above rather than borrowed unmodified from a fully-supervised setting.

5.2.1 Operating Context (RQ1)

The first question is whether the context vector $\boldsymbol{c}_t$ recovers the operating mode without supervision. The context vectors of the labeled windows are clustered with K-means at $K = 3$, matching the three steady prototype modes, and the clusters are scored against the recorded modes. Normalized mutual information is the headline measure, because it is invariant to the mismatch between the cluster count and the number of labels present in a given validation subset. The score is reported on two cohorts: a held-out *sanity cohort* of labeled windows reserved as the fast model-selection gate during development, and, more strictly, the full set of labeled D1 and D2 windows; in both cases arbitrary cluster identities are matched to mode names by a Hungarian assignment at scoring time only, which the permutation-invariant mutual information does not itself require but the purity figure does. Because the K-means step is the stochastic element, the score is read as a distribution over the five seeds rather than a single value, and both its central tendency and its spread are reported, since a score reached on one seed need not recur on another.

5.2.2 Anomaly Detection under Domain Shift (RQ2)

The second question is whether conditioning the detector on the operating context suppresses the false alarms that mode changes provoke. Because per-window anomaly labels do not exist for the field recordings, the question is answered along several complementary axes rather than by a single supervised score. The axes are read together, as converging evidence.

Multiple comparisons. This design naturally introduces the multiple-comparison problem that a single pre-registered metric would avoid: examining five axes over several

cohorts, and selecting among combiner rules on those same cohorts, multiplies the opportunities for a chance result. No formal correction is applied across the RQ2 axes, and where a headline is the best of a set of rules examined on the same data, as with the late-fusion conjunction of §5.3, it is reported as a selected minimum and read with that selection in mind. A formal false-discovery-rate correction is applied only to the SCADA channel ranking of RQ4, where the number of tests is large and the correction is affordable; the RQ2 axes, fewer in number and interpreted jointly, are not corrected in the same way.

Calibration on healthy data. By construction the per-cluster threshold places the healthy alert rate near five percent on the fit set, so the held-out healthy alert rate is expected to sit near five percent as well, and a departure from that target is itself the diagnostic that the held-out distribution has drifted from the training distribution. This calibration is stressed by a recording-level transfer over the pooled healthy corpus: the alert rate is reported both under an in-distribution threshold, fitted and evaluated on disjoint windows of the same recordings, and under a threshold transferred from a disjoint set of recordings, the latter being the direct measure of whether a boundary calibrated on one set of healthy recordings survives a move to another. Because each campaign contributes essentially one healthy recording per operating condition, the held-out recordings are a different operating and acquisition condition than the fit set, spanning operating mode, recording session, and the `speed1`–`speed3` fan-noise level of the later campaigns. The speed levels are pooled when the encoder and the healthy density model are fitted (§3.3.1) and are not used as the split axis; they enter only as one component of the held-out variation and as a separate noise-level ablation. The healthy rate is reported in aggregate and audited per cohort, so calibration can be checked across the discovered modes rather than only in the largest one. Because the transfer split can place an easy or a hard condition in the evaluation set, this axis is run over several split seeds and reported as a range, not a single rate.

Response to mode transitions. This is the failure mode the architecture is built to suppress. Since the corpus contains no recorded transition events (Chapter 3), the synthetic transition cohorts constructed there, the feature-level and the encoder-level crossfades, stand in for the test: both are healthy by construction across the transition, so any alert raised in the crossfade region is a false positive, and the false-positive rate over those windows is the domain-shift metric, held to the same near-five-percent target as the steady healthy pool. This construct carries one inherent limitation. A linear crossfade of

two healthy windows is itself mildly off-manifold, so the crossfade rate is an upper-bound stress rather than a like-for-like contest between detectors. A low rate on it is consistent with correct transition handling, but it is equally consistent with a detector that is simply insensitive to off-manifold inputs, and silence on the crossfade is therefore not by itself evidence of correct behavior. The axis is weighted accordingly, as one input among several rather than a decisive one.

Controlled detection curve. A held-out healthy window is perturbed by additive Gaussian noise of a calibrated magnitude in the latent feature space, which yields window-level ground truth, clean versus corrupted, on which the anomaly score becomes a genuine binary discriminator, and the area under the receiver operating characteristic is computed over a ladder of signal-to-noise ratios with a bootstrap interval at each rung. This probe has inherent limitations by design. The perturbation lives in latent space and does not correspond to any physical fault, and because the anomaly score is a negative log-likelihood, isotropic latent noise can move a window toward the higher-likelihood bulk of the density and so lower its score, which means the absolute area under the curve can fall below the 0.5 chance line even for a detector that behaves sensibly on real data. For this reason the curve is always read against the explicit 0.5 baseline, and the informative quantity is the lift the conditioning adds over its unconditional ablation at each rung rather than the absolute level. For fine comparisons between near-identical variants the validation negative log-likelihood is used alongside the curve, since it stays discriminative where the easy end of the ladder does not. The axis measures sensitivity to a controlled off-manifold perturbation, not to a real fault, so it complements and does not replace the cohort alert rates; a perturbation injected in the input signal, where it would carry physical meaning and the absolute level would be interpretable, is the fairer complement and is identified as future work. Because the lift is itself seed-sensitive, the ladder is reported across the five seeds, and a conditioning advantage that does not survive every seed is reported as partial rather than general.

Ranking of alert rates across the labeled anomaly cohorts. Because the anomaly density inside an anomaly recording is unknown, precision and recall pooled per window over a whole recording are not interpretable; what is interpretable is that the alert rate on an anomaly-bearing cohort should rank well above the healthy rate, in the order the qualitative descriptions of the faults imply.

Scoring against weak temporal labels. Although the corpus carries no per-window annotation, an impulsive fault is recoverable from the signal envelope without manual labeling, so for every anomaly-bearing recording a set of weak ground-truth intervals is derived from the multi-channel Hilbert envelope by iterative peak-picking. The detector is run at a fine inference stride, its per-cluster alerts are grouped into events, and each event is matched to the derived intervals: an interval with an overlapping alert is a true positive, an alert overlapping no interval is a false positive, and an unmatched interval is a false negative. This yields the event-level precision, recall, and F1 that the per-window pooling of the fourth axis cannot. Because the labels are envelope-derived rather than hand-annotated, and are moreover drawn from the same multi-channel Hilbert envelope that the vibration encoder ingests as its input channel 1 (Chapter 3), the F1 is not an independent validation but a consistency check between the learned detector and an envelope peak-picker; it is therefore read as a relative quantity for model selection across seeds and capacity settings, not as a calibrated absolute, and it too is reported with its seed-to-seed spread.

5.2.3 Source Localization (RQ3)

The third question is how accurately the head locates a detected anomaly, and how much the structure-borne path contributes. The error metric is the three-dimensional Euclidean distance between the predicted and the labeled position, reported as a mean in meters with a bootstrap interval, and additionally as a 95th percentile for the cross-session transfer test. The generalization protocols are the leave-one-recording-out, leave-one-position-out, and cross-dataset splits of §5.1. The contribution of each sensing path is isolated by re-running the same head in four input modes: the full acoustic-and-vibration model, an acoustic-only mode that uses the steered-power map alone, a classical vibration-only mode driven by accelerometer time-difference multilateration, and a learned vibration-only mode. The structure-borne wave speed in the printed casing is set to the material-correct value for the printed plastic ($2000\,\mathrm{m\,s^{-1}}$), not swept or tuned to the result; calibrating the exact wave speed of an individual printed part is left to future work. The localization error is additionally broken down per held-out position, so that a systematic failure can be attributed to the geometry of the sensor array rather than to signal quality.

5.2.4 Supervisory Signals (RQ4)

The fourth question cannot be answered on the prototype corpus, which carries no temperature, pressure, or SCADA channels (Chapter 1); it is therefore treated as exploratory and evaluated by two limited checks. The first exercises the supervisory conditioning pathway of §4.7 using the recorded operating-condition token as a stand-in supervisory variable, and measures whether feeding it through the conditioning path changes the localization error, which establishes that the mechanism is wired correctly. The second mines the real ROW II SCADA archive offline, with no model trained. After auditing which of the archive's channels carry a live signal, since the anomaly target depends on the instrumentation that was recording, each process channel is ranked by two complementary statistics, the mutual information and a rank-based area under the curve, for the dependence it shows on the operating mode and, separately, on historical anomaly events. Because the one-second SCADA is strongly autocorrelated, significance is assessed with a circular-shift permutation null and a Benjamini-Hochberg correction, and the anomaly test is also run within each operating mode to remove the between-mode confound. As a robustness check the anomaly target is rebuilt from the live accelerometer channels alone and the ranking re-run, so that a degraded legacy target cannot mask an association. The ranking identifies which channels a field deployment would route into the supervisory vector. Both checks frame RQ4 as a deployment recommendation to be validated on instrumented field data, not as a result established on the prototype.

5.3 Baselines and Fusion Paradigms

The proposed model is measured against established alternatives at every stage, so that each reported gain is relative to a credible reference rather than to nothing.

Anomaly baselines. The detection head is compared against the full set of unsupervised acoustic baselines that defined the prior ROW II study [65], reproduced here as the reference point this work must improve on.

1. **LSTM autoencoder.** A long short-term memory autoencoder scored by per-window reconstruction error.

2. ***k*-means.** A *k*-means model scored by the distance to the nearest healthy centroid.

3. **One-class SVM.** A one-class support vector machine scored by its signed distance to the healthy boundary.

4. **Kernel-density estimator.** A Gaussian kernel-density estimator scored by negative log-likelihood.

The autoencoder consumes the log-mel window sequence directly; the other three consume a time-aggregated embedding of the same windows, so a window seen by every baseline is the identical slice of signal and the comparison turns only on the scoring mechanism. Each baseline is fitted on healthy windows pooled across campaigns exactly as the proposed head is, and is run for the acoustic and the vibration streams alike, with one unavoidable exception noted in Chapter 6. The recurrent autoencoder requires fixed-length sequences and so cannot pool the variable-rate accelerometer stream across campaigns, for which the flat-feature scorers stand in. All four are thresholded by the same per-cluster percentile rule as the proposed head, fit on a held-out healthy cohort rather than on their own training scores, so the comparison isolates the scoring mechanism rather than the thresholding. They are then scored on the same RQ2 axes of §5.2.2: the within-campaign healthy-versus-anomaly ROC-AUC that the prior study reported as its headline number, and the held-out healthy alert rate under both an in-distribution threshold and a threshold transferred to an unseen operating condition, the latter read over several split seeds so that a single fragile split is not mistaken for a property of the scorer.

Context reference. For the operating-context question the label-free encoder is compared against two references on the hand-engineered features. The supervised reference is a gradient-boosted decision-tree classifier trained on hand-engineered acoustic and vibration features using the mode labels of D1 and D2, scored on held-out recordings rather than held-out windows, intended as the labeled upper reference. The unsupervised reference is a K-means clustering of the same hand-engineered features against the mode labels, a peer baseline that measures how separable the modes are in hand-engineered space. Both are computed and reported in §6.3, and both are encoder-independent, so they are computed once on the features and do not enter the per-seed sweep. The supervised reference is informative only where a labeled campaign is large enough to estimate it, which on this corpus neither is: one labeled campaign has too few recordings for a recording-level train and validation split, and the other collapses onto its majority class under a two-recording held-out split. The comparison therefore establishes that the mode signal is present in the features without claiming a stable supervised ceiling.

Localization baseline. For localization, the references are two classical estimators computed directly from the raw signals at the same ground-truth resolution and under

the same Euclidean error metric: the SRP-PHAT estimator on the microphone array, and an accelerometer time-difference multilateration solver on the structure-borne channels. These are the numbers the conditional head must improve on, and they stand in for the purely classical pipeline that the learned components are meant to augment [34].

Fusion paradigms. The architecture commits to intermediate fusion, combining the two modalities at the feature level. To justify that commitment, the same task is also solved by the two competing paradigms built from the same unimodal pipelines. Unimodal references use one modality alone. Late-fusion references combine the two unimodal outputs without a shared representation: for detection, by logical AND and OR of the per-modality alerts, whose per-cohort co-firing (acoustic-only, vibration-only, both, neither) is decomposed to show why a conjunction is specific, by a logistic combination of the standardized scores, and by their element-wise maximum; for localization, by a uniform average of the two position estimates, by a per-axis least-squares-weighted average, and by a confidence-gated choice that trusts whichever pipeline applied the smaller correction to its own initial estimate. Several combiners are therefore examined on the same cohorts, so a combiner carried forward as the operating point is the best of that set rather than a pre-registered choice, and the in-sample combiners, the score-weighted and element-wise-maximum rules whose weights are fitted on the evaluation cohort, are upper bounds rather than deployable rates. The three paradigms are compared head-to-head on the same error metric with bootstrap intervals, which is the test of whether the shared representation earns its added complexity; the contribution of context conditioning within the intermediate model is isolated separately by the paired-bootstrap ablation of §5.4.

5.4 Ablation Protocol

The ablations remove one design element at a time, leaving the rest of the chained system intact, so that each element's contribution is attributable. Two properties of the architecture make this clean. The conditioning projections are zero-initialized (§4.5, §4.6), so setting the context to zero reduces each head exactly to its unconditional form rather than to an untrained one; and the input modalities can be muted independently, so a modality's contribution can be read off directly. The first experiments below remove one such element at a time, conditioning and fusion and then modality; the last is a set of model-selection sweeps, grouped here because each is read under the same per-seed discipline.

Conditioning. The headline ablation removes context conditioning. The anomaly head is compared against an unconditional flow that sees a zero context vector, and the localization head against an unconditional residual, in both cases with all other weights held fixed. The comparison is made on paired per-window scores, so that the paired bootstrap of §5.1 can test significance at fixed training, and it is repeated across the five seeds, because the conditioning effect is one of the quantities that can move between initializations. A conditioning advantage is therefore claimed only where it survives that spread; a paired gain visible on a single seed but not across the set is treated as inconclusive. The fusion stage is ablated in the same spirit by scoring the operating-context metric on the per-modality summaries instead of the fused context vector, which isolates what cross-modal fusion adds to the representation. In addition, the dependence of the fused context vector on each input modality is probed directly by zeroing the vibration stream at the input and measuring the resulting change in the context vector, both as a cosine similarity and as the ratio of the vector's sensitivity to the two streams.

Modality contribution. Muting the vibration input while leaving the fusion machinery in place measures how much the structure-borne stream contributes to detection, and the four localization input modes of §5.2.3 do the same for localization, separating the classical and the learned use of the vibration timing.

Hyperparameter and capacity selection. The remaining experiments are model-selection sweeps rather than single-element removals. The encoder (its convolutional width), the conditional flow (its regularization and capacity), and the localization head (its regularization, capacity, residual range, and target-position augmentation) are each swept over a grid, with the winner chosen on a held-out selection metric and carried into the protocols above. The exact grids and the selected value on each axis are listed in Appendix A (Table A.4), and the complete per-stage hyperparameter configuration and trained parameter counts are tabulated there. Each selected configuration is then re-trained over the five seeds, and the reported quantity is the distribution of the selection metric rather than its value on a single run, so that an ordering which holds only on one seed is exposed rather than presented as a stable optimum. This is the same discipline applied to every headline number in the next chapter: a median with a range, and an explicit note wherever the seed-to-seed variation is large enough to qualify the conclusion.

Chapter 6

Results

This chapter reports the empirical results of the chained system. The five campaigns come from Chapter 3, the protocol and metrics from Chapter 5, and the structure from the four research questions of Chapter 1. Two of those questions, anomaly detection (RQ2) and source localization (RQ3), are where multimodality is expected to help. For both, the three standard fusion paradigms are compared head to head: unimodal, late fusion, and intermediate fusion. The headline is a paradigm comparison, not a single-architecture verdict. Different paradigms win on different cohorts and operating points, and the chapter reports which.

Every number that depends on a stochastic fit is reported as a distribution over the five seeds $\{42, 1337, 2024, 7, 99\}$ of §6.1, written as a median with the observed across-seed range in brackets. Where a quantity is stable across seeds it is stated as a finding; where it is not, the spread is reported and the instability is itself the result. Several cohorts recur across the RQ2 artifacts; Table 6.1 fixes their definitions and sizes once so that later tables can be read against it.

6.1 Experimental Setup

Training runs on a single consumer GPU (an NVIDIA RTX 3080), and is CPU-capable at a higher wall-clock cost. The anomaly stage takes about an hour and the localization stage about forty minutes. The threshold-fit discipline of Chapter 5 holds throughout: per-cluster percentile thresholds are fitted on a held-out healthy subset, and recall and false-positive rates are then reported on a disjoint subset. Splits are at the recording level everywhere. The five-seed sweep $\{42, 1337, 2024, 7, 99\}$ supports every stability statement; where a single representative run is shown, for example a score-distribution plot, the seed is named in the caption.

Table 6.1: Healthy and labeled cohorts used across the RQ1 and RQ2 artifacts. The RQ1 strict cohort and the RQ2 healthy hold-out both draw on the pooled labeled D1+D2 windows, which is why their sizes nearly coincide; they differ only in how each window is scored (cluster-versus-mode NMI as against the per-cluster alert rate). Window counts are identical across the five seeds; the seed varies the K-means assignment and the threshold, not the recording split.

Cohort	Role	n (windows)
RQ1 sanity gate	development model-selection NMI	4 018
RQ1 strict	headline NMI, all labeled D1+D2	6 773
RQ2 threshold-fit / val	per-cluster threshold fit, NLL paired test	$\approx$ 1 802
RQ2 healthy hold-out	per-cohort alert and specificity tables	6 774
RQ2 anomaly D2 / D3 / D4	per-cohort recall	1 069 / 266 / 26 996

One acquisition setting matters for what follows. The casing is 3D-printed plastic, so the structure-borne wave speed for the vibration time-difference solver is set to the plastic value of 2000 $\mathrm{m\,s^{-1}}$ rather than a metallic one. The choice is consequential at this scale: an over-estimated wave speed compresses the time-difference geometry and suppresses the vibration localization channel, so the material-correct value is used throughout.

6.2 Reference Baselines

Table 6.2 collects the non-learned V0 references (the baseline stage that precedes the learned V1 to V4 pipeline) each learned paradigm must beat. The classical SRP-PHAT estimator localizes to within roughly 0.25 to 0.42 m per recording. The accelerometer-TDOA multilateration solver, now using the corrected plastic wave speed, is of the same order.

Table 6.2: V0 reference baselines. ROC-AUC is within-campaign healthy-versus-anomaly; localization is per-recording 3-D Euclidean MAE in meters. A dagger marks the rate-incompatible pooled vibration autoencoder (footnote[1]).

Reference	Metric	D2	D3	D4
LSTM-AE (acoustic)	ROC-AUC	0.94	0.84	0.63
LSTM-AE (vibration)	ROC-AUC	$0.81^{\dagger}$	n/a	n/a
SRP-PHAT (acoustic)	MAE (m)	0.422	0.336	0.247
accel-TDOA (2000 $\mathrm{m\,s^{-1}}$)	MAE (m)	0.491	n/a	0.361

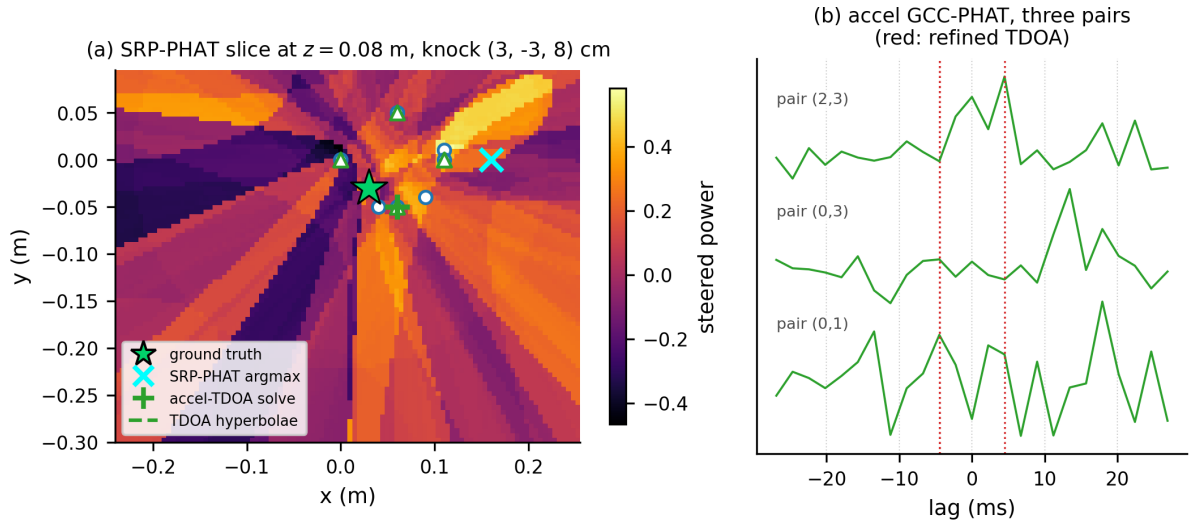


Figure 6.1: One real D5 knock burst through the classical front-end. (a) The SRP-PHAT steered-power slice at the knock height with the acoustic argmax, the accelerometer-TDOA solve, the ground truth, and the TDOA hyperbolae of the three strongest accelerometer pairs. (b) The accelerometer GCC-PHAT correlations those hyperbolae come from, with the parabolic sub-sample refinement marked. These are the inputs the learned V4 head starts from.

Table 6.3 reproduces the full prior-work acoustic-anomaly suite of Khamaisi et al. [65], namely an LSTM autoencoder, a k-means scorer, and a one-class SVM, plus a Gaussian kernel-density estimator. All are run on the proposed head's own RQ2 protocol and pooled across campaigns. Every scorer calibrates to the 0.05 target in-distribution, which validates the thresholding protocol; the intervals on the in-distribution rate confirm it. What the in-distribution rate does not show is whether that calibration survives a move to an unseen operating condition, which is the subject of §6.4.2.

Vibration detection (ROC-AUC 0.87 to 0.90) clearly beats acoustic (0.57 to 0.67) on this rig, consistent with the higher-rate D4 and D5 accelerometer cohorts. No unconditional baseline, however, is both a strong detector and robust to a threshold transfer, which is the gap the context-conditioned head and the late-fusion conjunction are built to close.

[1]The recurrent autoencoder needs fixed-length sequences, but the accelerometer rate varies across campaigns (4 Hz to 446 Hz), so the five-campaign pool yields incompatible window lengths. On the same-rate D1+D2 (4 Hz) cohort it reaches ROC-AUC 0.70. The flat-feature scorers below are unaffected because they time-aggregate each window to a fixed-width vector.

Table 6.3: Unconditional anomaly baselines, pooled across campaigns, at seed 42. ROC-AUC is the within-campaign detection metric; FPR in-dist is the held-out healthy alert rate under an in-distribution threshold, fit and evaluated on disjoint windows of the same held-out conditions. Intervals are 95% (bootstrap on ROC-AUC, Wilson on FPR). The transferred-threshold rate, where these scorers fail, is reported separately in Table 6.8.

Modality	Model	ROC-AUC	FPR in-dist
acoustic	LSTM-AE	0.57 [0.56, 0.58]	0.056 [0.045, 0.069]
acoustic	k-means	0.66 [0.66, 0.67]	0.042 [0.033, 0.054]
acoustic	OC-SVM	0.66 [0.66, 0.67]	0.044 [0.035, 0.056]
acoustic	KDE	0.67 [0.67, 0.68]	0.061 [0.049, 0.074]
vibration	k-means	0.87 [0.86, 0.87]	0.052 [0.041, 0.064]
vibration	OC-SVM	0.90 [0.89, 0.90]	0.045 [0.035, 0.057]
vibration	KDE	0.90 [0.89, 0.90]	0.054 [0.043, 0.066]

6.3 RQ1: Operational Mode Discovery

RQ1 asks one question: does the fused context vector $\boldsymbol{c}_t$ cluster into the operating modes without labels? Table 6.4 reports normalized mutual information against the recorded modes. The headline is the *strict* cohort, all labeled D1+D2 windows ($n = 6\,773$); the sanity-gate cohort that served as the development model-selection gate is reported alongside as a secondary, and necessarily optimistic, number, because selecting on it and then reading it back is not an out-of-sample estimate. Because the strict cohort contains the sanity-gate windows as a subset, the strict NMI is read as an across-seed stability estimate of a selected configuration rather than as a fully held-out generalization number.

Table 6.4: RQ1 mode-discovery NMI, median [min, max] over five seeds at $K = 3$. The headline column is the strict cohort (all labeled D1+D2 windows, $n = 6\,773$); the sanity-gate column ($n = 4\,018$) is the development selection gate and is reported only for context. The fusion ablation is defined on the sanity cohort only. The K-means floor is a single deterministic reference on the strict cohort; the supervised ceiling cannot be estimated on this corpus (see text).

Encoder	strict K=3 NMI	sanity-gate NMI
V1 acoustic	0.386 [0.370, 0.544]	0.902 [0.605, 1.000]
V1 vibration	0.117 [0.072, 0.136]	0.117 [0.045, 0.182]
V2 multimodal fusion	**0.409 [0.177, 0.563]**	0.842 [0.336, 0.869]
V2 minus vibration (ablation)	–	0.506 [0.000, 0.670]
K-means floor (handcrafted)	0.432	–

The headline is that the label-free encoder recovers operating mode at a strict NMI of about 0.41 (fused context vector, median over seeds), with a wide seed-to-seed spread. On the strict cohort the fused encoder (0.409) and the acoustic-only encoder (0.386) are statistically indistinguishable, their ranges overlapping almost completely, so the fused representation is acoustic-dominated and multimodal fusion does not deliver a clear, seed-robust gain for mode discovery at this data scale. The much higher sanity-gate numbers (0.842 fused, 0.902 acoustic) belong to the cohort used to select the model during development and are not the result; reporting them as the headline would be reporting performance on the selection set.

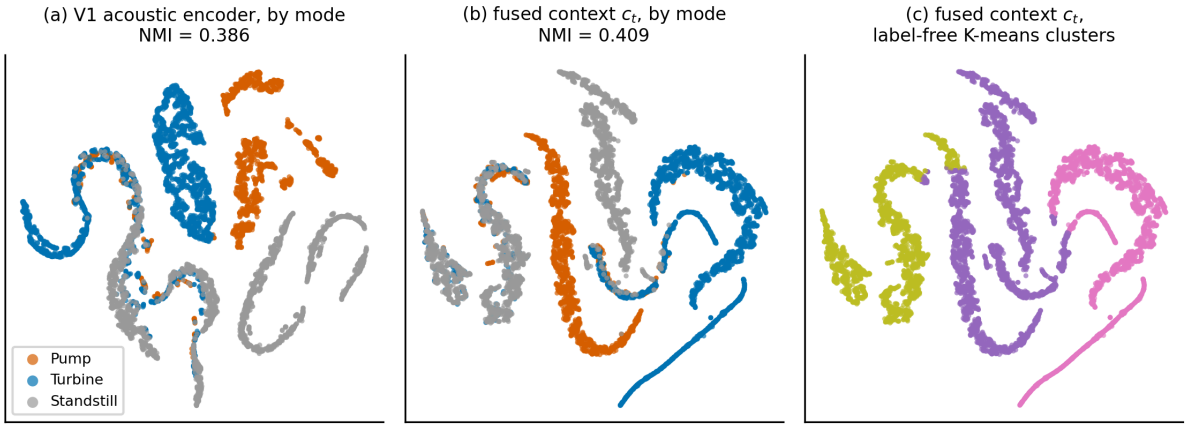


Figure 6.2: t-SNE of the label-free representations on all labeled D1+D2 windows (the strict cohort). (a) The V1 acoustic encoder separates the operating regimes; (b) the fused context vector $\boldsymbol{c}_t$ preserves but does not improve on that structure; (c) K-means on $\boldsymbol{c}_t$ recovers the mode partition without ever seeing a label. The panel NMIs are the strict-cohort values of Table 6.4.

The acoustic dominance is measured, not asserted. On the strict cohort the modality probe scores the fused vector, the acoustic stream alone, and the vibration stream alone: the fused and acoustic-only NMI differ by about 0.003, while vibration alone reaches only 0.04. Removing vibration from the fusion does lower the development-cohort NMI (median 0.842 $\rightarrow$ 0.506), so the vibration stream is not ignored outright; but on the out-of-sample strict cohort it adds nothing the acoustic trunk does not already carry. RQ1 is therefore a defensible negative on multimodality: at this data scale the fused pool clusters on its acoustic content.

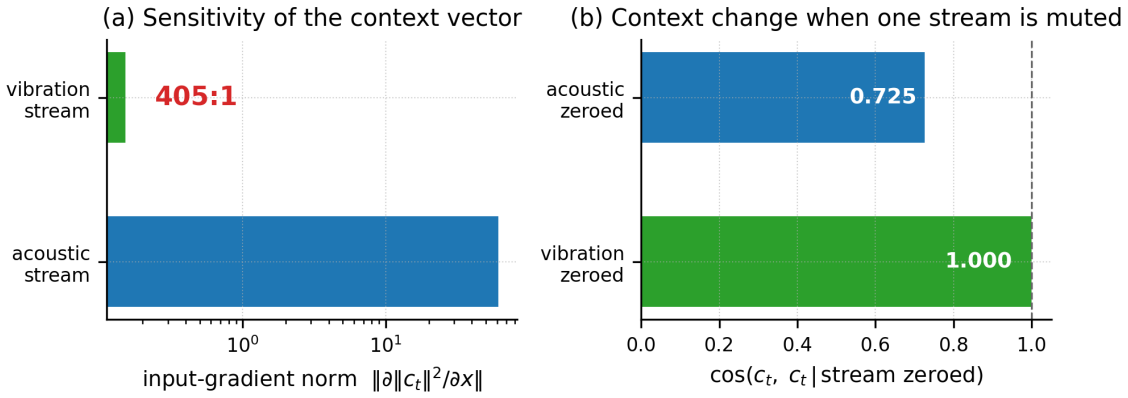


Figure 6.3: Evidence that the fused pool clusters on its acoustic content. (a) The fused vector's input-gradient sensitivity to the acoustic stream far exceeds its sensitivity to vibration. (b) Zeroing the vibration input leaves $\boldsymbol{c}_t$ almost unchanged, whereas zeroing the acoustic input moves it substantially.

This label-free number is read against two references on the same cohort, both computed on the hand-engineered features. The unsupervised K-means floor, a clustering of the hand-engineered features against the mode labels, reaches NMI 0.432 (purity 0.719). The learned encoder therefore matches this floor rather than clearing it: its strict NMI of 0.41 is comparable to the floor's 0.432, a touch below it at the median and above it on the higher-scoring seeds (up to 0.56 NMI and 0.83 purity). The reading is that the operating regimes are cleanly separable in this data, and that the label-free encoder recovers them at hand-engineered quality without any feature engineering or labels, not that it improves on hand-engineering for this particular task. The supervised ceiling cannot be estimated on this corpus: D1 carries too few mode-labeled recordings for a recording-level train and validation split, leaving the classifier with no training fold, and D2's two-recording held-out split collapses the classifier onto its majority class (macro-F1 0.333). The corpus is therefore too small to fix a supervised bound, and the strict NMI is best read on its own terms.

6.4 RQ2: Anomaly Detection under Domain Shift

RQ2 is answered in four parts, in the order the subsections below take them: which paradigm is most specific (this section), what conditioning adds to the flow (§6.4.1), whether the calibration survives a domain shift (§6.4.2), and what a mode transition does to it (§6.4.3).

Three anomaly heads share identical splits, with per-cluster thresholds fit unsupervised on the healthy subset: a unimodal acoustic flow, a unimodal vibration flow, and the intermediate-fusion flow conditioned on $\boldsymbol{c}_t$. The late-fusion rules combine the two unimodal decisions after training. Table 6.5 reports the per-cohort alert rates across the five seeds.

Table 6.5: RQ2 per-cohort alert rates, median [min, max] over five seeds. Healthy FPR is the held-out healthy alert rate (target 0.05); the cohort columns are the per-window alert rate on each anomaly-bearing cohort, not recall in the precision/recall sense, since those cohorts are sparsely anomalous (most windows are healthy background) so a true per-window recall is not directly interpretable. The `score_weighted` and `MAX` rows are in-sample upper bounds (the combiner is fitted on the test cohorts). Cohort sizes are in Table 6.1. The wide ranges are the result: the per-cohort alert rate is strongly seed-dependent (the mechanism is taken up in §6.7).

Paradigm	Rule	Healthy FPR	D2	D3	D4
Unimodal	V3-acoustic	0.131 [.01,.31]	0.094 [.00,1.0]	0.962 [.02,1.0]	0.546 [.00,.59]
Unimodal	V3-vibration	0.030 [.00,.04]	0.143 [.00,.27]	0.045 [.01,.15]	0.675 [.32,.74]
Late fusion	**AND**	**0.003 [.00,.01]**	0.007 [.00,.05]	0.045 [.00,.15]	0.200 [.00,.45]
Late fusion	OR	0.160 [.04,.31]	0.318 [.02,1.0]	0.962 [.04,1.0]	0.683 [.59,.89]
Late fusion	score_weighted*	0.050 [.05,.05]	0.320 [.05,.56]	0.959 [.00,1.0]	0.829 [.63,1.0]
Late fusion	MAX*	0.050 [.05,.05]	0.271 [.00,.45]	0.962 [.26,1.0]	0.685 [.41,.96]
Intermediate	**V3-fusion**	0.072 [.04,.09]	0.309 [.00,.74]	0.895 [.23,1.0]	0.319 [.09,.53]

Two findings survive the seed spread, and they are the ones worth reading. The first is specificity. Table 6.6 reports the fraction of windows on which only acoustic fires, only vibration fires, or both fire. On the healthy hold-out the share on which *both* modalities fire is 0.003 ($[0.001, 0.009]$ across seeds), so the parameter-free AND conjunction inherits a near-zero healthy alert rate (0.003, the tightest and lowest of any rule in Table 6.5). The second is the cost of that specificity: the AND rule's recall is low and seed-dependent (0.200 on D4 but ranging down to 0, 0.045 on D3, 0.007 on D2), because it fires only where two independent detectors agree and that agreement is itself unstable. The single-modality flows recover more faults but at a worse and more variable healthy rate (the acoustic flow alerts on 0.131 of healthy windows, ranging to 0.307).

Table 6.6: RQ2 specificity audit, median [min, max] over five seeds: fraction of windows per cohort on which only acoustic fires, only vibration fires, or both fire (the remainder is neither). The robust quantity is the near-zero both-fire share on healthy windows.

Cohort	a-only	v-only	both
healthy hold-out	0.122 [.01,.31]	0.025 [.00,.04]	**0.003 [.00,.01]**
D2 anomaly	0.050 [.00,1.0]	0.089 [.00,.23]	0.007 [.00,.05]
D3 anomaly	0.846 [.02,.99]	0.000 [.00,.02]	0.045 [.00,.15]
D4 anomaly	0.144 [.00,.36]	0.311 [.12,.68]	0.200 [.00,.45]

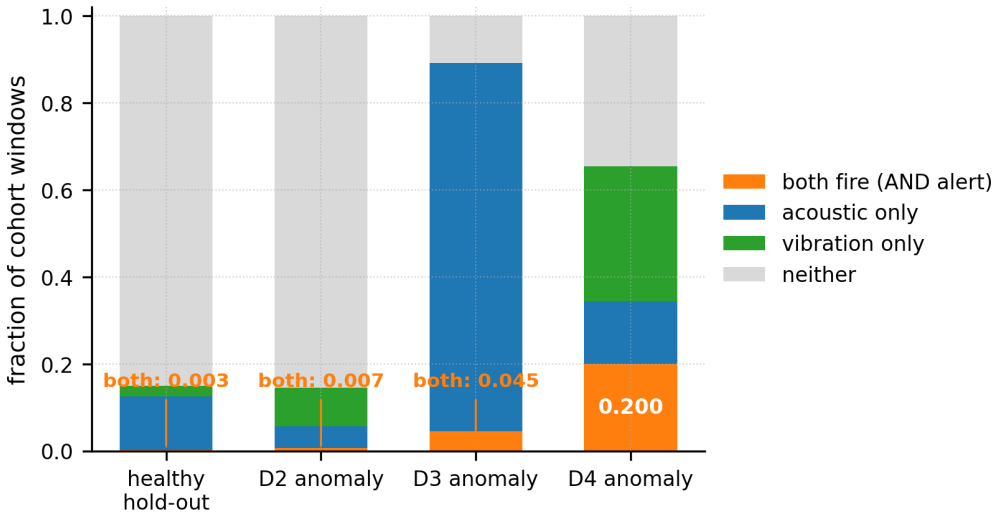


Figure 6.4: The specificity audit of Table 6.6, stacked per cohort. The two modalities almost never co-fire on held-out healthy windows (both-fire share 0.003), so the parameter-free AND conjunction inherits a near-zero healthy alert rate. Co-firing on the anomaly cohorts is real but seed-dependent.

The RQ2 verdict follows, and it is a paradigm trade-off rather than a single winner. The late-fusion AND rule is the high-specificity operating point: its healthy alert rate is near zero and, as §6.4.2 shows, it is the only rule that also stays calibrated under a threshold transfer. Its recall is the price, low and seed-sensitive. The single-modality flows and the OR rule detect more but raise several times the healthy alarms. The intermediate-fusion flow sits between them, with a moderate healthy rate (0.072) and a recall that collapses on the largest cohort (D4 median 0.319, ranging to 0.087). Its joint flow trains away part of the vibration anomaly signal there. One caveat of selection attaches to the AND verdict: it is the best of several late-fusion combiners examined on the same cohorts (Table 6.5), two of which are flagged as in-sample upper bounds.

Its near-zero healthy rate is therefore the minimum over that set and is read with that selection in mind. Its appeal rests on a structural argument rather than the number alone: a conjunction of two independent detectors can only lower the healthy alert rate, so the under-shoot is a property of the rule.

6.4.1 Impact of Conditioning on the Flow

A separate analysis isolates what conditioning adds to the intermediate flow itself. Removing it, by replacing the FiLM modulation with a zero context vector, degrades the held-out negative log-likelihood by a paired margin of $\Delta\mathrm{NLL} = 0.055$ (median over seeds, $[0.016, 0.116]$; at seed 42, 95% CI $[0.053, 0.057]$, $p < 0.001$, $n = 591$ paired windows). The direction is stable across all five seeds, so the conditional flow is a measurably better density model of the healthy manifold. The magnitude, however, is minute: against scores of order -241 the margin is about two hundredths of one percent, so the improvement is consistent but small, and it should be described that way rather than as a large gain.

The conditional flow also competes with its unconditional ablation on the controlled latent-perturbation curve (Table 6.7). This probe has inherent limitations by design: additive isotropic Gaussian noise in the latent space does not imitate a fault, and because the anomaly score is a negative log-likelihood it can move a window toward the higher-likelihood bulk of the density and so lower its score, which is why the absolute area under the curve can fall below the 0.5 chance line. The curve is therefore read against that explicit baseline. Across the five seeds the conditional flow's median area stays above 0.5 at every rung (0.56 to 0.78), while the unconditional ablation falls below chance on some seeds (median 0.49 to 0.58); the conditioning lift is positive at the median throughout ($+0.02$ to $+0.20$) but its range overlaps, so the advantage is real on average and seed-dependent in detail. The informative quantity is the lift, not the absolute level; an input-space perturbation, where the noise would carry physical meaning, is the fairer complement and is left to future work.

Table 6.7: RQ2 latent-perturbation ROC-AUC ladder, median [min, max] over five seeds: conditional flow against its unconditional ablation, and the median conditioning lift, at each latent signal-to-noise ratio. The 0.5 chance line is the reference; the conditional median is above it at every rung.

Latent SNR (dB)	Conditional	Unconditional	Lift
−10	0.777 [0.596, 0.957]	0.575 [0.373, 0.736]	+0.202
−5	0.692 [0.546, 0.973]	0.558 [0.378, 0.703]	+0.134
0	0.569 [0.518, 0.958]	0.552 [0.365, 0.711]	+0.017
+5	0.563 [0.554, 0.910]	0.518 [0.374, 0.731]	+0.045
+10	0.583 [0.532, 0.836]	0.491 [0.420, 0.665]	+0.092

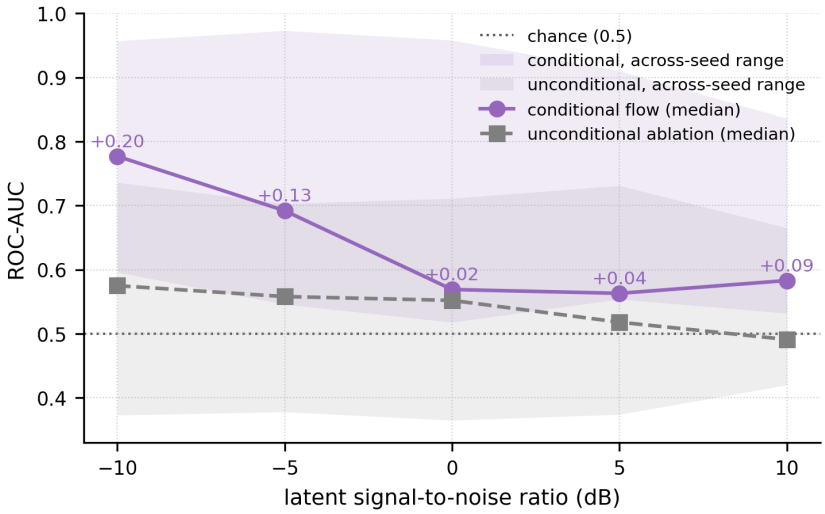


Figure 6.5: The latent-perturbation ROC-AUC ladder of Table 6.7, with the 0.5 chance line marked and a shaded across-seed band. The conditional median lies above chance at every rung; the unconditional ablation dips below it on some seeds. The probe is weak by design, so the informative quantity is the lift, not the absolute level.

6.4.2 DDomain-Shift Robustness

The decisive RQ2 measurement places baselines and heads on a single threshold-transfer axis: each per-cluster threshold is fitted on one set of healthy operating conditions and the held-out healthy alert rate is read off a disjoint set. Because the transfer split can land an easy or a hard condition in the evaluation set, the rate is reported over the five split seeds. One caveat bounds the interpretation. Because each campaign contributes essentially one healthy recording per operating condition, the fit and evaluation sides differ along several axes at once: operating mode, recording session, fan-noise level, and

the campaign's vibration-acquisition mode (peak-amplitude versus raw-waveform). The result therefore demonstrates robustness to a joint change of operating and acquisition condition, not to a change of operating mode in isolation; the two cannot be separated on this corpus. Table 6.8 gives the result, and it is the strongest in RQ2.

Table 6.8: RQ2 domain-shift robustness: held-out healthy false-positive rate after the per-cluster threshold is transferred to an unseen operating condition, median [min, max] over five split seeds, with the number of seeds on which the rate collapses above 0.2. Every unconditional baseline and every single acoustic or fusion head collapses on at least one seed; only the vibration-conditioned flow and the parameter-free AND rule never do, and the AND rule has the lowest and tightest range.

Method	Modality	FPR shift	seeds collapsed
OC-SVM	acoustic	0.00 [0.00, 1.00]	1/5
k-means	acoustic	0.00 [0.00, 0.34]	1/5
KDE	acoustic	0.00 [0.00, 0.37]	1/5
OC-SVM	vibration	0.06 [0.02, 0.92]	2/5
k-means	vibration	0.06 [0.04, 0.90]	1/5
KDE	vibration	0.25 [0.05, 0.93]	3/5
V3-acoustic (head)	acoustic	0.20 [0.11, 0.31]	3/5
V3-vibration (head)	vibration	0.04 [0.03, 0.12]	0/5
V3-fusion (head)	fusion	0.16 [0.03, 0.36]	2/5
Late-fusion AND (head)	fusion	**0.004 [0.001, 0.012]**	**0/5**

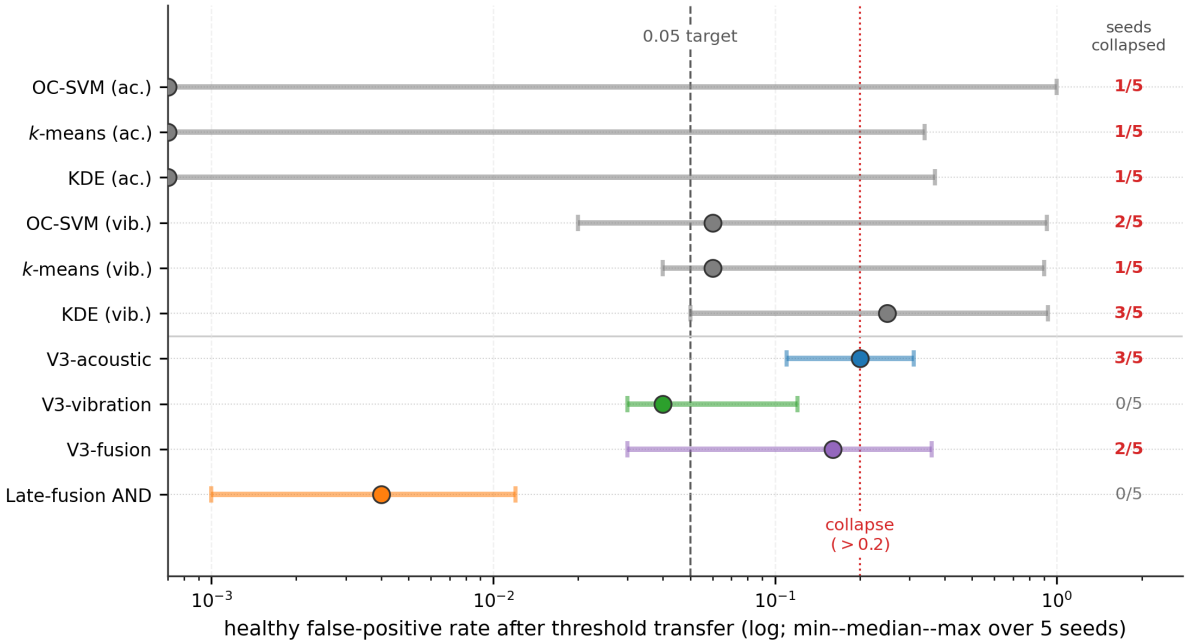


Figure 6.6: Held-out healthy false-positive rate after the per-cluster threshold is transferred to an unseen operating condition (log scale; matched protocol of Table 6.8). Every method calibrates near the 0.05 target in distribution, but under transfer the unconditional scorers and the acoustic and fusion heads collapse on at least one split seed, while the vibration-conditioned flow and the deployed late-fusion AND rule stay low; the AND rule holds at about 0.004.

The old single-number story, in which one scorer collapses and the rest transfer gracefully, does not survive multi-seed scrutiny. Under transfer the acoustic OC-SVM collapses completely on one split seed (FPR $\rightarrow$ 1.0) while sitting near zero on the others, and the supposedly graceful density scorers are no better: the vibration KDE collapses on three of five seeds. No unconditional baseline is reliably domain-robust. The conditioned heads are mixed, the acoustic and fusion flows collapsing on several seeds, but the vibration-conditioned flow never collapses and the parameter-free AND rule never collapses and stays at about 0.004, two orders of magnitude below the worst baseline. This is the steady-state robustness result that motivates deploying the AND rule, and it is distinct from the mode-transition limitation taken up next: conditioning helps keep the threshold calibrated as the operating condition changes, but a mode *transition* still requires the conjunction.

6.4.3 The Mode-Transition Limitation

Under steady operation the conditional flow alerts on 0.072 of held-out healthy windows (Wilson 95% interval $[0.066, 0.078]$ at seed 99, range $[0.044, 0.094]$ across seeds), against a 0.05 target. That is about 44% above target and is better described as over-alerting modestly than as calibrated. On the synthetic mode-crossfade cohorts the picture is worse: the false-positive rate is 1.000 on every seed, so the flow fires on every crossfade window. Part of this is construction, since a linear crossfade of two healthy windows is itself mildly off-manifold. The message still holds. Conditioning sharpens the on-manifold density estimate, but it does not by itself suppress the transient alarms a mode change provokes, which is why the deployed anomaly decision is the late-fusion AND rule. One asymmetry should be stated plainly: because the crossfade is partly off-manifold by construction, the AND rule's silence on it is not by itself evidence of correct behavior, since a detector that was simply insensitive to off-manifold inputs would pass the same test. The crossfade is therefore an upper-bound stress on the flow, and the AND rule's case rests on its healthy-cohort specificity and the structural argument above, not on this figure.

Complementary high-recall operating point. The recall this specificity costs is the rule's real limitation, and it should be read as one operating point rather than the whole detection story. A near-zero false-alarm rate is not by itself a deployment criterion, since a detector that almost never fires inherits one for free. The complementary operating point, a usable recall held at the same 0.05 healthy target, is supplied by an impulse-aware one-class flow developed alongside the chained system and reported in Appendix B. Evaluated under the same five-seed protocol, that head reaches a recording-level ROC-AUC up to about 0.98 with the healthy false-positive rate calibrated to 0.05, and flags between 0.91 and 0.96 of anomaly recordings on the cohorts the array resolves well, an order of magnitude above the AND rule's recall. It is a different architecture from the paradigm comparison and so is kept out of the headline tables, but it shows that the specificity of the conjunction and a usable recall need not be traded against each other once the head preserves the transient structure the contrastive embedding discards.

6.5 RQ3: Source Localization

Three localization heads are trained: a unimodal acoustic head (steered-power volume only), a unimodal vibration head (multilateration init plus learned TDOA tokens), and the

intermediate-fusion head (volume, TDOA tokens, and $\boldsymbol{c}_t$). Three generalization protocols are reported, in increasing order of strictness.

6.5.1 Leave-One-Recording-Out

Table 6.9 reports leave-one-recording-out cross-validation over the localization validation recordings, with the late-fusion weights re-fit per fold on the calibration recordings only. The macro-mean weights every recording equally.

Table 6.9: RQ3 leave-one-recording-out paradigm comparison, macro MAE in meters, median [min, max] over five seeds.

Paradigm	Row	macro MAE
Late fusion	**confidence-gated**	**0.181 [0.160, 0.218]**
Late fusion	uniform avg	0.194 [0.163, 0.215]
Intermediate	V4-fusion	0.189 [0.163, 0.278]
Unimodal	V4-acoustic	0.198 [0.158, 0.286]
Unimodal	V4-vibration	0.231 [0.211, 0.334]
Late fusion	weighted avg	0.252 [0.246, 0.254]

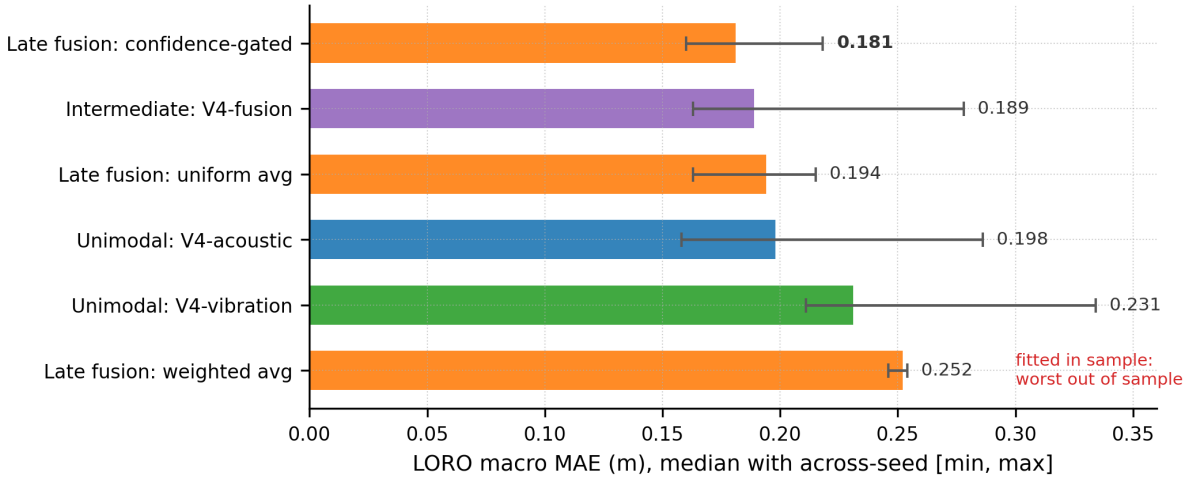


Figure 6.7: The leave-one-recording-out paradigm comparison of Table 6.9. Confidence-gated late fusion has the lowest macro median (0.181 m), but the paradigm bands overlap heavily, so no paradigm is clearly tighter than the others. The in-sample-fitted weighted-average rule is the worst out of sample, the leakage the protocol is built to catch.

Late fusion (confidence-gated) has the lowest macro median, 0.181 m, with a free post-hoc gate that trusts whichever pipeline moved its prediction less far from its own initial estimate. The other learned paradigms follow within the overlap of their seed

ranges, so leave-one-recording-out does not separate them cleanly; the fitted weighted-average rule is the worst, the expected out-of-sample collapse of a rule whose weights are fit in sample. Conditioning the intermediate head on $\boldsymbol{c}_t$ does *not* help here. The FiLM ablation reverses the earlier expectation: the conditioned head reaches a holdout MAE of 0.304 m (median over seeds, $[0.254, 0.332]$) against the unconditional head's seed-invariant 0.247 m, a degradation of 0.057 m ($[0.008, 0.085]$) that is in the same direction on all five seeds ($p < 0.001$). Conditioning injects encoder-seed variance into the localizer without improving it, so $\boldsymbol{c}_t$ earns its place in the anomaly head, not the localization head.

6.5.2 Leave-One-Position-Out and Cross-Session Transfer

Holding out recordings still lets two recordings at the same physical position share a fold. Leave-one-position-out cross-validation over all 16 labeled positions closes that gap (Table 6.10), and it produces a different paradigm winner.

Table 6.10: RQ3 leave-one-position-out cross-validation, by channel mode, over 16 labeled positions (16 folds each). MAE in meters, median [min, max] over five seeds. The ranges are tight, so the position-held-out ordering is seed-robust.

Channel mode	mean MAE
tdoa-only	**0.129 [0.121, 0.132]**
both (fusion)	0.171 [0.167, 0.176]
srp-only (acoustic)	0.184 [0.168, 0.196]
vibration-only (learned)	0.258 [0.241, 0.274]

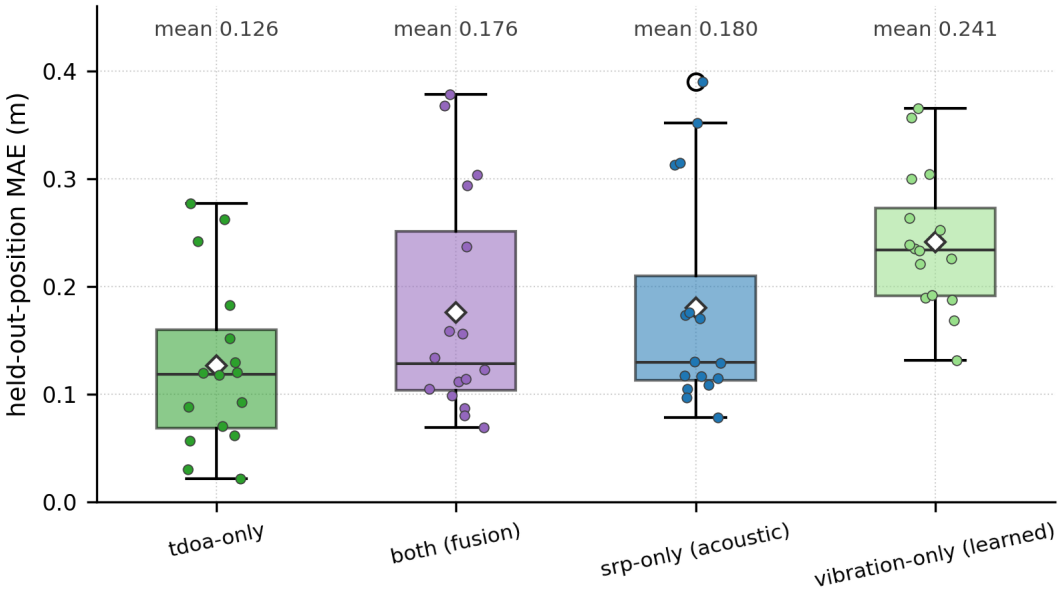


Figure 6.8: Leave-one-position-out error distributions over the 16 folds, by channel mode. The accelerometer-TDOA pathway alone generalizes best to unseen positions; adding the acoustic steered-power pathway does not improve on it, reversing the leave-one-recording-out ordering.

Here the accelerometer-TDOA pathway alone is the most accurate (0.129 m), ahead of the full fusion model (0.171 m) and the acoustic steered-power pathway (0.184 m). Adding the acoustic pathway to the TDOA pathway does not help position-held-out generalization: the volume's soft-argmax initialization overfits the training positions. This reverses the leave-one-recording-out ordering of §6.5.1, and the position-held-out test is the stronger claim. Cross-session transfer confirms the picture (Table 6.11). Trained on the older campaigns and tested on the never-seen D5 session, the TDOA pathway localizes D5 knocks at 0.113 m and the full fusion model at 0.145 m.

Table 6.11: RQ3 cross-session transfer, train on D2/D3/D4 and test on the unseen D5 ($n_{\text{test}} = 63$). MAE and 95th percentile in meters, median [min, max] over five seeds.

Channel mode	val MAE	val p95
tdoa-only	**0.113 [0.107, 0.147]**	0.259 [0.206, 0.390]
both (fusion)	0.145 [0.117, 0.173]	0.282 [0.214, 0.341]
srp-only	0.151 [0.140, 0.176]	0.320 [0.266, 0.347]
vibration-only (learned)	0.291 [0.245, 0.304]	0.441 [0.373, 0.536]

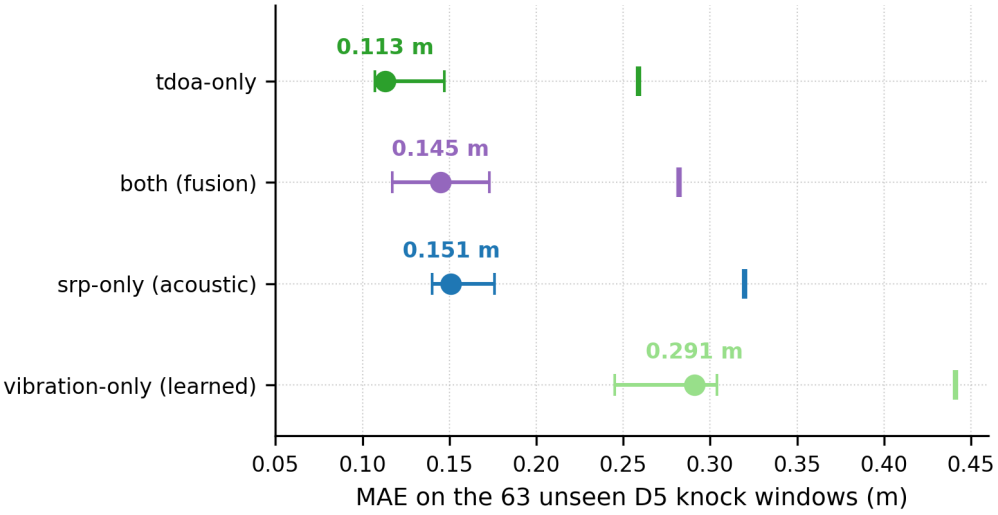


Figure 6.9: Cross-session transfer: trained on D2/D3/D4 and tested on the 63 knock windows of the never-seen D5 session (Table 6.11). The TDOA pathway localizes unseen-session knocks at 0.113 m mean error (median over five seeds), the strongest generalization claim in the thesis.

A per-position breakdown shows the failures are geometric, not signal-quality driven. The worst positions lie on or outside the convex hull of the sensor array, at held-out MAE around 0.25 to 0.32 m; every position inside the hull is localized within roughly the 0.07 to 0.20 m band. The limit is therefore the array envelope: positions the microphones and accelerometers do not bracket cannot be triangulated reliably.

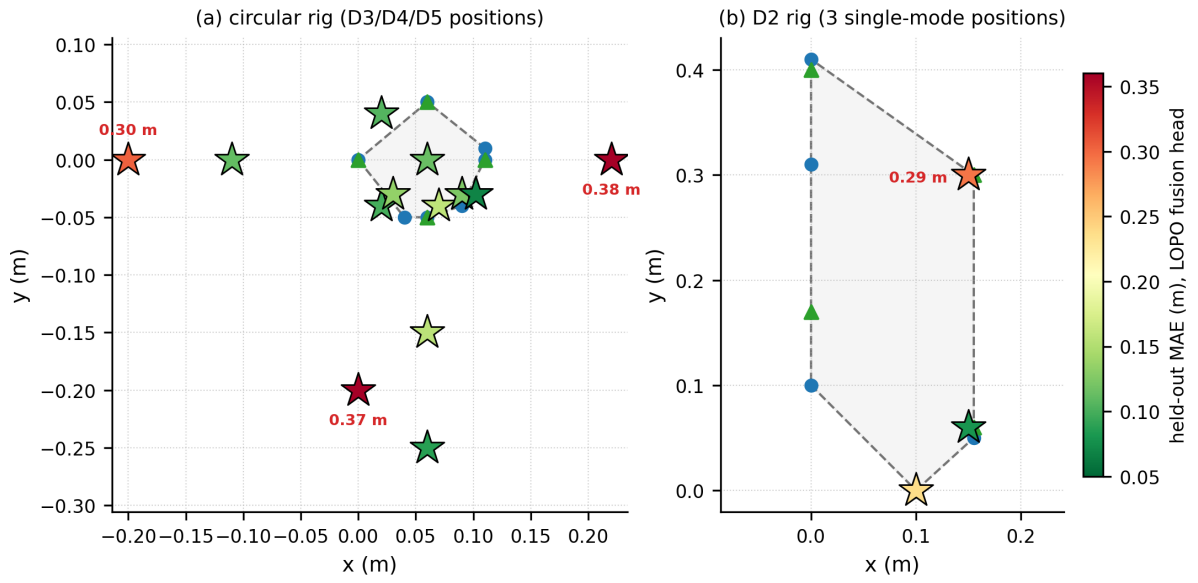


Figure 6.10: Held-out localization error per labeled position (leave-one-position-out, fusion head), overlaid on the sensor arrays of Figure 3.2. Positions inside the dashed sensor convex hull localize well; the failures are the positions on or outside the hull. The limit is the array envelope, not signal quality.

Table 6.12 places the classical baselines against the best learned result under each protocol on one error axis. The classical estimators localize to roughly 0.25 to 0.5 m; the best learned configurations reach 0.11 to 0.19 m, an improvement under every protocol. The protocols differ, as the table states, so each block is read on its own terms rather than as a single controlled comparison across the classical-to-learned boundary.

Table 6.12: RQ3 consolidated localization error: classical baselines against the best learned result under each protocol, on one MAE axis. The protocol column states the split each number is computed under; the classical estimators are per-recording within a campaign, the learned heads are cross-validated. MAE in meters; all learned-head rows are five-seed medians. The full per-paradigm breakdowns are in Tables 6.9, 6.10, and 6.11.

Method	Protocol	MAE (m)
SRP-PHAT (acoustic, classical)	per-recording	0.247–0.422
accel-TDOA (vibration, classical)	per-recording	0.361–0.491
Late fusion (confidence-gated)	LORO macro	0.181
tdoa-only	LOPO	**0.129**
tdoa-only	cross-session → D5	**0.113**

6.6 RQ4: Supervisory Conditioning

RQ4 is exploratory: the prototype corpus carries no temperature, pressure, or SCADA channels (Chapter 5). It is evaluated in two parts, a prototype stand-in and an offline analysis of the real plant archive.

6.6.1 Prototype stand-in

Exercising the supervisory pathway with the recorded fan-noise level (one instance of the generic operating-condition token of Chapter 5) as a stand-in does not improve the localization MAE (0.303 m median over seeds, against the same conditioned head without it). This is the planned wiring check of §5.2.4, and its outcome is a null: the token does not measurably help. The result is consistent with $\boldsymbol{c}_t$ already encoding the operating-context information the token would carry, so adding it explicitly is redundant, though, being a null, it does not establish that on its own. What it does confirm is that the conditioning pathway runs end to end without error.

6.6.2 Real plant SCADA

The real ROW II archive was subsequently delivered and mined offline. No model is trained here; the analysis is a pure information-theoretic ranking that asks which process channels a field deployment should route into the supervisory vector $\boldsymbol{s}_t$. The `ROWII_Allg_M1` stream covers 633 600 samples at 1 Hz (about 176 h) across 30 channels, of which 28 carry signal. It has no temperature channel, so the thermal element of RQ4 does not apply to this archive. An instrumentation audit qualifies it further: of the twelve accelerometer channels carried by the machine's four triaxial accelerometers only 6 were live, and of the eight installed microphones none were live during the campaign, so the legacy anomaly target was derived from a degraded sensor set. Mutual information between each channel and a target is estimated with a nearest-neighbor estimator, and significance is taken from a circular-shift permutation null, which preserves the strong 1 Hz autocorrelation, with a Benjamini-Hochberg false-discovery-rate (FDR) correction across channels, reported as an adjusted q-value.

The primary result is positive, and it is about operating state rather than anomaly. The SCADA channels identify the operating mode strongly. To make the magnitude interpretable, the mode label here has an entropy of $H(\text{mode}) \approx 0.95$ nats (computed from the mode occupancy in Table 6.14), which is the ceiling any mutual information with the mode can reach. The leading channels of Table 6.13 reach 0.70 to 0.96 nats,

that is, essentially the full mode entropy, so the operating mode is almost completely determined by these process channels.[2] This is precisely the role $\boldsymbol{s}_t$ plays, telling the detector what state the machine is in so that a mode change reads as a context shift rather than a fault, and it confirms on real-turbine data that wiring these channels into $\boldsymbol{s}_t$ is well founded.

Table 6.13: RQ4 real-SCADA analysis: mutual information between each ROW II SCADA channel and the operating-mode label, leading channels. The mode-label entropy is $H(\text{mode}) \approx 0.95$ nats, so these values are at or near the ceiling.

Channel	Family	MI with mode (nats)
`1_P_Ist` (active power)	electrical	0.96
`1_21kV Gen. Spg.` (gen. voltage)	electrical	0.93
`1_Drehzahl_Ist` (speed)	rotational	0.93
`1_KS Stellung` (valve)	other	0.91
`1_Erregerstrom` (excitation)	electrical	0.90
`1_Leitapparat Stell.` (guide vane)	hydraulic	0.90
`1_Laufraddruck` (runner pressure)	pressure	0.86
`1_Spiraldruck` (spiral-case pressure)	pressure	0.81
`1_Q_Ist` (flow)	hydraulic	0.70

[2]The nearest-neighbor estimate for the top channel marginally exceeds $H(\text{mode})$, which is an estimator artifact rather than a true information value, since mutual information is bounded by the label entropy; the practical reading is that the channel is at the ceiling. One nat is about 1.44 bits.

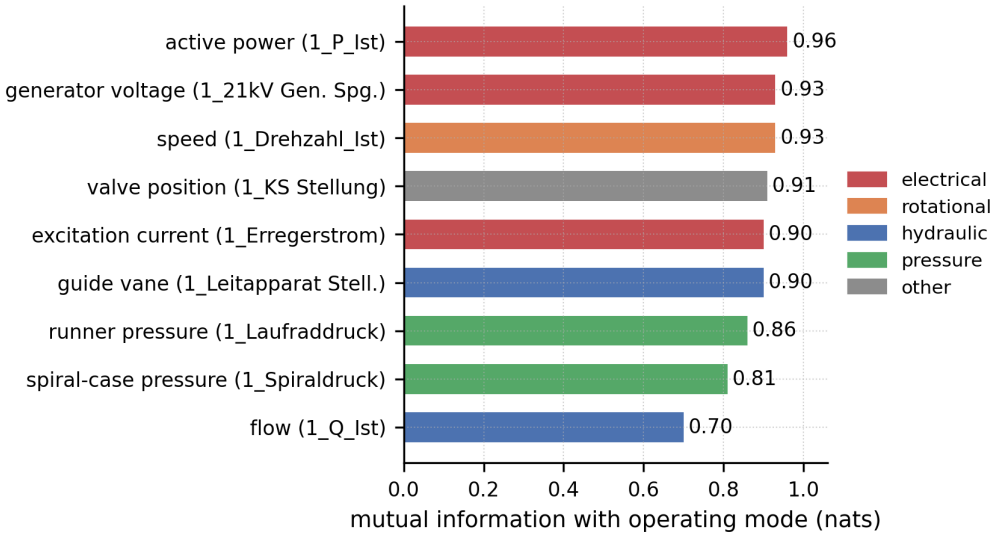


Figure 6.11: Mutual information between the leading ROW II SCADA channels and the operating-mode label (Table 6.13), colored by channel family, against the $H(\text{mode}) \approx 0.95$ nat ceiling. The electrical, rotational and hydraulic process channels identify the mode at 0.70–0.96 nats, the operating-context role the supervisory vector $\boldsymbol{s}_t$ is designed to carry.

The anomaly-association question is the secondary one, and the answer is more guarded. Against the specific anomaly events, no channel is significant after FDR correction under a global, per-second co-occurrence test. Three channels reach uncorrected significance, namely runner pressure `1_Laufraddruck` ($p = 0.005$), flow `1_Q_Ist` ($p = 0.030$), and turbine flow `Durchfluss TU` ($p = 0.045$), but the smallest corrected q is 0.140. Rebuilding the target from the six live accelerometers alone gives the same outcome (0 of 28 channels after FDR, per-channel AUC 0.487 to 0.520), which rules out the degraded target as the explanation. The result is therefore that no association is detected by this global test, not that none exists.

Stratifying by operating mode removes the dominant between-mode variance and changes that picture (Table 6.14). The two largest modes remain null, but pump mode surfaces flow and guide-vane position, and phase-shifter mode surfaces headwater level and pressure, which are the physically correct channels for an anomaly in each regime. These hits are suggestive rather than established, and the caveat must travel with the numbers: FDR is applied within each mode but *not* across the four modes, so the per-mode q-values (for example the phase-shifter $q = 0.007$) are not corrected for the mode

that produced them. The pump-mode dependence is moreover not directional, and the phase-shifter test rests on 169 anomaly seconds, which is small.

Table 6.14: RQ4 real-SCADA analysis: within-mode anomaly-association test. Significance is after per-mode FDR control only; no correction is applied across the four modes, so the per-mode q-values should be read with that in mind. PU is significant under the MI test; the 169-second PH cohort is too small to estimate MI reliably, so only the rank-AUC test is reported there.

Mode	n (s)	anomaly s	significant (leading channels)
ST (steady gen.)	372 960	8 280	0
TU (turbine)	189 006	6 420	0
PU (pump)	61 979	1 029	4: flow, guide vane, pump flow, gen. voltage
PH (phase-shift)	6 957	169	4: headwater level, headwater pressure, pump flow

The deployment recommendation for RQ4 is therefore concrete and stays within what the data supports. The mode-defining channels of Table 6.13 should be routed into $\boldsymbol{s}_t$ for operating context, and, analyzed per operating mode, the pump-mode flow and guide-vane channels and the phase-shifter headwater channels are the leading anomaly-relevant supervisory inputs to validate on instrumented field data. RQ4 remains exploratory: no detector or localizer is trained on these signals in this thesis.

6.7 Robustness and Hyperparameter Sensitivity

Across the five seeds the density objective is essentially seed-invariant (validation NLL -241.05, range $[-241.09, -241.03]$), confirming that the flow fits the healthy manifold the same way every time. What varies is the event-level real-anomaly F1, which couples to the K-means cluster assignment: it has a median of 0.989 but a mean of 0.936 (± 0.103 across the five seeds), pulled down by one low-tail outlier (seed 99, F1 $= 0.756$) from a degenerate cluster initialization. The selected configuration reaches F1 $= 0.993$ at seed 42. The contrast between the stable density objective and the unstable F1 is the robustness story: the model fits reliably, and the metric that swings is the one that inherits the clustering's seed sensitivity. Selection of the regularization cell and residual range was made on a seed-averaged selection metric, not on the most favorable single seed.

In the model-selection regularization sweep (Appendix A.2), conducted under an earlier single-seed protocol and used only to choose the dropout and weight-decay cell rather than to fix a headline number, the real-anomaly F1 stayed in $[0.909, 0.961]$ with

precision at or above 0.994 on every cell, so recall was the discriminator; the five-seed headline F1 reported above is the figure of record.

On the localization side the residual range is the most consequential swept axis. Widening the bounded correction from its ±0.20 m default (Eq. 4.7) to ±0.30 m lowers the holdout MAE in the sweep, but the gain is not uniform: it is concentrated on the few knock positions that lie outside the sensor footprint, which are the worst-localized in the corpus (Figure 6.10). Their true locations sit beyond the reach of a 0.20 m correction, so only the wider bound can push the estimate toward them, whereas every position the array brackets is already well localized at either bound. The ±0.20 m default is the headline value and is used for every localization number reported in this thesis. It is tied to the physical scale of the prototype, on which an induced fault was expected to lie within the instrumented region, and holding it fixed keeps every reported figure on the same residual scale and therefore comparable across campaigns. The handful of positions that benefit from the wider bound sit outside the sensor footprint at 0.20 to 0.30 m and occur across the campaigns, the farthest of them in the early D2 session, so the ±0.30 m result is the headroom left on this axis, an extrapolation onto those out-of-footprint outliers, rather than a competing headline. Target-position augmentation beyond about half a centimeter backfires, eroding the geometric prior the steered-power initialization carries. Restricting the training windows to those the anomaly head gates as alerts, rather than to a weak impulse-envelope heuristic, roughly halves the holdout error on the small gated cohort.

6.8 Summary of Headline Numbers

Table 6.15 consolidates the generalization claims in order of evidentiary strength. The strongest is cross-session transfer: trained without ever seeing the newest recording session, the system localizes its knocks at 0.113 m (TDOA) to 0.145 m (fusion), median over five seeds, on the 63 held-out D5 knock windows. Under leave-one-position-out, folded over all 16 labeled positions, it localizes at 0.171 m with the full model and 0.129 m with the accelerometer-TDOA pathway alone. Both rest on small labeled cohorts, so the medians carry the wide intervals reported in §6.5 and are read as relative orderings rather than field specifications. For mode discovery the label-free encoder reaches a strict NMI of 0.41 (median over seeds). The anomaly head's deployed operating point, the late-fusion AND rule, holds a near-zero healthy alert rate (0.003) and is the only method that stays calibrated under a threshold transfer. The event-level F1 against envelope-derived weak labels (five-seed median 0.989) is reported in §6.7 as a consistency

check for model selection rather than a headline metric, because its targets share the Hilbert-envelope source used as a vibration encoder input (§5.2.2), so it is omitted from the consolidated table below.

Table 6.15: Consolidated headline numbers, by evidentiary strength. All values are five-seed medians.

Claim	Method	Value
Cross-session transfer (D2/D3/D4 → D5), TDOA	MAE	0.113 m
Cross-session transfer (D2/D3/D4 → D5), fusion	MAE	0.145 m
Leave-one-position-out, TDOA	MAE	0.129 m
Leave-one-position-out, fusion	MAE	0.171 m
Mode discovery (fused $\boldsymbol{c}_t$), strict	NMI	0.41
Anomaly detection (AND rule), healthy alert	FPR	0.003

Taken together, the results give a paradigm-specific answer, not a single-model one. For anomaly detection, the parameter-free late-fusion AND rule is the operating point with a near-zero healthy alert rate and the only one robust to a threshold transfer, at the cost of a low and seed-sensitive recall. For localization, late fusion wins on average accuracy under leave-one-recording-out, while under the stricter position-held-out and cross-session tests the accelerometer-TDOA pathway alone is the most accurate, and conditioning the localizer on $\boldsymbol{c}_t$ does not help. For mode discovery, the fused encoder recovers operating mode at a strict NMI of about 0.41 but does not clearly beat the acoustic encoder alone. For the supervisory question, the real plant SCADA identifies the operating mode almost to its entropy ceiling, which supports the context-conditioning pathway, even though no anomaly-specific channel survives a global significance test. Chapter 7 takes up the interpretation of these findings, their limitations, and the conditions under which they would transfer to the full-scale machine.

Chapter 7

Discussion

This chapter interprets the results of Chapter 6 against the central question of this thesis: whether latent representations of operational context can be recovered from acoustic and vibration streams, and whether conditioning anomaly detection and source localization on those representations improves performance under domain shift. Rather than revisit each result in isolation, the chapter builds a single argument. It opens with the verdict that runs through every research question, develops that verdict through the four questions in turn, draws the threads into the general principles they share, and ends with the implications that follow for a field deployment.

One framing point governs everything below and is made once, here. The evidence comes from a 3D-printed, instrumented scale prototype built to represent the ROW II machine, not from the full-scale turbine. This bounds the magnitude of every absolute number and the realism of every injected fault, but it does not weaken the methodological claims, which concern *how* multimodal context should be learned and used and are evaluated under protocols (Chapter 5) designed to remain honest at any scale. The point is not repeated section by section; where the prototype scale changes a specific interpretation, that is flagged in place.

7.1 Performance Trade-Offs Across Operating Points

The study was designed as a head-to-head contest of three fusion paradigms (unimodal, late, and intermediate) across detection and localization. Its most important outcome is that the contest has no single winner; it has a map of operating points (Figure 7.1). Label-free mode discovery rests on the acoustic-dominated context encoder, and fusion adds nothing seed-robust there (§6.3); the near-zero-false-alarm operating point for anomaly detection is won by the parameter-free late-fusion AND rule (§6.4); and for localization, late fusion wins average accuracy while the accelerometer-TDOA pathway alone wins

the strictest generalization (§6.5). A method that leads on one axis rarely leads on another. The contribution is to establish which paradigm to reach for under which demand, not to crown one architecture.

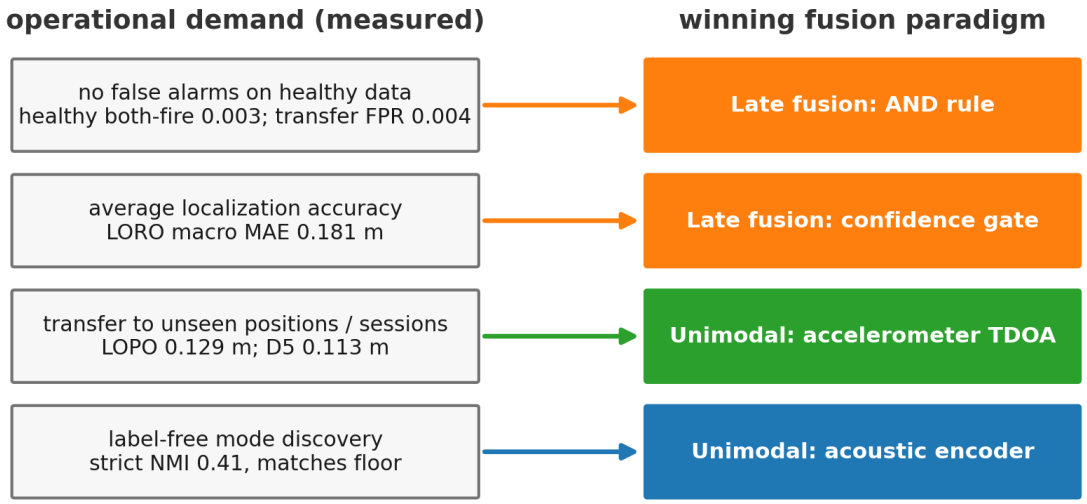


Figure 7.1: The central contribution in one view: no single fusion paradigm wins everywhere. Each measured operational demand selects a different winner: the parameter-free AND rule, the confidence gate, the accelerometer-TDOA pathway, or the acoustic encoder. The verdict of this thesis is demand-specific, not architectural.

One thread runs beneath that map and is the most transferable finding of the thesis: at this data scale, the parameter-free or single-pathway combination repeatedly matches or beats the jointly trained intermediate model. The AND rule holds a near-zero healthy alert rate (about 0.003 on the healthy hold-out) where the joint intermediate flow's recall on the largest labeled cohort is low and seed-dependent; the accelerometer-TDOA pathway localizes held-out positions to about 0.13 m where full fusion reaches only 0.17 m. Fusion is not thereby worthless (it still improves the density model, if modestly), but its value is conditional, and §7.6 returns to why. The central question therefore receives the nuance the evidence demands: operating context can be extracted and is useful, but the deployable gains at this scale follow from disciplined, parameter-free use of the learned components rather than from end-to-end joint training.

7.2 Acoustic-Trunk Context Learning (RQ1)

The first research question asks which fusion strategy best captures cross-modal dependencies for context learning. The answer is positive on the premise and negative on the means. Stated without softening, the learned encoder does not beat the classical reference: at a strict NMI of about 0.41 (Table 6.4) it sits level with a hand-engineered feature baseline at the median and marginally below its unsupervised clustering floor. What it adds is not accuracy but economy, since it reaches that level with no labels and no feature engineering, and the contribution of RQ1 is bounded by exactly that: operating mode is recoverable from raw streams, at hand-engineered quality, for free. On the means the result is negative twice over, because fusion does not improve on the acoustic encoder either. That negative is usable rather than disappointing because its cause is measured, not assumed: zeroing the vibration stream leaves the context vector almost unchanged, so the joint cross-attention pool has discarded vibration along the very directions the clustering uses. Fusion has nothing left to combine.

This reading is narrow and should not be over-read. The finding is not that vibration is uninformative, since RQ2 shows it to be the stronger detection modality on this rig, but that label-free discrimination of steady operating *modes* rests on acoustic structure onto which the representation collapses. The cross-modal alignment objective is therefore retained for its role in keeping the two per-modality representations stable and aligned, not for a quantified lift on this corpus. A firm supervised ceiling cannot be read here, since one labeled campaign has too few recordings for a recording-level split and the other collapses onto its majority class, so the label-free NMI is read on its own terms. This acoustic dominance is the first appearance of a pattern the next two questions sharpen: the two modalities carry different information, and the most effective designs exploit that asymmetry rather than averaging it away.

The verdict is, moreover, a property of the acquisition environment as much as of the modalities. The prototype is recorded in a comparatively quiet acoustic setting; a full-scale turbine hall is not, and its broadband machine noise, reverberation, and neighboring-unit signatures would degrade the acoustic channel while the structure-borne vibration path stays comparatively local and shielded. The ordering seen here (acoustic carrying the mode structure, vibration adding little) may therefore invert at full scale. The fusion machinery is retained precisely so the representation can lean on whichever stream the deployment environment makes more reliable.

7.3 Conditioning and the Conjunction Rule (RQ2)

Where RQ1 found vibration weak for mode discovery, RQ2 finds it the stronger of the two detectors, the first half of the modality asymmetry. The problem it must solve is set by the unconditional baselines (Table 6.3): no prior-work scorer is simultaneously a strong detector and domain-robust. Under a threshold transfer to an unseen condition the one-class SVM collapses on some seeds, its held-out false-positive rate reaching 1.0, and the density and distance scorers are no safer. Closing that gap is the operational form of the central question, and the role the context-conditioned head and the conjunction rule are built for.

Two regimes of domain shift must be kept apart, because the system answers them differently. The first is a *steady-state* change of operating condition, in which the machine is held in a new but stationary regime. Here conditioning helps, but the conjunction carries the result. The conditioned flow is a measurably better density model of the healthy manifold than its unconditional ablation, by a held-out paired margin of $\Delta\mathrm{NLL} = 0.055$ (median over seeds, $[0.016, 0.116]$; at seed 42, 95% CI $[0.053, 0.057]$, $p < 0.001$, $n = 591$ paired windows): the direction is stable across all five seeds, but against log-likelihoods of order -241 the gain is about two hundredths of one percent, statistically real and practically minute. On the protocol that matches the baselines (Table 6.8) conditioning alone is then not enough, since the acoustic and fusion heads still collapse on some seeds. The robust operating point is the parameter-free late-fusion AND rule, which holds a transferred false-positive rate of about 0.004 under that steady-state shift and never collapses across the five seeds, because it requires two independent detectors to agree and a healthy window rarely fools both at once. Its near-zero rate undershoots the 0.05 target by construction, since a conjunction can only lower a false-positive rate, so it is read as a property of the rule rather than as drift.

One qualification travels with this claim. On this corpus the transfer split moves operating mode, recording session, fan-noise level, and vibration-acquisition mode together, so the result establishes robustness to a *joint* change of operating and acquisition condition rather than to an isolated change of operating mode. Disentangling the two would require several healthy recordings per condition, which the prototype does not provide.

The second regime is a *transient* operating-mode transition, and it is where the present evidence stops, for a reason that must be stated before any number. The corpus records no real mode transitions, so the only available probe is a synthetic crossfade between two

healthy windows, which is off-manifold by construction and therefore non-diagnostic of field behavior. That the conditional flow fires on every crossfade window is consistent with the probe being degenerate rather than with the detector being miscalibrated; it confirms that conditioning does not by itself absorb an off-manifold input, and nothing more. Robustness to genuine transitions is thus neither shown nor refuted here: the steady-state conclusion above does not transfer to this case, which remains open and requires a campaign that records real transitions.

That specificity is bought with recall, so a deployment should treat the conjunction as one of two operating points, not the whole detection answer: a detector that almost never fires meets a near-zero false-alarm target trivially. The instructive failure of the alternative is the intermediate flow's low, seed-dependent recall on the largest cohort (median about 0.32, falling to about 0.09 across seeds), where joint training erodes the vibration anomaly signal, which is precisely why the deployed decision conjoins two specialized detectors rather than relying on one jointly trained one. The complementary recall is supplied by an impulse-aware one-class flow developed alongside the chained system (Appendix B), which under the same five-seed protocol reaches a recording-level ROC-AUC up to about 0.98 at the calibrated 0.05 healthy false-positive rate, an order of magnitude above the conjunction's recall. One scope limitation must be conceded squarely here. This flow belongs to a different architecture class than the three fusion paradigms under test, so it was held outside the headline comparison; the practical consequence is that the detector this thesis ultimately recommends is not contained, in full, in the controlled contest that the rest of the chapter rests on. The exclusion is principled rather than convenient (the impulse-aware flow is not a fusion paradigm and could not be slotted into the unimodal/late/intermediate axis without changing what that axis measures) and it is mitigated by evaluating the flow under the identical five-seed, recording-level protocol, so the comparison across the two heads is at least like-for-like even though it is not in-contest. With that limitation stated, the deployable detection answer is a pair: the conjunction as a guaranteed-specificity gate, and the impulse-aware flow for the recall the gate cannot provide.

7.4 Paradigm- and Geometry-Bound Localization (RQ3)

The same data-scale-and-complexity pattern recurs in localization, where it can be watched directly as the test gets stricter. Three generalization protocols of increasing strictness return different winners, and the progression is itself the finding. Late

fusion wins average accuracy under leave-one-recording-out; under the stricter leave-one-position-out test the accelerometer-TDOA pathway alone is most accurate (about 0.13 m against 0.17 m for fusion), a verdict the cross-session transfer to the unseen D5 session confirms.

The reversal is interpretable and carries a general lesson. The acoustic steered-power volume is initialized by a soft-argmax over a grid, which memorizes the training positions; the residual head cannot fully undo that bias on a position it has never seen, so adding the acoustic pathway *harms* position-held-out generalization. The accelerometer time-differences, by contrast, are a geometric invariant (the same hyperboloids regardless of which positions occupied the training fold), so the TDOA pathway transfers. The best pathway is thus a function of the generalization actually demanded: fusion buys stability when the test resembles training, the geometric pathway buys transfer when it does not. One caution bounds the lesson. The acoustic pathway's failure to transfer is tied to its soft-argmax grid initialization, an implementation choice, so the evidence shows that *this* acoustic front-end harms position transfer, not that acoustic localization must in principle. A front-end without the memorized grid prior might transfer better, which makes the claim one about memorized spatial priors rather than about the acoustic modality as such.

Two observations bound the practical accuracy. First, errors are set by the array envelope, not by signal quality: every failure lies on or outside the convex hull of the sensor array, while every position the sensors bracket localizes well, so field accuracy is governed by sensor placement rather than by the learner. Second, a methodological lesson: an inherited steel wave-speed constant, too high by roughly $2.5\times$, had silently suppressed the entire vibration localization channel until it was replaced by the material-correct plastic value of $2000\,\mathrm{m\,s^{-1}}$. The correction was physics, not tuning: in a chained sensing-to-estimation system a single wrong constant can mask a genuine signal.

7.5 Regime Identification from Supervisory Signals (RQ4)

The fourth question is exploratory by construction, since the prototype carries no temperature, pressure, or SCADA channels, but its two available checks validate the premise the earlier questions relied on. The first is a wiring check: exercising the supervisory pathway with the recorded operating-condition token leaves the localization error unchanged. Read correctly, this is the predicted null returning as expected, since the context vector already encodes what the token would carry, so the mechanism runs end to end as redundancy rather than as a performance gain.

The substantive result comes from mining the real ROW II archive offline, and two findings stand in sharp contrast. First, the process channels encode the operating mode almost completely: their mutual information with the mode label rises to the mode-label entropy ceiling of about 0.95 nats (Table 6.13), the leading channel's point estimate sitting fractionally above that bound as a known positive bias of the nearest-neighbor estimator rather than a true excess, so the channels saturate the available mode information. This corroborates a narrower claim than it might first seem. It establishes that operating mode is recoverable from the process channels, which is the warrant for routing them in as the supervisory conditioning input $\boldsymbol{s}_t$. It does not bear on the acoustic-vibration encoder, whose own recoverability result rests on the prototype evidence of §7.2, and it does not show that conditioning improves detection, which §7.3 found to be a modest effect. The archive validates the conditioning *input*, not the encoder and not the size of the conditioning gain. Second, no process channel is a significant anomaly detector. None survives a global, FDR-controlled test, and the per-anomaly information sits roughly two orders of magnitude below the per-mode information. Stratifying by mode surfaces a few physically sensible hits (pump-mode flow and guide-vane position, phase-shifter headwater level and pressure), but only on small cohorts, and an instrumentation audit qualifies even those: of the machine's twelve accelerometer channels only six were live, and none of its eight microphones were, so the legacy anomaly target was derived from a degraded sensor set, and the null speaks as much to field-instrumentation readiness as to the channels themselves.

The contrast settles how the supervisory signal should be used, and it justifies the architecture directly. Because the SCADA channels carry operating-mode information up to saturation but carry no globally significant anomaly signal, they belong on the conditioning path, not the detection path: they should enter the model as the context vector $\boldsymbol{s}_t$ that tells the anomaly head which regime it is in, rather than being thresholded as detectors in their own right. This is exactly the role the chained architecture assigns them, and the archive analysis is the empirical warrant for that choice on real-turbine data. The within-mode anomaly associations are left as a hypothesis to be validated on properly instrumented field data before any channel is trusted as a detector.

7.6 Cross-Cutting Synthesis

The four questions converge on three observations that cut across them, each already visible in the sections above. Stated together, they are the most transferable reading the evidence supports, and each is bounded by the limitations that follow in §7.7.

No single paradigm wins everywhere. The paradigm map is not an artifact of one metric: every operational demand measured here (label-free mode discovery, near-zero-false-alarm detection, average localization accuracy, transfer to unseen positions) selects a different winner. The design question is therefore not *whether* to fuse but *which* combination to reach for under a stated demand, and a system that fixes one architecture for all demands will be wrong for most of them.

Modality complementarity is real but asymmetric. The acoustic stream carries the mode-discriminative structure and detection recall; the vibration stream carries detection specificity and, through its time-differences, the localization transfer. RQ1, RQ2, and RQ3 each saw one side of this asymmetry. The most effective designs in this thesis place each modality where it is strong (conjoining acoustic recall with vibration specificity in the AND rule, and reserving the geometric TDOA pathway for spatial transfer) rather than forcing a symmetric fusion that averages their strengths away. The asymmetry is itself environment-dependent: which stream dominates is a property of the acquisition setting (§7.2), which is why the architecture keeps both pathways available.

Complexity must earn its place. On this corpus, a parameter-free or single-pathway method repeatedly matched or beat the jointly trained model. The likely mechanism is a small-data effect: end-to-end joint training has more parameters to fit, so on a corpus this size it can overfit the dominant modality or the largest cohort before it learns a genuinely shared representation. This is a testable hypothesis, not a result; the thesis does not vary dataset size, so the scaling behavior of fusion against the simpler combinations remains open, and the advantage of joint training may yet grow with data. The firm claim is therefore narrow: at the scale studied here, added complexity did not pay its way, and no scaling advantage is asserted beyond it.

This ordering runs against the grain of the multimodal-fusion literature, and the contrast, rather than the headline number, is the interesting object. Prior work on rotating machinery reports that feature-level fusion exceeds either modality alone, often by wide margins [22, 42], and that fused models stay robust even with very few labeled samples per fault [37]. The present study finds the opposite ordering on its corpus. Two differences

plausibly account for the divergence. First, that literature is almost entirely supervised, multi-class fault classification with balanced labels, whereas the task here is label-free context learning followed by one-class detection under domain shift, a regime that offers far less signal for fitting a shared representation. Second, the comparison here is leakage-controlled at the recording level, which removes within-recording correlations that can flatter a high-capacity fusion model. The small-data reading offered above is, moreover, not uncontested: [37] reports fusion robustness precisely at small sample sizes, which cuts against it, and that unresolved tension is exactly why the scaling question is left open rather than answered.

A separate finding sits beside those three and is too important to leave implicit: it concerns the architecture's own central bet. The system is built around context conditioning, yet conditioning was the smallest of the gains it produced. It made the healthy density model measurably but only modestly better ($\Delta\text{NLL} = 0.055$, median over seeds, $p < 0.001$; §7.3), and the robustness a deployment would actually rely on came from a parameter-free conjunction that learns no combiner at all. The central architectural hypothesis is therefore only weakly supported on this corpus: learning a context representation is useful, but conditioning the detector on it was not where the decisive robustness came from. This does not void the architecture as a contribution, since it remains the experimental vehicle that made the leakage-controlled paradigm comparison possible; it narrows the claim from "conditioning drives robustness" to "a learned context representation is worth building, but how it should enter the decision is itself a design variable." The result is stated plainly rather than buried, because it is the one most likely to redirect a follow-on design.

Two boundaries travel with these observations and are detailed in §7.7: the detection-side reading rests on induced knocks rather than incipient faults, and the modality ordering is a property of this acquisition environment. They bound how far the observations reach without overturning them. Running underneath them is also a methodological stance. Each null is reported with its mechanism attached: the RQ1 fusion null with its measured acoustic dominance, the RQ3 acoustic-harms-transfer result with its soft-argmax explanation, the RQ4 anomaly null with its instrumentation audit. The contribution is not that the results are reported faithfully, which is the baseline expectation of empirical work, but that the mechanism attached to each null converts it into actionable design guidance: a measured cause tells a practitioner what to change, where an unexplained null tells them only to stop.

7.7 Limitations

The primary limitation concerns the anomalies themselves. Every induced fault is a casing impact, a knock, not a naturally occurring incipient fault. This matters more than the scale of the rig: a knock is a broadband impulsive event that excites both channels strongly and simultaneously, whereas the faults the work ultimately targets (vaporous cavitation, propagating micro-cracks, early bearing wear) are narrow-band, low-energy, and modality-selective in their early stages (Chapter 2). A detector that separates knocks from healthy background has therefore been tested on a comparatively easy target, and the detection results must be read as a method and a relative ranking rather than as evidence of sensitivity to real incipient damage. The localization conclusions are less exposed, since a knock is a legitimate impulsive source of known position, but the detection conclusions inherit this in full. Closing the gap requires a campaign that records genuine fault signatures, and it is the first thing a field deployment must establish.

The prototype scale, established at the outset, bounds every absolute number: each headline value is evidence for a method and a relative ordering, not a field specification. The remaining limitations are more specific and follow from the data.

The labeled cohorts are small, which widens every interval; the bootstrap and Wilson intervals and the multi-seed stability checks are reported precisely so this uncertainty stays visible. The corpus records no genuine mode transitions, so the transition stress relies on synthetic crossfades that are partly off-manifold by construction, and the operating-condition shift is approximated by controlled noise levels rather than a campaign driven across regimes. The real-archive analysis cannot fully separate a weak anomaly signal from a degraded instrumentation target, though rebuilding the target from the live sensors alone rules that out as the sole explanation. Finally, faithful re-scoring of the conditional flow requires persisting the learned window-pooling summary alongside the flow weights, which the released pipeline does; the headline tables are five-seed medians, while the per-position map and the score-distribution plot are drawn from a single reference seed.

A final scoping note bounds what the frozen comparison covers. The evaluated conditional flow conditions its affine-coupling layers on the operating context through FiLM over a fixed standard-normal base; refinements to that head were deliberately held out of the evaluation and are taken up as future work in §8.3. The impulse-aware one-class flow that supplies the high-recall operating point is a separate architecture: it has been evaluated over all five seeds and is reported in Appendix B, but it too sits outside

the frozen paradigm comparison, so it informs deployment without entering the headline verdicts.

7.8 Implications

For a near-term deployment on instrumented hardware, the results translate into three recommendations. *Detection* should run two operating points, not one: the parameter-free AND rule as the high-specificity gate, paired with the impulse-aware flow of Appendix B for the recall the gate cannot supply. *Localization* should rely on the accelerometer-TDOA pathway and treat sensor placement as the design variable that sets accuracy, since the array envelope, not the learner, bounds the achievable error. *Supervisory conditioning* should route the mode-defining SCADA channels into $\boldsymbol{s}_t$ and analyze anomalies per operating mode rather than pooled, since that is where the real archive exposed physically meaningful associations.

One conjecture worth testing beyond this setting follows from the pattern, and it is offered as a conjecture rather than a result. In other constrained-data multimodal problems, it may be more productive to separate representation learning from decision fusion, learning each modality well and then combining specialized detectors with discipline, than to assume that deeper end-to-end integration improves robustness by default. Whether that holds outside the present corpus is untested here. What this study establishes within its scope is narrower and firmer: on this prototype, multimodal architectures were better understood as a set of demand-specific operating points than as a search for one universally optimal fusion strategy, and the useful deliverable was less a single best model than a leakage-controlled map of which design serves which demand, reported with the intervals and caveats that let it be reused once the hardware, the faults, and the scale are real.

Chapter 8

Conclusion

This thesis asked whether latent representations of operational context can be recovered from acoustic and vibration streams, and whether conditioning anomaly detection and source localization on those representations improves performance under domain shift. It answered the question by building a single chained system (from self-supervised context encoding, through a context-conditioned anomaly head, to a fusion localization head) and evaluating that system on a five-campaign corpus collected from a 3D-printed scale prototype of the ROW II machine. The four research questions were posed as a contest between the standard fusion paradigms, and the evaluation was held throughout to protocols designed to remain honest at prototype scale: recording-level splits, held-out calibration, bootstrap and Wilson intervals, and baselines drawn from the prior work the system is meant to improve on.

8.1 Summary of Findings

The four research questions returned a consistent and, in places, unexpected picture. Operating context proved recoverable from the sensor streams without labels, though carried mainly by the acoustic channel, and joint fusion did not improve on it. Conditioning the detector on that context helped only modestly; the dependable robustness under steady-state domain shift came instead from a parameter-free conjunction, with a complementary high-recall head supplying what the conjunction trades away. Localization was paradigm- and geometry-bound, the geometric time-difference pathway generalizing most reliably to unseen positions and accuracy set by sensor placement. The supervisory channels of the real archive identified operating mode almost to its entropy ceiling while flagging no anomaly on their own. No paradigm dominated overall: each owned a different operating point, and at this scale the simpler parameter-free combinations were not outperformed

by joint training, whose advantage at larger scale is left open as a hypothesis rather than a finding.

8.2 Contributions

The thesis makes four contributions, of four different kinds.

A scientific finding. The thesis contributes the demand-specific verdict itself (§7.1): the best paradigm is selected by the operational demand, not fixed by architecture, and at this scale the simpler parameter-free combinations are not outperformed by joint training. The accompanying small-data reading is framed as a falsifiable hypothesis rather than a settled result, set explicitly against fusion-literature evidence that points the other way, and the premise that operating context is recoverable is independently corroborated, for the supervisory channels, on the real ROW II archive.

An evaluation methodology. The thesis contributes a leakage-controlled, head-to-head comparison of the unimodal, late, and intermediate fusion paradigms, conducted under recording-level splits, held-out calibration, multi-seed stability reporting, and prior-work baselines, which turns the usual question of *whether* fusion helps into the more useful question of *which* paradigm to use under which demand. Part of the method is a reporting discipline: negative and null results are carried with their mechanisms attached, as first-class evidence rather than as omissions.

A system architecture. The thesis contributes the chained context-conditioning design (a self-supervised operating-context representation that conditions both a normalizing-flow anomaly head, through FiLM, and a fusion localization head), which is the vehicle that makes the paradigm comparison possible and is built to be carried directly onto instrumented hardware.

Deployment guidance. The thesis contributes a concrete, deployable recipe, set out in full in §7.8: a two-operating-point detector (the parameter-free AND gate with the impulse-aware flow for recall), an accelerometer time-difference pathway for spatial transfer with sensor placement as the governing design variable, and the mode-defining supervisory channels routed into the detector's context rather than thresholded as detectors in their own right.

8.3 Outlook

The prototype could establish the relative orderings between paradigms but not the absolute performance of any of them, and the work it leaves to do divides into open scientific questions and the engineering that would answer them.

Scientific open questions. Whether the advantage of jointly trained fusion grows with data, as the small-data reading tentatively suggests and the fusion literature partly disputes, is the question above all others. Beyond it lie two more: whether the dominance of acoustic over vibration inverts at full scale, where broadband hall noise degrades the acoustic channel; and whether context conditioning holds under a genuine, *isolated* change of operating mode rather than the joint operating-and-acquisition shift this corpus could provide. Each is a hypothesis the prototype could raise but not settle.

Engineering work. A larger, fully instrumented campaign on the full-scale machine is what would answer them. It must record genuine fault signatures and real mode transitions rather than synthetic crossfades, bring the full sensor complement live (the microphones and all accelerometer channels the archive audit found dormant), and carry the temperature, pressure, and SCADA channels the fourth research question could only mine offline. Two anomaly-head extensions are ready to fold into it: the three refinements held out of the present evaluation (a context-conditional flow base, an injected impulse-and-spectral anchor, and an empirical-Bayes shrinkage of the per-cluster thresholds), which a single-seed probe suggests help mode discovery but which must be validated under the multi-seed protocol before any gain is claimed; and the impulse-aware one-class flow of Appendix B, already validated over five seeds, to be promoted into the deployed detector.

Such a campaign would convert the relative orderings established here into field specifications, would test the conditioning claim against real domain shift, and would settle whether the acoustic-trunk verdict of the prototype survives the broadband noise of a production hall or gives way to the structure-borne vibration path. The chained, context-conditioned design and the evaluation discipline developed in this thesis are built to be carried directly into it. The contribution of this work is not a finished monitor but the validated map that tells a larger campaign where to look: which paradigm serves which demand, which claims are settled and which remain open, and which design choices a full-scale deployment must revisit before any field-grade guarantee can be made.

Bibliography

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks". In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (2019), pp. 34–48. ISSN: 1941-0484. DOI: 10.1109/JSTSP.2018.2885636.

[2] Muhammad Ahsan, Jose Rodriguez, and Mohamed Abdelrahem. "Bearing Fault Diagnosis in Induction Motors Using Low-Cost Triaxial ADXL355 Accelerometer and a Hybrid CWT-DCNN-LSTM Model". In: *IEEE Access* 13 (2025), pp. 101037–101050. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3577672.

[3] Allen-Bradley. *Applying Condition Monitoring to Various Machinery*. Application note. Allen-Bradley, 2019.

[4] ANDRITZ AG. *Austria - At the heart of Europe*. en. 2026. URL: https://www.andritz.com/hydro-en/hydronews/hn-europe/austria (visited on 03/08/2026).

[5] Jérôme Antoni. "The spectral kurtosis: a useful tool for characterising non-stationary signals". In: *Mechanical Systems and Signal Processing* 20.2 (2006), pp. 282–307. DOI: 10.1016/j.ymssp.2004.09.001.

[6] Blondelle Melina Atsafack, Charles Kabiri, and Gerard Rushingabigwi. "Predictive Maintenance for Hydraulic Turbine Unit: A Comparative Deep Learning Approach Using Internet of Things Data in Real-Time". In: *IEEE Access* 13 (2025), pp. 158340–158352. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3607733.

[7] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language". In: *Proceedings of the 39th International Conference on Machine Learning (ICML)*. 2022, pp. 1298–1312.

[8] Alison Bartle. "Hydropower potential and development activities". In: *Energy Policy* 30.14 (2002), pp. 1231–1239. ISSN: 03014215. DOI: 10.1016/S0301-4215(02)00084-8.

[9] Alessandro Betti, Emanuele Crisostomi, Gianluca Paolinelli, Antonio Piazzi, Fabrizio Ruffini, and Mauro Tucci. "Condition monitoring and predictive maintenance methodologies for hydropower plants equipment". In: *Renewable Energy* 171 (2021), pp. 246–253. ISSN: 0960-1481. DOI: 10.1016/j.renene.2021.02.102.

[10] Michel Blain, Denis Thibault, Ariane Immarigeon, Rafaël Guay, Jérôme Lonchampt, and Pascal Hernu. *Probabilistic technico-economic analysis of hydroelectric power unit operation and maintenance including prognostic*. en. Available via ResearchGate. Varennes (Québec), Aug. 2021. URL: https://www.researchgate.net/publication/355944698_Probabilistic_technico-economic_analysis_of_hydroelectric_power_unit_operation_and_maintenance_including_prognostic (visited on 03/14/2026).

[11] Wahyu Caesarendra and Tegoeh Tjahjowidodo. "A Review of Feature Extraction Methods in Vibration-Based Condition Monitoring and Its Application for Degradation Trend Estimation of Low-Speed Slew Bearing". In: *Machines* 5.4 (2017), p. 21. ISSN: 2075-1702. DOI: 10.3390/machines5040021.

[12] Juan Luis Ferrando Chacon. *Fault Detection in Rotating Machinery Using Acoustic Emission*. en. 2015. DOI: 10.13140/RG.2.1.2369.5444. URL: https://www.researchgate.net/doi/10.13140/RG.2.1.2369.5444 (visited on 03/14/2026).

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 2020, pp. 1597–1607.

[14] Zhang Chen, Yucong Zhang, Xiaoxiao Miao, and Ming Li. *Toward Multimodal Industrial Fault Analysis: A Single-Speed Chain Conveyor Dataset with Audio and Vibration Signals*. arXiv:2603.07130 [cs] version: 1. Mar. 2026. DOI: 10.48550/arXiv.2603.07130. URL: http://arxiv.org/abs/2603.07130 (visited on 03/15/2026).

[15] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. "Convolutional Recurrent Neural Networks for Music Classification". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 2392–2396. DOI: 10.1109/ICASSP.2017.7952585.

[16] Zhenyun Chu, Shuo Xing, Baokun Han, and Jinrui Wang. "A Novel Cross-Domain Mechanical Fault Diagnosis Method Fusing Acoustic and Vibration Signals by Vision Transformer". In: *Sensors (Basel, Switzerland)* 24.16 (2024), p. 5120. ISSN: 1424-8220. DOI: 10.3390/s24165120.

[17] Cora. *Phased Array - Principles and Applications of Microphone Arrays and Speaker Arrays*. en. Blog. Manufacturer blog post (grey literature). July 2025. URL: https://sponcomm.com/info-detail/phased-array (visited on 03/14/2026).

[18] Gonzalo Corral-García, Diana Tejera-Berengué, Manuel Utrilla-Manso, Manuel Rosa-Zurera, Roberto Gil-Pita, and FangFang Zhu-Zhou. "Enhanced acoustic monitoring for industrial predictive maintenance using semi-coprime microphone array and joint spatial-frequency filtering". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2026.1 (2025), p. 1. ISSN: 1687-4722. DOI: 10.1186/s13636-025-00432-3.

[19] DCASE. *Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques - DCASE*. en. 2022. URL: https://dcase.community/challenge2022/task-unsupervised-anomalous-sound-detection-for-machine-condition-monitoring-results (visited on 03/15/2026).

[20] Xin Deng, Hong Hua, Chaoshun Li, Shuman Wei, Zhu Yan, Wanquan Deng, Jiayang Pang, Yufan Xiong, Lihao Li, and Xiaobing Liu. "Deformation and Stress of a Runner in Large Francis Turbines Under Wide-Load Operating Conditions". In: *Water* 17.16 (2025), p. 2374. ISSN: 2073-4441. DOI: 10.3390/w17162374.

[21] David Diaz-Guerra, Antonio Miguel, and Jose R. Beltran. "Robust Sound Source Tracking Using SRP-PHAT and 3D Convolutional Neural Networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 300–311. ISSN: 2329-9304. DOI: 10.1109/TASLP.2020.3040031.

[22] Shaohu Ding, Guangsheng Zhou, Xinyu Wang, and Weibin Li. "Fault Diagnosis of Wind Turbine Rotating Bearing Based on Multi-Mode Signal Enhancement and Fusion". In: *Entropy* 27.9 (2025), p. 951. ISSN: 1099-4300. DOI: 10.3390/e27090951.

[23] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations (ICLR)*. 2017.

[24] Roger F. Dwyer. "Detection of non-Gaussian signals by frequency domain kurtosis estimation". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 8. 1983, pp. 607–610. DOI: 10.1109/ICASSP.1983.1172264.

[25] Electric Power Research Institute. *Acoustic Emissions Monitoring Study: Application of Airborne Acoustic Sensing for Motor Generator Fault Detection in a Laboratory Environment*. June 2022. URL: https://www.epri.com/research/products/000000003002023744 (visited on 03/08/2026).

[26] Xavier Escaler, Eduard Egusquiza, Mohamed Farhat, François Avellan, and Miguel Coussirat. "Detection of cavitation in hydraulic turbines". In: *Mechanical Systems and Signal Processing* 20.4 (2006), pp. 983–1007. ISSN: 08883270. DOI: 10.1016/j.ymssp.2004.08.006.

[27] Yongxin Fan, Yiming Dang, and Yangming Guo. "Fault Identification Model Using Convolutional Neural Networks with Transformer Architecture". In: *Sensors (Basel, Switzerland)* 25.13 (2025), p. 3897. ISSN: 1424-8220. DOI: 10.3390/s25133897.

[28] Hubert Fechner. *National Survery Report of PV Power Applications in Austria 2024*. Tech. rep. IEA PVPS, 2025.

[29] K. Warren Frizell. *Cavitation Detection in Hydraulic Turbines*. Hydraulics Branch report PAP-0685. U.S. Bureau of Reclamation, Hydraulics Branch, Denver, CO, 1995.

[30] Cihun-Siyong Gong, Chih-Hui Simon Su, Kuo-Wei Chao, Cheng-Yen Lin, and Yu-Hua Chen. "Acoustic-Based Fault Detection for Robotic Arms". In: *IEEE Journal of Selected Areas in Sensors* 3 (2026), pp. 1–22. ISSN: 2836-2071. DOI: 10.1109/JSAS.2025.3638975.

[31] Rahul Goyal and Bhupendra K. Gandhi. "Review of hydrodynamics instabilities in Francis turbine during off-design and transient operations". In: *Renewable Energy* 116 (2018), pp. 697–709. ISSN: 09601481. DOI: 10.1016/j.renene.2017.10.012.

[32] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 21271–21284.

[33] Eric Grinstein, Elisa Tengan, Bilgesu Çakmak, Thomas Dietzen, Leonardo Nunes, Toon van Waterschoot, Mike Brookes, and Patrick A. Naylor. "Steered Response Power for Sound Source Localization: a tutorial review". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2024.1 (2024), p. 59. ISSN: 1687-4722. DOI: 10.1186/s13636-024-00377-z.

[34] Pierre-Amaury Grumiaux, Srđan Kitić, Laurent Girin, and Alexandre Guérin. "A survey of sound source localization with deep learning methods". In: *The Journal of the Acoustical Society of America* 152.1 (2022), pp. 107–151. ISSN: 0001-4966. DOI: 10.1121/10.0011809.

[35] Jun Gu, Yuxing Peng, Hao Lu, Xiangdong Chang, and Guoan Chen. "A novel fault diagnosis method of rotating machinery via VMD, CWT and improved CNN". In: *Measurement* 200 (2022), p. 111635. ISSN: 02632241. DOI: 10.1016/j.measurement.2022.111635.

[36] Siwei Guan, Zhiwei He, Shenhui Ma, and Mingyu Gao. "Conditional normalizing flow for multivariate time series anomaly detection". In: *ISA Transactions* 143 (2023), pp. 231–243. ISSN: 00190578. DOI: 10.1016/j.isatra.2023.09.002.

[37] Junchao Guo, Qingbo He, Dong Zhen, Fengshou Gu, and Andrew D. Ball. "Multi-sensor data fusion for rotating machinery fault detection using improved cyclic spectral covariance matrix and motor current signal analysis". In: *Reliability Engineering & System Safety* 230 (2023), p. 108969. ISSN: 09518320. DOI: 10.1016/j.ress.2022.108969.

[38] Tao Guo, Jinming Zhang, and Zhumei Luo. "Analysis of Channel Vortex and Cavitation Performance of the Francis Turbine under Partial Flow Conditions". In: *Processes* 9.8 (2021), p. 1385. ISSN: 2227-9717. DOI: 10.3390/pr9081385.

[39] Fatemeh Hajimohammadali, Emanuele Crisostomi, Mauro Tucci, and Nunzia Fontana. "Evaluating Deep Learning Networks Versus Hybrid Network for Smart Monitoring of Hydropower Plants". In: *Energies* 17.22 (2024), p. 5670. ISSN: 1996-1073. DOI: 10.3390/en17225670.

[40] Tibor Haller and Jonas Länzlinger. *Algorithmic Framework for Sound Localization in Polyphonic Noisy Environments*. Integrative Masters Project (IMP). University of St. Gallen (HSG), Jan. 2025. (Visited on 03/13/2026).

[41] Changheon Han, Yuseop Sim, Hoin Jung, Jiho Lee, Hojun Lee, Yun Seok Kang, Sucheol Woo, Garam Kim, Hyung Wook Park, and Martin Byung-Guk Jun. *IMPACT: Industrial Machine Perception via Acoustic Cognitive Transformer*. arXiv:2507.06481 [cs] version: 1. July 2025. DOI: 10.48550/arXiv.2507.06481. URL: http://arxiv.org/abs/2507.06481 (visited on 03/14/2026).

[42] Jian Hao, Yaqiong Lv, Jialun Liu, and Yu-Chen Liu. "Dynamic weighted multimodal fusion for fault diagnosis of marine rotating machinery under noisy and low-sample conditions". In: *Ocean Engineering* 339 (2025), p. 122082. ISSN: 00298018. DOI: 10.1016/j.oceaneng.2025.122082.

[43] Changjun He, Hua Geng, Kaushik Rajashekara, and Ambrish Chandra. "Analysis and Control of Frequency Stability in Low-Inertia Power Systems: A Review". In: *IEEE/CAA Journal of Automatica Sinica* 11.12 (2024), pp. 2363–2383. ISSN: 2329-9274. DOI: 10.1109/JAS.2024.125013.

[44] Dan Hendrycks and Kevin Gimpel. *Gaussian Error Linear Units (GELUs)*. arXiv:1606.08415 [cs]. 2016. DOI: 10.48550/arXiv.1606.08415. URL: https://arxiv.org/abs/1606.08415.

[45] Fucai Hu, Xiaohui Song, Ruhan He, and Yongsheng Yu. "Sound source localization based on residual network and channel attention module". In: *Scientific Reports* 13.1 (2023), p. 5443. ISSN: 2045-2322. DOI: 10.1038/s41598-023-32657-7.

[46] Lei Hu, Lingjie Tan, Xiangyan Meng, Jiyu Zeng, Peng Luo, and Yi Yang. "A Framework for Anomaly Detection and Evaluation of Rotating Machinery Based on Data-Accumulation-Aware Generative Adversarial Networks and Similarity Estimation". In: *Machines* 14.1 (2026), p. 61. ISSN: 2075-1702. DOI: 10.3390/machines14010061.

[47] Valentin Huber and Stefan Krummenacher. *A Case Study of Acoustic-based Anomaly Detection for Industrial Predictive Maintenance*. Integrative Masters Project (IMP). University of St. Gallen (HSG), Jan. 2025.

[48] Hydro Review Content Directors. *Voith Hydro wins equipment contract for Rodund II hydro plant rehab*. en-US. July 2010. URL: https://www.renewableenergyworld.com/hydro-power/dams-civil-structures/voith-hydro-wins-equipment/ (visited on 03/08/2026).

[49] Illwerke vkw. *Rodundwerk II*. de-AT. 2026. URL: https://www.illwerkevkw.at/rodundwerk-ii (visited on 03/08/2026).

[50] International Energy Agency. *Renewable electricity – Renewables 2025 – Analysis*. en-GB. 2025. URL: https://www.iea.org/reports/renewables-2025/renewable-electricity (visited on 03/08/2026).

[51] International Hydropower Association. *Global hydropower generation rebounds in 2024 and pumped storage development surges Flagship 2025 World Hydropower Outlook Out Now*. en. June 2025. URL: https://www.hydropower.org/news/flagship-2025-world-hydropower-outlook-out-now (visited on 03/08/2026).

[52] International Hydropower Association. *Hydropower in Europe | IHA Regional ProfileHydropower in Europe*. en. 2025. URL: https://www.hydropower.org/region-profiles/europe (visited on 03/08/2026).

[53] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. 2015, pp. 448–456.

[54] Giovanni Jacovitti and Gaetano Scarano. "Discrete time techniques for time delay estimation". In: *IEEE Transactions on Signal Processing* 41.2 (1993), pp. 525–533. DOI: 10.1109/78.193195.

[55] Reza Jalayer, Masoud Jalayer, and Amirali Baniasadi. "A Review on Sound Source Localization in Robotics: Focusing on Deep Learning Methods". In: *Applied Sciences* 15.17 (2025), p. 9354. ISSN: 2076-3417. DOI: 10.3390/app15179354.

[56] Gabriel Jekateryńczuk and Zbigniew Piotrowski. "A Survey of Sound Source Localization and Detection Methods and Their Applications". In: *Sensors* 24.1 (2023), p. 68. ISSN: 1424-8220. DOI: 10.3390/s24010068.

[57] Minyuan Jiang, Min Luo, Chaoyong Zhang, Min Shu, and Guohao Sun. "Rolling bearing fault diagnosis based on acoustic-vibration data fusion and mode decomposition combined with the crested porcupine optimization algorithm". In: *Heliyon* 10.22 (2024), e40351. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2024.e40351.

[58] D. N. Joanes and C. A. Gill. "Comparing measures of sample skewness and kurtosis". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 183–189. DOI: 10.1111/1467-9884.00122.

[59] Gbanaibolou Jombo and Yu Zhang. "Acoustic-Based Machine Condition Monitoring—Methods and Challenges". In: *Eng* 4.1 (2023), pp. 47–79. ISSN: 2673-4117. DOI: 10.3390/eng4010004.

[60] Innocent Kafodya, Samuel Ndovie, Adetayo Onososen, Innocent Musonda, Jabulani Matsimbe, Francis Masi, Juliana Masi, Jolly Chibwana, German Nkhonjera, and Innocent Kafodya. *Deep Time Series in Structural Health Monitoring of Civil Structures: A Review of Architectures, Applications and Challenges*. en. June 2025. DOI: 10.20944/preprints202505.2422.v2. URL: https://www.preprints.org/manuscript/202505.2422 (visited on 03/15/2026).

[61] Aaditya Karna, Sailesh Chitrakar, Hari Prasad Neopane, and Ole Gunnar Dahlhaug. "A review of condition monitoring in Francis turbines for predictive maintenance". In: *Journal of Vibration and Control* (2025), p. 10775463251315816. ISSN: 1077-5463, 1741-2986. DOI: 10.1177/10775463251315816.

[62] Antanas Kascenas, Pedro Sanchez, Patrick Schrempf, Chaoyang Wang, William Clackett, Shadia S. Mikhael, Jeremy P. Voisey, Keith Goatman, Alexander Weir, Nicolas Pugeault, Sotirios A. Tsaftaris, and Alison Q. O'Neil. "The role of noise in denoising models for anomaly detection in medical images". In: *Medical Image Analysis* 90 (2023), p. 102963. ISSN: 13618415. DOI: 10.1016/j.media.2023.102963.

[63] Yohei Kawaguchi, Keisuke Imoto, Yuma Koizumi, Noboru Harada, Daisuke Niizumi, Kota Dohi, Ryo Tanabe, Harsh Purohit, and Takashi Endo. *Description and Discussion on DCASE 2021 Challenge Task 2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring under Domain Shifted Conditions*. Version Number: 2. 2021. DOI: 10.48550/ARXIV.2106.04492. URL: https://arxiv.org/abs/2106.04492 (visited on 03/15/2026).

[64] Najibullah Khadem, Asmatullah Nashir, and Shamsullah Rahmatyar. "The Role of Deep Learning in Advancing Computer Vision Applications: A Comprehensive Systematic Review". In: *Journal of Advanced Computer Knowledge and Algorithms* 3.1 (2025), pp. 1–8. ISSN: 3031-8955. DOI: 10.29103/jacka.v3i1.24732.

[65] Karim Khamaisi, Nicolas Keller, Stefan Krummenacher, Valentin Huber, Bernhard Fässler, and Bruno Rodrigues. *From Noise to Knowledge: A Comparative Study of Acoustic Anomaly Detection Models in Pumped-storage Hydropower Plants*. arXiv:2509.22881 [cs]. Sept. 2025. DOI: 10.48550/arXiv.2509.22881. URL: http://arxiv.org/abs/2509.22881 (visited on 03/08/2026).

[66] Seon-Gyu Kim, Taeyong Park, Ryuna Kang, and Yunsang Kwak. "Domain-collaborative multimodal transformer for fault diagnosis of rotating machines under noisy environments". In: *Advanced Engineering Informatics* 68 (2025), p. 103763. ISSN: 14740346. DOI: 10.1016/j.aei.2025.103763.

[67] Charles Knapp and Glifford Carter. "The generalized correlation method for estimation of time delay". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.4 (1976), pp. 320–327. DOI: 10.1109/TASSP.1976.1162830.

[68] Omer Kullu and Eyup Cinar. "A Deep-Learning-Based Multi-Modal Sensor Fusion Approach for Detection of Equipment Faults". In: *Machines* 10.11 (2022), p. 1105. ISSN: 2075-1702. DOI: 10.3390/machines10111105.

[69] David Latil, Raymond Houé Ngouna, Kamal Medjaher, and Stéphane Lhuisset. "Vibration-based Data-driven Fault Diagnosis of Rotating Machines Operating Under Varying Working Conditions: A Review and Bibliometric Analysis". In: *International Journal of Prognostics and Health Management* 16.2 (2025). ISSN: 2153-2648. DOI: 10.36001/ijphm.2025.v16i2.4208.

[70] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. "Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019.

[71] Jie Li, Yu Bao, WenXin Liu, PengXiang Ji, LeKang Wang, and Zhongbing Wang. "Twins transformer: Cross-attention based two-branch transformer network for rotating bearing fault diagnosis". In: *Measurement* 223 (2023), p. 113687. ISSN: 02632241. DOI: 10.1016/j.measurement.2023.113687.

[72] Ruixue Li, Guohai Zhang, Yi Niu, Kai Rong, Wei Liu, and Haoxuan Hong. "A Multi-Channel Multi-Scale Spatiotemporal Convolutional Cross-Attention Fusion Network for Bearing Fault Diagnosis". In: *Sensors* 25.18 (2025), p. 5923. ISSN: 1424-8220. DOI: 10.3390/s25185923.

[73] Xu Lin, Ruichun Dong, and Zhichao Lv. "Deep Learning-Based Classification of Raw Hydroacoustic Signal: A Review". In: *Journal of Marine Science and Engineering* 11.1 (2023), p. 3. ISSN: 2077-1312. DOI: 10.3390/jmse11010003.

[74] Dong Liu, Lijun Kong, Bing Yao, Tangming Huang, Xiaoqin Deng, and Zhihuai Xiao. "Hydroelectric Unit Vibration Signal Feature Extraction Based on IMF Energy Moment and SDAE". In: *Water* 16.14 (2024), p. 1956. ISSN: 2073-4441. DOI: 10.3390/w16141956.

[75] Stuart P. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[76] Alberto Luna-Ramírez, Alfonso Campos-Amezcua, Oscar Dorantes-Gómez, Zdzislaw Mazur-Czerwiec, and Rodolfo Muñoz-Quezada. "Failure analysis of runner blades in a Francis hydraulic turbine — Case study". In: *Engineering Failure Analysis* 59 (2016), pp. 314–325. ISSN: 1350-6307. DOI: 10.1016/j.engfailanal.2015.10.020.

[77] Huiying Ma, Tao Shang, Gufeng Li, and Zhaokun Li. "Sound Source Localization Method Based on Time Reversal Operator Decomposition in Reverberant Environments". In: *Electronics* 13.9 (2024), p. 1782. ISSN: 2079-9292. DOI: 10.3390/electronics13091782.

[78] Youwan Mahé, Elise Bannier, Stéphanie Leplaideur, Elisa Fromont, and Francesca Galassi. *Unsupervised Deep Generative Models for Anomaly Detection in Neuroimaging: A Systematic Scoping Review*. arXiv:2510.14462 [cs] version: 1. Oct. 2025. DOI: 10.48550/arXiv.2510.14462. URL: http://arxiv.org/abs/2510.14462 (visited on 03/15/2026).

[79] Sergio Martín del Campo Barraza. *Unsupervised feature learning applied to condition monitoring*. 2017. ISBN: 978-91-7583-896-0.

[80] Iñigo Martinez. "Diffeomorphic Transformations for Time Series Analysis: An Efficient Approach to Nonlinear Warping". PhD thesis. 2003. DOI: 10.15581/10171/67243.

[81] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and Music Signal Analysis in Python". In: *Proceedings of the 14th Python in Science Conference (SciPy)*. 2015, pp. 18–25. DOI: 10.25080/Majora-7b98e3ed-003.

[82] Ministry of Energy and Infrastructure in UAE. *Sub-Synchronous Shaft Vibration Analysis Using Multi-Channel Methods in Rotating Machinery*. en-US. Mar. 2025. URL: https://roticsymposium.com/blogpost/sub-synchronous-shaft-vibration-analysis-using-multi-channel-methods-in-rotating-machinery/ (visited on 03/14/2026).

[83] Chaoquan Mo, Ke Huang, and Houxin Ji. "A lightweight and precision dual track 1D and 2D feature fusion convolutional network for machinery equipment fault diagnosis". In: *Scientific Reports* 14 (2024), p. 31666. ISSN: 2045-2322. DOI: 10.1038/s41598-024-81118-2.

[84] Koketso Mohale and Michelle Lochner. "Enabling unsupervised discovery in astronomical images through self-supervised representations". In: *Monthly Notices of the Royal Astronomical Society* 530.1 (2024), pp. 1274–1295. ISSN: 0035-8711. DOI: 10.1093/mnras/stae926.

[85] Ayantha Senanayaka Mudiyanselage, Qing Lee, Nayeon Lee, Sungkwang Mun, Amin Amirlatifi, Joseph Jabour, Thomas Arnold, and Maria Seale. "Frequency domain tensor-based 1D-convolutional neural network and multilinear principal component analysis for machinery fault detection". In: *Annual Conference of the PHM Society* 16.1 (2024). ISSN: 2325-0178. DOI: 10.36001/phmconf.2024.v16i1.3871.

[86] Christophe NICOLET. "HYDROACOUSTIC MODELLING AND NUMERICAL SIMULATION OF UNSTEADY OPERATION OF HYDROELECTRIC SYSTEMS". Doctorat. EPFL, 2007.

[87] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei Kawaguchi. *Description and Discussion on DCASE 2025 Challenge Task 2: First-shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring*. arXiv:2506.10097 [cs] version: 2. Sept. 2025. DOI: 10.48550/arXiv.2506.10097. URL: http://arxiv.org/abs/2506.10097 (visited on 03/15/2026).

[88] Ahmed Ramadhan Al-Obaidi and Anas Alwatban. "Analysis of hydraulic mechanism of dynamics flow visualization in an axial pump with impeller blades based on novel transient characteristics conditions and vibration techniques". In: *Scientific Reports* 16 (2026), p. 6416. ISSN: 2045-2322. DOI: 10.1038/s41598-026-36822-6.

[89] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. "Attentive Statistics Pooling for Deep Speaker Embedding". In: *Proc. Interspeech 2018*. 2018, pp. 2252–2256. DOI: 10.21437/Interspeech.2018-993.

[90] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*. arXiv:1807.03748 [cs]. 2018. DOI: 10.48550/arXiv.1807.03748. URL: https://arxiv.org/abs/1807.03748.

[91] Ana I. Oviedo, John F. Vargas, Roberto C. Hincapie, Andres F. Molina, Edimerk A. Vergara, and Diana M. Tello. "Unsupervised Real-Time Anomaly Detection in Hydropower Systems via Time Series Clustering and Autoencoders". In: *Technologies* 13.11 (2025), p. 534. ISSN: 2227-7080. DOI: 10.3390/technologies13110534.

[92] Oxmaint. *The Economic Impact of Predictive Maintenance on Global Industries*. en. Sept. 2025. URL: https://www.oxmaint.com/blog/post/economic-impact-predictive-maintenance (visited on 03/14/2026).

[93] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition". In: *Proc. Interspeech 2019*. 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680.

[94] Haiwoong Park and Hyeryung Jang. "Enhancing Time Series Anomaly Detection: A Knowledge Distillation Approach with Image Transformation". In: *Sensors (Basel, Switzerland)* 24.24 (2024), p. 8169. ISSN: 1424-8220. DOI: 10.3390/s24248169.

[95] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. *FiLM: Visual Reasoning with a General Conditioning Layer*. arXiv:1709.07871 [cs]. Dec. 2017. DOI: 10.48550/arXiv.1709.07871. URL: http://arxiv.org/abs/1709.07871 (visited on 03/15/2026).

[96] Adrian Alan Pol, Victor Berger, Cecile Germain, Gianluca Cerminara, and Maurizio Pierini. "Anomaly Detection with Conditional Variational Autoencoders". In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 2019, pp. 1651–1657. ISBN: 978-1-7281-4550-1. DOI: 10.1109/ICMLA.2019.00270.

[97] Alexandre Presas, Yongyao Luo, Zhengwei Wang, and Bao Guo. "Fatigue life estimation of Francis turbines based on experimental strain measurements: Review of the actual data and future trends". In: *Renewable and Sustainable Energy Reviews* 102 (2019), pp. 96–110. ISSN: 13640321. DOI: 10.1016/j.rser.2018.12.001.

[98] Patricio A. Ravetta, Jorge Muract, and R. Burdisso. "Feasibility Study Of Microphone Phased Array Based Machinery Health Monitoring". In: 2007.

[99] Rafel Roig, Jian Chen, Oscar de la Torre, and Xavier Escaler. "Understanding the Influence of Wake Cavitation on the Dynamic Response of Hydraulic Profiles under Lock-In Conditions". In: *Energies* 14.19 (2021), p. 6033. ISSN: 1996-1073. DOI: 10.3390/en14196033.

[100] Rafel Roig, Xavier Sánchez-Botello, Oscar de la Torre, Xavier Ayneto, Carl-Maikel Högström, Berhanu Mulu, and Xavier Escaler. "Fatigue damage analysis of a Kaplan turbine model operating at off-design and transient conditions". In: *Structural Health Monitoring* 23.3 (2024), pp. 1687–1700. ISSN: 1475-9217. DOI: 10.1177/14759217231191417.

[101] Marcelo Romanssini, Paulo César C. De Aguirre, Lucas Compassi-Severo, and Alessandro G. Girardi. "A Review on Vibration Monitoring Techniques for Predictive Maintenance of Rotating Machinery". In: *Eng* 4.3 (2023), pp. 1797–1817. ISSN: 2673-4117. DOI: 10.3390/eng4030102.

[102] Ellen Rushe and Brian Mac Namee. "Deep Contextual Novelty Detection with Context Prediction". In: *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*. 2022, pp. 127–138.

[103] Muhammad Ismail Saleem, Sajeeb Saha, Tushar Kanti Roy, and Subarto Kumar Ghosh. "Assessment and management of frequency stability in low inertia renewable energy rich power grids". In: *IET Generation, Transmission & Distribution* 18.7 (2024), pp. 1372–1390. ISSN: 1751-8695. DOI: 10.1049/gtd2.13129.

[104] Elisa Sanchez and Axel Busboom. *Cavitation Monitoring in Rotating Hydraulic Machines Using Machine Learning - A Review*. en. Mar. 2026. DOI: 10.20944/preprints202603.1027.v1. URL: https://www.preprints.org/manuscript/202603.1027 (visited on 03/15/2026).

[105] Prajwal Sapkota, Subarna Paudel, Ravi Poudel, Sailesh Chitrakar, Hari Prasad Neopane, and Bhola Thapa. "Experimental investigation of vibrational signal in a fault induced Francis runner". In: *Frontiers in Mechanical Engineering* 11 (2025), p. 1536603. ISSN: 2297-3079. DOI: 10.3389/fmech.2025.1536603.

[106] Maximilian Schmidt and Marko Simic. *Normalizing flows for novelty detection in industrial time series data*. arXiv:1906.06904 [cs]. June 2019. DOI: 10.48550/arXiv.1906.06904. URL: http://arxiv.org/abs/1906.06904 (visited on 03/15/2026).

[107] Otto H. Schmitt. "A thermionic trigger". In: *Journal of Scientific Instruments* 15.1 (1938), pp. 24–26. DOI: 10.1088/0950-7671/15/1/305.

[108] Senseven Marketing Team. *Vibration Analysis and Acoustic Emission Testing: Benefits and limitations for predictive maintenance*. en-gb. June 2023. URL: https://blog.senseven.ai/vibration-analysis-and-acoustic-emission-testing-benefits-and-limitations-for-predictive-maintenance (visited on 03/14/2026).

[109] Rajesh Shah, Vikram Mittal, and Michael Lotwin. "Recent Advances in Vibration Analysis for Predictive Maintenance of Modern Automotive Powertrains". In: *Vibration* 8.4 (2025), p. 68. ISSN: 2571-631X. DOI: 10.3390/vibration8040068.

[110] Vignesh V. Shanbhag, Surya T. Kandukuri, Grunde Olimstad, and Rune Schlanbusch. "Predictive Maintenance of Critical Components in Hydroelectric Turbines: A Review". In: *IEEE Sensors Journal* 25.17 (2025), pp. 31959–31979. ISSN: 1558-1748. DOI: 10.1109/JSEN.2025.3577692.

[111] Qiaorui Si, Ding Tian, Zhen Zhang, Yu Lu, Zhanxiong Lu, and Peng Wang. "Investigation of flow characteristics and acoustic responses in hydraulic machinery based on a nonlinear cavitation model and vortex-sound coupling method under cavitating conditions". In: *Engineering Applications of Computational Fluid Mechanics* 19.1 (2025), p. 2547991. ISSN: 1994-2060. DOI: 10.1080/19942060.2025.2547991.

[112] Abderrahmane Smahi and Salim Makhloufi. "The Power Grid Inertia With High Renewable Energy Sources Integration: A Comprehensive Review". In: *Journal of Engineering* 2025.1 (2025), p. 7975311. ISSN: 2314-4912. DOI: 10.1155/je/7975311.

[113] Harald Strahberger and Florian Sesztak. *Renewable Energy 2025 - Austria | Global Practice Guides | Chambers and Partners*. Sept. 2025. URL: https://practiceguides.chambers.com/practice-guides/renewable-energy-2025/austria (visited on 03/12/2026).

[114] Wentao Su, Weijun Meng, Yue Zhao, Jian Wu, and Yuekun Sun. "Research on an acoustic judgment method for the incipient cavitation of model hydraulic machinery runner blades". In: *Matéria (Rio de Janeiro)* 28.3 (2023), e20230154. ISSN: 1517-7076. DOI: 10.1590/1517-7076-rmat-2023-0154.

[115] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. "Integral Human Pose Regression". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 529–545. DOI: 10.1007/978-3-030-01231-1_33.

[116] Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, Ngan Le, and Khoa Luu. "Fast Flow Reconstruction via Robust Invertible n × n Convolution". In: *Future Internet* 13.7 (2021), p. 179. ISSN: 1999-5903. DOI: 10.3390/fi13070179.

[117] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis". In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*. 2019, pp. 253–257.

[118] D Valentín, A Presas, M Egusquiza, E Egusquiza, and Jl Drommi. "Innovative Approaches to Hydraulic Turbine Advanced Condition Monitoring". In: *IOP Conference Series: Earth and Environmental Science* 1411.1 (2024), p. 012019. ISSN: 1755-1307, 1755-1315. DOI: 10.1088/1755-1315/1411/1/012019.

[119] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017, pp. 5998–6008.

[120] Victor Velasquez and Wilfredo Flores. "Machine Learning Approach for Predictive Maintenance in Hydroelectric Power Plants". In: *2022 IEEE Biennial Congress of Argentina (ARGENCON)*. 2022, pp. 1–6. DOI: 10.1109/ARGENCON55245.2022.9939782.

[121] Santhosh Krishnan Venkata and Swetha Rao. "Fault Detection of a Flow Control Valve Using Vibration Analysis and Support Vector Machine". In: *Electronics* 8.10 (2019), p. 1062. ISSN: 2079-9292. DOI: 10.3390/electronics8101062.

[122] Maheshwari Vigneswar and Karthikeyan Paramasivam. *Representing Audio Data: An In-Depth Look at STFT and MFCC*. en-US. Blog. 2026. URL: https://www.ideas2it.com/blogs/mfcc-stft-from-audio-data (visited on 03/14/2026).

[123] Anindita Adikaputri Vinaya and Tiffani Febiola Aciandra. "MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC) FEATURE FOR PUMP ANOMALY DETECTION IN NOISY ENVIRONMENTS". In: *Jurnal Rekayasa Mesin* 15.2 (2024), pp. 1175–1186. ISSN: 2477-6041, 2338-1663. DOI: 10.21776/jrm.v15i2.1815.

[124] Hanyang Wang, Yuxuan Yang, Hongjun Wang, and Lihui Wang. *Unsupervised Multi-Attention Meta Transformer for Rotating Machinery Fault Diagnosis*. arXiv:2509.09251 [cs]. Sept. 2025. DOI: 10.48550/arXiv.2509.09251. URL: http://arxiv.org/abs/2509.09251 (visited on 03/15/2026).

[125] Xin Wang, Dongxing Mao, and Xiaodong Li. "Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network". In: *Measurement* 173 (2021), p. 108518. ISSN: 02632241. DOI: 10.1016/j.measurement.2020.108518.

[126] Kunbo Xu, Zekai Zong, Dongjun Liu, Ran Wang, and Liang Yu. "Deep Learning-Based Sound Source Localization: A Review". In: *Applied Sciences* 15.13 (2025), p. 7419. ISSN: 2076-3417. DOI: 10.3390/app15137419.

[127] Lianchen Xu, Xiaohui Jin, Zhen Li, Wanquan Deng, Demin Liu, and Xiaobing Liu. "Particle Image Velocimetry Test for the Inter-Blade Vortex in a Francis Turbine". In: *Processes* 9.11 (2021), p. 1968. ISSN: 2227-9717. DOI: 10.3390/pr9111968.

[128] Fangyong Xue, Chang Liu, Feifei He, and Zeping Bai. "Acoustics-Augmented Diagnosis Method for Rolling Bearings Based on Acoustic–Vibration Fusion and Knowledge Transfer". In: *Sensors* 25.16 (2025), p. 5190. ISSN: 1424-8220. DOI: 10.3390/s25165190.

[129] Junyang Yang, Jiuxin Cao, and Chengge Duan. "Towards Realistic Industrial Anomaly Detection: MADE-Net Framework and ManuDefect-21 Benchmark". In: *Applied Sciences* 15.20 (2025), p. 10885. ISSN: 2076-3417. DOI: 10.3390/app152010885.

[130] Weijia Yang, Zhigao Zhao, Juan Ignacio Pérez-Díaz, Julian David Hunt, Elena Vagnoni, Jonas Kristiansen Nøland, Emanuele Quaranta, Ran Wang, Xudong Li, and Yongguang Cheng. "Pumped storage hydropower operation for supporting clean energy systems". In: *Nature Reviews Clean Technology* 1.7 (2025), pp. 454–473. ISSN: 3005-0685. DOI: 10.1038/s44359-025-00057-x.

[131] Woonghee Yeo and Mitsuharu Matsumoto. "Fault Diagnosis Systems for Robots: Acoustic Sensing-Based Identification of Detached Components for Fault Localization". In: *Applied Sciences* 15.12 (2025), p. 6564. ISSN: 2076-3417. DOI: 10.3390/app15126564.

[132] Yueyun Yu, Mingyuan Gao, Benjamin K. Ng, and Chan-Tong Lam. "SRP-DPCRN-IASDNet: A Blind Sound Source Location Method Based on Deep Neural Networks". In: *Mathematics* 14.4 (2026), p. 698. ISSN: 2227-7390. DOI: 10.3390/math14040698.

[133] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. "Deep Sets". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[134] Zilong Zhang, Zhibin Zhao, Xingwu Zhang, Chuang Sun, and Xuefeng Chen. *Industrial Anomaly Detection with Domain Shift: A Real-world Dataset and Masked Multi-scale Reconstruction*. arXiv:2304.02216 [cs]. Sept. 2023. DOI: 10.48550/arXiv.2304.02216. URL: http://arxiv.org/abs/2304.02216 (visited on 03/15/2026).

[135] Weiqiang Zhao, Alexandre Presas, Mònica Egusquiza, David Valentín, Eduard Egusquiza, and Carme Valero. "On the use of Vibrational Hill Charts for improved condition monitoring and diagnosis of hydraulic turbines". In: *Structural Health Monitoring* 21.6 (2022), pp. 2547–2568. ISSN: 1475-9217, 1741-3168. DOI: 10.1177/14759217211072409.

[136] Zilong Zhou, Yichao Rui, Xin Cai, Riyan Lan, and Ruishan Cheng. "A Closed-Form Method of Acoustic Emission Source Location for Velocity-Free System Using Complete TDOA Measurements". In: *Sensors* 20.12 (2020), p. 3553. ISSN: 1424-8220. DOI: 10.3390/s20123553.

[137] Zilong Zhou, Yichao Rui, Xin Cai, and Jianyou Lu. “Constrained total least squares method using TDOA measurements for jointly estimating acoustic emission source and wave velocity”. In: *Measurement* 182 (2021), p. 109758. ISSN: 0263-2241. DOI: 10.1016/j.measurement.2021.109758.

Appendix A
Hyperparameters and Reproducibility

This appendix records the complete hyperparameter configuration of the four learned stages (the per-modality encoder V1, the fusion encoder V2, the conditional-flow anomaly head V3, and the localization head V4), the model-selection grids over which the capacity and regularization settings were chosen, and the trained parameter count of each stage. The values are those recorded in the run manifests of the five-seed evaluation of Chapter 6; they are the configuration that produced every reported number, read back directly from those manifests. Unless stated otherwise, every stage uses the AdamW optimizer, a recording-level training/validation split with a validation ratio of 0.3, and seed 42 for the headline numbers; the multi-seed robustness sweep additionally uses seeds $\{1337, 2024, 7, 99\}$. The acoustic analysis grid ($n_{\text{fft}} = 4096$, hop $= 2048$, $n_{\text{mels}} = 96$) is shared by all stages and is justified by the empirical sweep of Chapter 3.

A.1 Stage hyperparameters

A.2 Model-selection grids

The capacity and regularization settings of §A.1 were chosen by the sweeps in Table A.4, each evaluated on a held-out selection metric and then re-trained over the multiple seeds above. Capacity is written as the swept pair of dimensions.

A.3 Parameter counts

Table A.5 reports the trained tensor count of each stage. The per-modality trunks are pretrained in V1 and their weights are inherited by the V2 fusion encoder, so they are not counted again in the deployed total: the chained inference pipeline is V2 $\rightarrow$ V3 + V4, about 1.15 million parameters in all.

Table A.1: Encoder hyperparameters. V1 pretrains the two per-modality trunks contrastively; V2 inherits them and trains the cross-attention fusion block and the context pool jointly. "n/a" marks a setting that does not apply to that stage.

Hyperparameter	V1	V2
Window length / stride (s)	2.0/1.0	2.0/1.0
Epochs	12	12
Batch size	16	16
Learning rate	10^{-3}	10^{-3}
Weight decay	10^{-4}	10^{-4}
Feature / embed dim	64/64	64/64
Attention heads	4	4
Projection dim	32	32
CWT scales	32	32
Acoustic CNN width multiplier	1	1
Gain jitter (dB)	9.0	9.0
Channel dropout	0.2	0.2
SpecAugment freq / time mask	12/16	12/16
Contrastive (NT-Xent) weight / temperature	1.0/0.1	1.0/0.1
Latent-mask fraction / weight	n/a	0.3/1.0
Cross-modal-alignment weight / temperature	n/a	0.5/0.1
Vibration modality dropout	n/a	0.5
Context PMA seeds	n/a	2

Table A.2: Conditional-flow anomaly head (V3) and per-cluster thresholds.

Hyperparameter	Value
Affine coupling layers	6
Scale / translation hidden dim	64
Hidden layers per scale/translation net	2
Scale bound $s_{\max}$	2.0
Coupling-MLP dropout	0.1
Context PMA-2 pool heads	4
Epochs	15
Batch size	32
Learning rate	10^{-3}
Weight decay	10^{-4}
Threshold clusters K	3
Threshold percentile	95

Table A.3: Localization head (V4). The residual half-range r is the architecture default used for every reported localization number; see §A.2 and §6.7 for its sweep.

Hyperparameter	Value
3-D CNN feature dim	64
TDOA feature dim	32
Residual-MLP hidden dim	64
TDOA pool heads	2
Soft-argmax temperature τ	1.0
Residual half-range r (m)	0.20
Epochs	30
Batch size	8
Learning rate	10^{-3}
Weight decay	10^{-4}
Head dropout	0.1
Smooth-L1 β (cm)	1.0
Target-position noise (m)	0.002
SRP-volume noise std	0.02
TDOA jitter (m)	0.001

Table A.4: Model-selection grids and selected values. The residual half-range is selected at 0.20 m, the architecture default that produces every reported localization number; the robustness sweep of §6.7 notes that widening it to 0.30 m yields a marginally lower held-out MAE, but this value was not adopted.

Stage	Axis	Grid	Selected
Encoder	Acoustic CNN width mult.	$\{1,2\}$	1
V3 flow	Coupling-MLP dropout	$\{0.0,0.1,0.2,0.3\}$	0.1
V3 flow	Weight decay	$\{10^{-4},5\times10^{-4},10^{-3}\}$	10^{-4}
V3 flow	Capacity (layers, hidden)	$\{(4,32),(6,64),(8,128)\}$	$(6,64)$
V4 head	Head dropout	$\{0.0,0.1,0.2,0.3\}$	0.1
V4 head	Weight decay	$\{10^{-4},5\times10^{-4},10^{-3}\}$	10^{-4}
V4 head	Capacity (CNN, hidden)	$\{(32,32),(64,64),(128,128)\}$	$(64,64)$
V4 head	Residual half-range (m)	$\{0.10,0.20,0.30\}$	0.20
V4 head	Target-position noise (m)	$\{0.002,0.010,0.020\}$	0.002
V4 head	SRP-volume noise std	$\{0.02,0.10,0.20\}$	0.02

Table A.5: Trained parameter count per stage (state-dict tensors).

Stage	Parameters
Per-modality encoder, acoustic (V1)	190 340
Per-modality encoder, vibration (V1)	149 284
Fusion encoder, incl. both trunks (V2)	440 616
Conditional flow (V3)	501 120
Context pool for the flow input (V3)	41 984
Localization head (V4)	162 471
Deployed pipeline (V2 + V3 + V4)	1 146 191

Appendix B
Supplementary Anomaly-Head Experiments

This appendix records two lines of work on the anomaly head that sit *outside* the frozen paradigm comparison of Chapter 6. Neither changes a research-question verdict, since each is a different architecture from the chained system those verdicts rest on. They are documented here because each points to a concrete next step for a field deployment, and because the second of them now carries five-seed evidence that bears directly on the recall limitation of the deployed anomaly rule. Both were developed after the paradigm-comparison protocol was frozen, so they are kept apart from the headline tables to keep every reported comparison on a single fixed protocol (Chapter 5); the scoping note of §7.7 introduces them in the discussion, and the detail follows here.

B.1 Post-Freeze Conditioning Refinements

The conditional flow evaluated in the main text conditions its affine-coupling layers on the operating context through FiLM while keeping a fixed standard-normal base, and scores the pooled fused tokens directly (§4.5). Three refinements of that head were implemented afterward. Each targets a specific weakness visible in the results of Chapter 6, in particular the modest conditioning gain (§6.4.1) and the seed-sensitive per-cohort recall (Table 6.5).

Context-conditional flow base. The fixed standard-normal base is replaced by a base whose mean and variance are predicted from the context vector $\boldsymbol{c}_t$ by a small, zero-initialized network, so the center of the healthy density moves with the operating regime and the score is regime-normalized by construction rather than only through the coupling layers. This targets the cross-regime confound in which a single-regime fault is scored against a pooled, multi-regime baseline.

Impulse-and-spectral anchor. A diagnostic probe shows that the self-supervised encoder optimizes the impulsiveness of a knock out of its embedding: a ridge regressor cannot recover the crest factor from the flow-input embedding ($R^2 \approx 0$). The anchor restores that information by concatenating the validated hand-crafted impulse and spectral feature families (crest, impulse, clearance and shape factors, kurtosis, knock count, and spectral-kurtosis, alongside spectral centroid, spread, flatness, entropy, rolloff, and band-energy ratios) to the flow input, computed on the same windowed features the en-

coder already consumes so no raw re-windowing is introduced. These features cannot be optimized away, so the conditional density sees the anomaly while remaining conditioned on $\boldsymbol{c}_t$.

Empirical-Bayes threshold shrinkage. Each per-cluster percentile threshold is pulled toward the global pooled percentile by an empirical-Bayes weight $w = n_k/(n_k + \lambda)$, so a small or noisy cluster borrows strength from the stable global boundary while a large cluster keeps its own. This guards against the failure mode in which a single small cluster's mis-fit percentile inflates the held-out healthy alert rate.

The outcome of a single-seed probe is mixed, and it is the reason these refinements were not adopted for the headline evaluation. Integrated together into the chained pipeline, they drive the RQ1 mode-discovery score to near-perfect on the labeled cohorts, but they degrade the other research questions relative to the frozen configuration, so the net effect on the system as a whole is negative. Because the effect was not validated under the same five-seed protocol that governs every reported number, no gain is claimed from it, and the simpler fixed-base head remains the system of record. A multi-seed re-evaluation of these refinements remains an open step for the anomaly head.

B.2 Deep Impulse-Aware Flow

The anchor of §B.1 treats a symptom of the self-supervised objective: that contrastive and cross-modal-alignment training discard transient structure. A separate, standalone model addresses the cause directly. Instead of scoring a contrastively trained embedding, it reads the raw waveform window through a per-modality one-dimensional CNN that is trained end to end with a conditional normalizing flow on the healthy negative log-likelihood, a one-class density objective rather than a contrastive one, so transients are preserved rather than collapsed. The validated hand-crafted impulse and spectral features are again concatenated as a recall anchor, and a learned low-dimensional context head with a context-conditional flow base normalizes the healthy density per operating regime so that one global threshold transfers across campaigns. The anomaly score is the per-modality negative log-likelihood of the anchored embedding under the context, sum-fused across modalities, fit on healthy windows only.

This model is a different architecture trained under a different objective from the chained system of Chapter 4, and it does not slot into the unimodal/late/intermediate paradigm comparison, so it is reported here as a forward-looking supplement rather than as a research-question result. It is, however, evaluated under the same five-seed discipline

that governs the main text, on the seed set $\{42, 1337, 2024, 7, 99\}$, so its stability can be read directly. Table B.1 gives its recording-level detection: the flow is fit on the pooled healthy windows of D2, D3, and D4, calibrated to a 0.05 healthy false-positive rate (achieved 0.050 at the median, with no spread across seeds), and tested on each campaign, with D5 held out entirely from fitting.

Two readings matter. The first is detection quality. The held-out D5 recording-level ROC-AUC reaches a five-seed median of 0.973, and rises to 0.982 once a few-shot adaptation on a small healthy slice of D5 is allowed, at a held-out false-positive rate of 0.059; the in-fit campaigns D2 and D3 reach 0.981 and 0.964. These are clean recording-level detection numbers of the kind the sparse per-window cohorts of Chapter 6 could not yield for the chained head. The second reading concerns recall, and it is the one that bears on the deployed anomaly rule. At the calibrated 0.05 operating point the fraction of anomaly recordings flagged is 0.91 on D2, 0.94 on D3, and 0.96 on the held-out D5 after adaptation, an order of magnitude above the late-fusion AND rule's recall on the same cohorts (Table 6.5). The exception is D4, the sparse cohort whose knock recordings are mostly healthy background, where the flagged fraction falls to 0.35. The impulse-aware flow therefore supplies the usable-recall operating point that the parameter-free conjunction of Chapter 6 trades away for its near-zero false-alarm rate, holding the same 0.05 healthy target while recovering most anomaly recordings on every cohort the array resolves well.

Table B.1: Deep impulse-aware flow, recording-level detection, median [min, max] over the five seeds $\{42, 1337, 2024, 7, 99\}$. The flow is fit on the pooled healthy windows of D2, D3, and D4 at a 0.05 healthy false-positive target (achieved 0.050, no spread); D5 is held out from fitting. ROC-AUC and PR-AUC are recording level; *Flag@thr* is the fraction of anomaly recordings flagged at the calibrated operating point, the recording-level recall. The D5 row is reported both zero-shot and after a few-shot adaptation on a small healthy slice of D5 (held-out false-positive rate 0.059).

Campaign	Held out	ROC-AUC	PR-AUC	Flag@thr
D2	no	0.981 [0.98, 0.99]	0.963 [0.96, 0.98]	0.907 [0.87, 0.96]
D3	no	0.964 [0.92, 0.99]	0.895 [0.85, 0.94]	0.944 [0.88, 0.99]
D4	no	0.856 [0.78, 0.88]	0.928 [0.89, 0.94]	0.352 [0.31, 0.37]
D5 (zero-shot)	**yes**	0.973 [0.78, 0.99]	0.996 [0.97, 1.00]	1.000 [1.00, 1.00]
D5 (few-shot)	**yes**	**0.982 [0.92, 0.99]**	0.998 [0.99, 1.00]	0.960 [0.83, 0.97]

Table B.2: Deep impulse-aware flow, selected configuration.

Hyperparameter	Value
Mel bands / time frames	64/64
Window length / stride (s)	1.0/0.5
Embedding dim	32
Context dim	16
Anchor features	16
Flow layers / hidden dim	8/32
Dropout	0.1
Epochs / batch size	40/64
Learning rate / weight decay	$3 \times 10^{-4}/10^{-4}$
Augmentation	strong
Healthy FPR target	0.05

One caution bounds this result. Like every detection result in this thesis, it is measured against deliberately induced casing impacts rather than naturally occurring incipient faults (§7.7), so the recording-level ROC-AUC reflects separability of knocks from healthy background, not sensitivity to real damage. The recall numbers inherit the same bound, and the D4 shortfall is a reminder that even on knocks a sparse cohort remains hard. With that caveat, the deep impulse-aware flow is the most promising single direction for the anomaly head: it pairs the AND rule's near-zero false-alarm rate with a usable recall the conjunction cannot provide, and it is the natural candidate to carry, under the full protocol, into the instrumented campaign outlined in Chapter 8.

Appendix C
Source Code Availability

The complete source code developed for this thesis is publicly available in the following repository:

https://github.com/ZakariaOmarar/Master-Thesis

All experiments, models, and figures reported in this thesis were produced with the code hosted at this address. The repository contains the full processing pipeline, the implementations of the four learned stages whose configurations are recorded in Appendix A (the per-modality encoder, the fusion encoder, the conditional-flow anomaly head, and the localization head), the supplementary anomaly-head experiments of Appendix B, and the scripts that regenerate every reported number and figure. The version of record is the state of the repository at the time of submission; together with the run manifests and hyperparameter grids of Appendix A, it is intended to make every result in this thesis reproducible.

The repository shows only a single, final commit because the GitHub account originally used during development was suspended by GitHub; the code was therefore re-published under the account above as one consolidated commit, and the absence of an earlier commit history is a consequence of that suspension rather than of the development process itself.

Declaration of Authorship

‘I hereby declare,

- that I have written this thesis independently;
- that I have written the articles and contributions using only the aids listed in the index;
- that all parts of the thesis produced with the help of aids (incl. AI-Tools) have been precisely declared;
- that I have mentioned all sources used and cited them correctly according to established academic citation rules; this also includes the comprehensible disclosure of all personal publications;
- that I have acquired all immaterial rights to any materials I may have used, such as images or graphics, or that these materials were originally created by myself;
- that I am aware of the legal provisions regarding the publication and dissemination of parts or the entire thesis and that I comply with them accordingly;
- that I am aware that the University will prosecute a violation of this Declaration of Authorship and that disciplinary as well as criminal consequences may result, which may lead to expulsion from the University or to the withdrawal of my title.’

By submitting this thesis, I confirm through my conclusive action that I am submitting the statutory declaration, that I have read and understood it, and that it is true.

St. Gallen, June 30, 2026 Signature: Zakaria Omarar

Declaration of Auxiliary Aids

The creation of this dissertation involved a lot of third-party software, including AI-based tools. I declare all uses of third-party software in the dissertation text that contributed non-trivially to the dissertation's content and are non-standard in such research. Further, by the University of St. Gallen's decree II.B.1.20 section 5.3, I explicitly declare the following uses:

- Private proofreading, DeepL, Google Gemini, and OpenAI ChatGPT (Plus), Antidote in their current versions available between between February 2026 and June 2026, have been used for spellchecking, grammar checking, and enhancing the conciseness and clarity of the dissertation text.
- Claude Opus 4.7, Gemini Pro in their current versions available between between February 2026 and June 2026, have been used for brainstorming.