Learning the Language of Vibration: A Self-Supervised Transformer Foundation Model for PHM

Giuseppe Mannone[1], Paula Fischer[1], Martin Dazer[1]

[1] *Institute of Machine Components (IMA), University of Stuttgart, Stuttgart, Germany*
giuseppe.mannone@ima.uni-stuttgart.de
paula.fischer@ima.uni-stuttgart.de
martin.dazer@ima.uni-stuttgart.de

ABSTRACT

Industrial prognostics and health management (PHM) increasingly relies on vibration-based deep learning, yet field deployment remains limited by two practical constraints: (i) fault labels are scarce and expensive, and (ii) models trained on one machine or dataset often degrade under distribution (domain) shift (different sensors, sampling rates, loads, and signal conventions). These constraints motivate *vibration foundation models*: reusable encoders trained once on large collections of unlabeled raw vibration recordings and adapted to new assets with minimal supervision. This paper presents **VibFM**, a Transformer encoder trained via self-supervised *masked spectrogram modeling* in the spirit of masked language modeling and masked autoencoders. Raw waveforms from 16 open datasets totaling $\approx$ 400 hours are standardized into 128×128 log-magnitude short-time Fourier transform (STFT) spectrograms and paired with a compact conditioning vector that encodes sampling rate and time/frequency resolution. Pre-training reconstructs masked time–frequency patches, encouraging the encoder to capture transferable vibration primitives such as persistent narrowband ridges, modulation sidebands, and impulsive transients. Transfer is evaluated on the held-out Paderborn University and KAt-DataCenter bearing benchmark (excluded from pre-training) using leakage-resistant *bearing-level* splits. On three-class fault diagnosis, frozen VibFM features substantially improve over training from scratch, while end-to-end fine-tuning provides the strongest performance. For reconstruction-based anomaly detection, adapting a decoder on healthy target data yields reconstruction-error scores that separate healthy from damaged states across operating conditions. Masked reconstructions and pooling-attention visualizations provide qualitative audits of learned time–frequency structure, and the limits of these interpretability probes are discussed.

1. INTRODUCTION

Rotating machinery is pervasive in industrial systems, and unexpected failures can cause safety incidents, production downtime, and high maintenance costs (Jardine, Lin, & Banjevic, 2006). Vibration monitoring is therefore widely deployed because it is inexpensive and sensitive to bearing, gearbox, and drivetrain damage. Despite strong progress in deep learning for vibration diagnostics (R. Zhao et al., 2019) and a proliferation of architectures, two bottlenecks repeatedly limit field adoption:

- **Label scarcity:** Damage events are rare, and collecting or annotating fault/damage data for each new asset is expensive.
- **Distribution (domain) shift:** Vibration signatures depend on sensor placement, sampling rate, load/speed regime, and even preprocessing conventions; models trained on one dataset often transfer poorly (Pan & Yang, 2010; Xiao et al., 2025).

A recurring observation in vibration analysis is that many tasks share the same physical building blocks, for example, rotation harmonics, amplitude modulation with sidebands, and impulsive transients (Randall, 2011). Standard supervised pipelines often relearn these primitives from scratch for each dataset or asset, which is computationally inefficient (wasting compute and labeled data) and can yield representations that overfit dataset-specific artifacts (resolution, bandwidth, and noise texture).

These considerations motivate a *train once, adapt many times* paradigm for vibration PHM: pre-train a general-purpose encoder on diverse vibration recordings treated as *unlabeled* for self-supervised training (labels, when present, are ignored) and reuse it via frozen features or fine-tuning on new assets (Li, Wang, & Sun, 2024; Lai et al., 2024). The objective is not language-model scale, but transferable representations under real-world heterogeneity. A key technical challenge is that vibration datasets are not naturally aligned: different sampling

rates and window lengths mean that a fixed-size spectrogram grid can correspond to different physical time and frequency resolutions. To reduce this ambiguity, inputs are standardized and the model is explicitly conditioned on acquisition metadata.

The language of vibration is learned via *masked spectrogram modeling*. Inspired by masked language modeling (Devlin, Chang, Lee, & Toutanova, 2019) and masked autoencoders (He et al., 2022), a large fraction of time–frequency patches is masked and a Transformer encoder–decoder is trained to reconstruct them. Because reconstruction requires global context, the encoder is encouraged to internalize regularities that recur across datasets and machines.

The main contributions are:

- **Vibration foundation encoder:** VibFM is introduced as a Transformer-based encoder pre-trained with masked spectrogram reconstruction on a heterogeneous multi-source corpus.
- **Standardized time–frequency input with conditioning:** Raw waveforms are mapped to fixed-size 128×128 spectrogram images and paired with a compact conditioning vector $\mathbf{c}$ specifying sampling rate and time/frequency resolution, mitigating ambiguity induced by resizing.
- **Transfer to two PHM tasks:** Transfer is evaluated on a held-out bearing benchmark for (i) three-class diagnosis using lightweight heads and (ii) reconstruction-based anomaly detection via a transferred encoder and an adapted decoder.
- **Audits of learned structure:** Masked reconstructions, latent-space visualization, and pooling-attention maps provide qualitative audits of learned time–frequency structure, and the limits of these probes are made explicit.

The downstream evaluation is designed to test whether multi-source self-supervised pre-training improves transfer under unit-level generalization, rather than whether the architecture is optimized for a single benchmark. Quantitative results are reported in Section 5, where frozen transfer, end-to-end fine-tuning, and healthy-only anomaly detection are evaluated under bearing-level splits.

2. RELATED WORK

2.1. Time–frequency representations for vibration

Vibration diagnostics has traditionally relied on physically interpretable representations such as Fourier spectra, envelope spectra, order spectra, and cyclic or modulation spectra (Randall, 2011; Tandon & Choudhury, 1999). These tools are highly effective when bearing geometry, speed regime, resonance bands, and expected fault frequencies are known. The objective of VibFM is different: the encoder is trained on a heterogeneous multi-source corpus where datasets differ in sampling rate, sensor placement, component type, operating regime, and available metadata.

In this setting, a log-magnitude short-time Fourier transform (STFT) spectrogram provides a practical common representation. It preserves joint time–frequency localization of narrowband ridges, resonance-dominated energy, modulation-related structure, and impulsive bursts, while remaining generic enough to apply across datasets without dataset-specific band selection. The spectrogram is therefore not used merely because it is image-like or convenient for Transformers. It provides a representation on which masked reconstruction can be posed: missing local patches can only be reconstructed well when the encoder has learned broader time–frequency context.

At the same time, spectrograms are not claimed to replace envelope or cyclic analyses. Rather, VibFM is complementary to these methods: classical signal processing provides physics-based interpretability and fault-frequency validation, while the learned encoder provides reusable features for label-scarce transfer and healthy-only anomaly detection. A known risk of spectrogram-based deep learning is that models may exploit test-rig-specific frequency response functions (FRFs) or average spectral profiles rather than fault-related mechanisms. Liefstingh, Taal, Echeverri Restrepo, and Azarfar (2021) showed that removing time-domain information from spectrograms had limited impact on classification performance, suggesting that some networks learned setup-specific frequency-response signatures. VibFM is designed to reduce, but not eliminate, this failure mode by pre-training on 16 heterogeneous datasets, limiting source domination through dataset-aware sampling, using per-window amplitude standardization and acquisition conditioning, and evaluating transfer on the held-out Paderborn benchmark with bearing-level splits.

2.2. Transformers on spectrograms

Vision Transformers (Dosovitskiy et al., 2021) operate on sequences of image patches and are well suited to spectrogram inputs, where long-range dependencies exist both in time and frequency. In audio, spectrogram Transformers such as AST (Gong, Chung, & Glass, 2021) demonstrate that patch-wise attention can model global time–frequency structure effectively. The proposed approach adopts a similar architectural bias but targets vibration PHM, where domain shift and evaluation leakage are central concerns.

2.3. Self-supervised and masked modeling objectives

Self-supervised learning reduces label dependence by learning from unlabeled data (Wang, Liu, Ge, & Peng, 2022). Masked modeling objectives are especially attractive because

they scale efficiently and encourage learning global structure by reconstructing missing content (Devlin et al., 2019; He et al., 2022; Gong, Lai, Chung, & Glass, 2022; P.-Y. Huang et al., 2022; Chen et al., 2023). In contrast to contrastive objectives, masked reconstruction provides a direct pixel-level target in spectrogram space, which is convenient for reconstruction-based anomaly detection.

2.4. Leakage-resistant evaluation in PHM

In PHM, evaluation protocols can dominate reported performance. Segment-level random splits often leak unit identity into the test set, inflating accuracy (Hendriks, Dumond, & Knox, 2022). The Paderborn bearing benchmark provides a bearing-level protocol that enforces generalization to unseen bearing codes (Lessmeier, Kimotho, Zimmer, & Sextro, 2016), which is adopted here to make transfer claims meaningful.

3. METHODOLOGY: VIBFM PRETRAINING AND TRANSFER

This section describes the multi-source pre-training corpus, the standardized spectrogram representation with acquisition conditioning, and the masked modeling objective used to train VibFM, followed by the downstream transfer setup.

3.1. Overview and design goals

The objective is to learn a *transferable* representation for vibration-based PHM that can be reused across datasets differing in sensor hardware, mounting, sampling rate, and operating regime. The target setting assumes that labels on the *target* machine are scarce, while large amounts of unlabeled vibration are available from other machines.

A practical vibration foundation model should satisfy the following requirements:

- **Multi-source compatibility:** Heterogeneous datasets should be ingested without dataset-specific feature engineering.
- **Scale-awareness:** Visually similar spectrogram patterns should not be conflated when they correspond to different physical time and frequency scales (e.g., due to different sampling rates f_s or short-time Fourier transform (STFT) settings).
- **Transfer-friendly embeddings:** A compact feature vector should be produced for lightweight frozen transfer or efficient fine-tuning.
- **Auditable behavior:** Beyond aggregate metrics, reconstructions and attention should be inspectable to verify that the model focuses on physically meaningful time–frequency structure rather than dataset textures.

Each vibration segment is treated as a time–frequency image and learned via *masked spectrogram modeling*: a large subset of spectrogram patches is masked and a Transformer is trained to reconstruct them (masked-autoencoder-style) (He et al., 2022; Gong et al., 2022; P.-Y. Huang et al., 2022). The high masking rate makes reconstruction impossible by local interpolation alone; instead, the encoder must infer missing content from global context. Intuitively, this is analogous to language models that infer a hidden word from the surrounding sentence. Here, VibFM observes only part of a vibration spectrogram and must reconstruct missing patches from neighboring ridges, modulation patterns, broadband bursts, and longer-range recurrence. In vibration signals, this encourages the encoder to internalize recurring physical motifs, such as persistent narrowband structures, modulation sidebands, and broadband impulsive transients, that are shared across machines and fault mechanisms (Randall, 2011).

Figure 1 provides the full pipeline. The remainder of this section details the pre-training corpus and sampling strategy, the standardized spectrogram representation and acquisition conditioning, the encoder–decoder architecture, and the self-supervised objective and optimization.

3.2. Pre-training corpus and sampling

A heterogeneous unlabeled corpus is assembled from 16 open datasets spanning bearings, gearboxes, and drivetrain components, totaling approximately 400 hours of vibration data. The Paderborn benchmark used for downstream evaluation (Section 4.2) is **excluded** from pre-training so that downstream results reflect transfer rather than accidental reuse of evaluation data. To encourage broad reuse, the training stream is kept diverse rather than being dominated by a single source.

Each pre-training sample is a single-channel vibration window $x[n]$ of duration T seconds ($T = 1.5\,\mathrm{s}$); for recordings with multiple channels (e.g., triaxial accelerometers), each channel is treated as an independent sample rather than assuming consistent axis alignment across datasets. This increases diversity and avoids imposing a sensor-coordinate convention. Lightweight quality control is applied to avoid degenerate inputs that can destabilize self-supervised training:

- windows with missing values (NaNs) are discarded to avoid undefined STFT inputs and unstable gradients;
- windows with near-zero variance (idle or disconnected sensors) are discarded;
- optional saturation checks can be enabled for datasets with known clipping artifacts.

A common failure mode in multi-source self-supervised learning is dataset domination: the largest dataset overwhelms the training stream. To mitigate this, dataset-aware sampling is used. Let n_d be the number of eligible windows

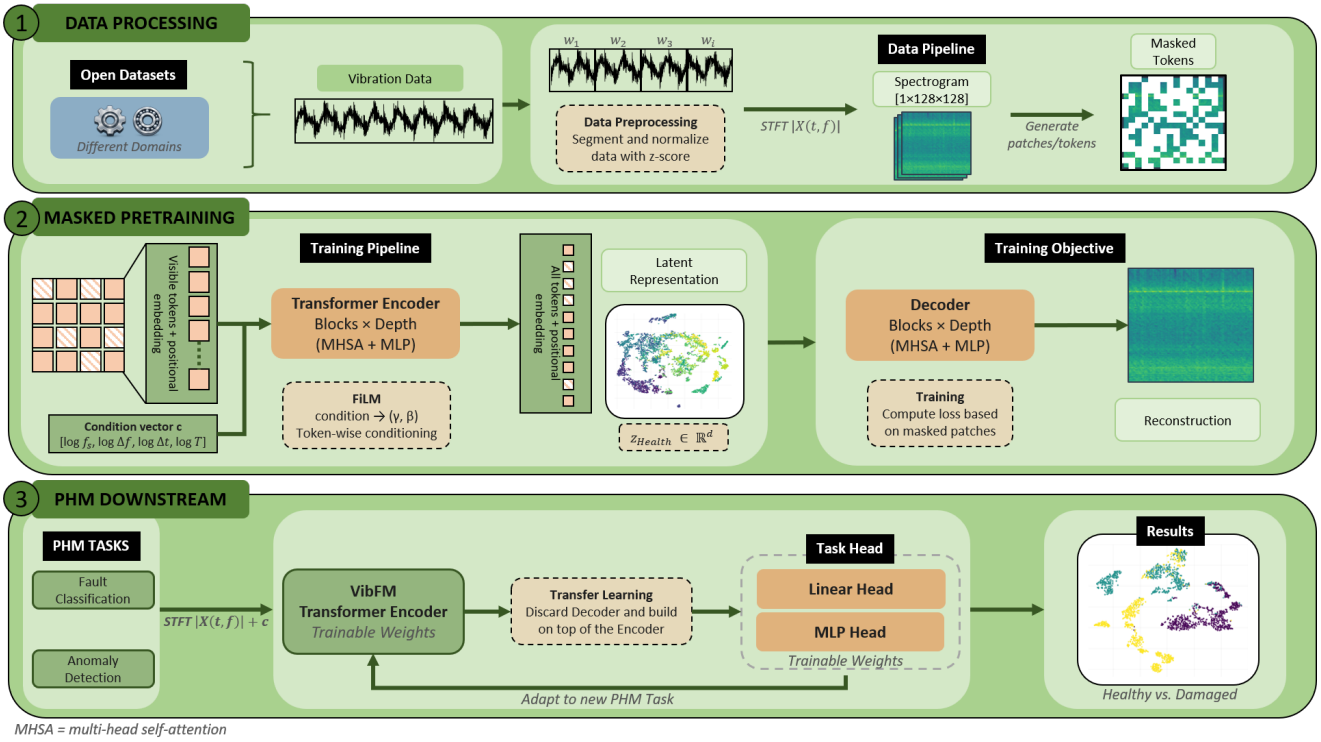


Figure 1. VibFM pipeline: (1) standardize waveforms to conditioned spectrograms, (2) masked spectrogram pre-training, (3) transfer to diagnosis and anomaly detection on a held-out benchmark. Abbreviations in the schematic: MHSA = multi-head self-attention; MLP = multilayer perceptron.

in dataset d. Datasets are sampled with probability

$$p(d) \propto n_d^{\alpha}, \tag{1}$$

where $\alpha \in [0, 1]$ controls the balance between proportional sampling ($\alpha = 1$) and near-uniform per-dataset sampling ($\alpha \to 0$).

Within each sampled dataset, a recording is selected and a random time offset is used to extract a window. This yields a training stream that is diverse in operating regimes and measurement styles while remaining simple and reproducible. Figure 2 shows representative standardized STFT inputs from this corpus.

3.3. From waveforms to standardized spectrogram images

The central preprocessing challenge is that open vibration datasets vary widely in sampling rate and time–frequency resolution. VibFM therefore uses a two-part strategy: (i) a standardized spectrogram image $X \in [0, 1]^{1\times128\times128}$ for batching and patch tokenization, and (ii) an explicit conditioning vector $\mathbf{c}$ that tells the model what physical time/frequency scale the image corresponds to. Concretely, each waveform window is mapped to $(X, \mathbf{c})$ using the following steps.

1. **Segmentation and amplitude normalization:** Fixed-duration windows of length T seconds are extracted from each raw vibration recording. Let $x[n]$ denote one window sampled at effective rate f_s with $N = Tf_s$ samples. The DC component is removed and amplitude is standardized per window:

$$\begin{aligned} \tilde{x}[n] &= \frac{x[n] - \mu_x}{\sigma_x + \epsilon_x}, \\ \mu_x &= \frac{1}{N}\sum_{m=0}^{N-1} x[m], \\ \sigma_x &= \sqrt{\frac{1}{N}\sum_{m=0}^{N-1} \big(x[m] - \mu_x\big)^2}. \end{aligned} \tag{2}$$

   with a small ϵ_x (e.g., 10^{-8}) for numerical stability. This step reduces sensitivity to sensor gain and mounting differences while preserving relative spectral structure.

2. **Sampling-rate handling:** To reduce unnecessary heterogeneity while avoiding upsampling artifacts, recordings may be downsampled to a set of common buckets (e.g., 12.8/25.6/51.2/102.4 kHz) while retaining sufficient Nyquist bandwidth for the component.

3. **STFT and log-magnitude scaling:** For each normalized window $\tilde{x}[n]$, a one-sided STFT is computed us-

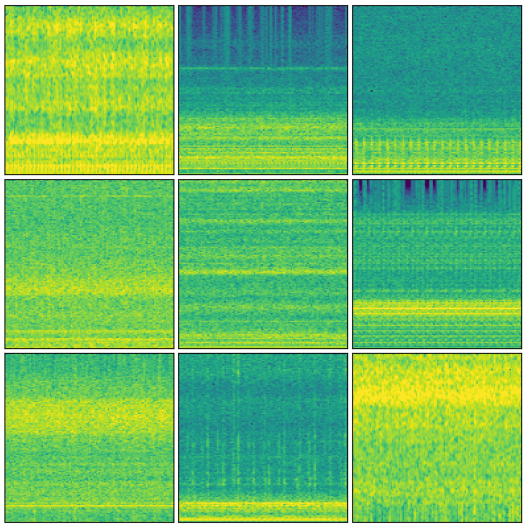


Figure 2. Representative standardized log-magnitude STFT inputs (128×128, scaled to $[0, 1]$) from the pre-training corpus.

ing a Kaiser window with $\beta = 14.0$, window length 12.5 ms and hop 6.25 ms. Denoting the complex STFT by $\mathcal{X}(t, k)$, log-magnitude is taken in decibels:

$$S_{\mathrm{dB}}(t, k) = 20 \log_{10}(|\mathcal{X}(t, k)| + \epsilon), \tag{3}$$

with a small ϵ for numerical stability and an explicit floor at -120 dB. The floor is applied immediately after log scaling to avoid extreme negative values before interpolation; the later clipping step sets the fixed reconstruction range $[-100, 0]$. This compression stabilizes dynamic range across datasets and makes transient and narrowband spectral structures simultaneously visible. The 12.5 ms window is a compromise between transient localization and frequency resolution, not an optimal setting for resolving all bearing characteristic frequencies in the classical spectral sense. For example, the Paderborn operating points use 900 and 1500 rpm, corresponding to shaft frequencies of 15 and 25 Hz and shaft periods of approximately 67 and 40 ms. Standard kinematic estimates for the 6203 bearing used in the benchmark place the ball pass frequency outer race (BPFO) and ball pass frequency inner race (BPFI) in the tens to low hundreds of hertz, so their periods can be comparable to or longer than a single STFT window. VibFM is therefore not expected to infer fault type from one sharply resolved low-frequency line in one frame. Instead, the full spectrogram segment captures repeated energy patterns, resonance-band excitation, modulation-related structure, and impulsive localized events across many frames. Multi-resolution STFTs, envelope spectra, or hybrid waveform/spectrogram tokenization are natural extensions when precise fault-frequency resolution is required.

4. **Standardization to a fixed grid:** The STFT produces dataset-dependent dimensions (number of time frames and frequency bins). To standardize token count and enable batching across datasets, both axes of S_{dB} are resampled to a fixed 128×128 grid by bilinear interpolation:

$$S_{\mathrm{grid}} \in \mathbb{R}^{128 \times 128}. \tag{4}$$

The frequency axis is mapped to $[0, f_s/2]$ (one-sided spectrum) and the time axis to the full window duration $[0, T]$. This resizing preserves *coarse pattern topology*, such as narrowband ridges, modulation-related spacing, and burst-like transients, while sacrificing the direct pixel-to-physics mapping, which is partly disambiguated via conditioning.

5. **Clipping and mapping:** To make reconstruction targets consistent across datasets and windows, values are clipped to a fixed dB range and linearly mapped to $[0, 1]$:

$$\begin{aligned} S_{\mathrm{clip}}(t, k) &= \mathrm{clip}\big(S_{\mathrm{grid}}(t, k),\ -100,\ 0\big), \\ X(t, k) &= \frac{S_{\mathrm{clip}}(t, k) + 100}{100}. \end{aligned} \tag{5}$$

Per-window min–max scaling is avoided because it can erase amplitude structure and encourage reliance on texture-like cues; the fixed range in Equation 5 yields a consistent pixel meaning under per-window waveform standardization.

6. **Acquisition-conditioning vector:** Resizing removes a direct mapping from pixel coordinates to physical units. To reduce scale ambiguity, a compact conditioning vector is attached:

$$\mathbf{c} = [\log_{10} f_s,\ \log_{10} \Delta f,\ \log_{10} \Delta t,\ \log_{10} T] \in \mathbb{R}^4, \tag{6}$$

where $\Delta f = \frac{f_s/2}{128}$ and $\Delta t = \frac{T}{128}$ are the effective frequency and time resolutions implied by the standardized grid. Conditioning provides the encoder with the missing physical context needed to interpret similar visual motifs at different scales.

This standardization makes it possible to pre-train a single encoder across many acquisition conventions, while $\mathbf{c}$ preserves the physical context needed for scale-aware transfer.

3.4. Architecture and self-supervised objective

Given the standardized spectrogram X and acquisition-conditioning vector $\mathbf{c}$ (Section 3.3), VibFM follows a masked-autoencoder-style encoder–decoder design (He et al., 2022): the encoder learns from visible patches, and a lightweight decoder reconstructs masked patches during pre-training. The decoder is discarded after pre-training, leaving a reusable encoder for downstream tasks.

Given $X \in [0,1]^{1\times128\times128}$, it is split into non-overlapping 8×8 patches with stride 8. The resulting patch grid has

$$N = \left(\frac{128}{8}\right)^2 \tag{7}$$

tokens. Each patch is flattened to $\mathbb{R}^{64}$ and mapped to a 384-dimensional embedding with a learned linear projection. Fixed 2D positional embeddings are added to obtain a token sequence $\{\mathbf{e}_i\}_{i=1}^{N}$. The encoder is a Vision Transformer-style stack (Dosovitskiy et al., 2021) with 8 layers and 8 attention heads. Each block alternates multi-head self-attention and a multilayer perceptron (MLP) with residual connections using the standard pre-norm formulation:

$$\begin{aligned}\mathbf{H}^{(\ell+1)} &= \mathbf{H}^{(\ell)} + \mathrm{MSA}\Big(\mathrm{LN}\Big(\mathbf{H}^{(\ell)}\Big)\Big),\\ \mathbf{H}^{(\ell+2)} &= \mathbf{H}^{(\ell+1)} + \mathrm{MLP}\Big(\mathrm{LN}\Big(\mathbf{H}^{(\ell+1)}\Big)\Big).\end{aligned} \tag{8}$$

where $\mathbf{H}^{(\ell)}$ stacks all tokens and $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^{384}$ denotes token i at layer ℓ.

Conditioning can be injected additively or via modulation. A feature-wise linear modulation (FiLM) mechanism is used: a small MLP maps $\mathbf{c}$ to token-wise scale and shift parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta}) \in \mathbb{R}^{384}$, and patch embeddings are modulated as

$$\tilde{\mathbf{e}}_i = \boldsymbol{\gamma}(\mathbf{c}) \odot \mathbf{e}_i + \boldsymbol{\beta}(\mathbf{c}). \tag{9}$$

This modulation allows the encoder to adapt its internal feature scaling depending on the physical meaning of the standardized grid (e.g., different f_s implies different physical frequency spacing in bins). Conditioning is applied at the encoder input; stronger variants apply conditioning inside each Transformer block.

During pre-training, a fraction 0.70 of patch positions is randomly masked. Let $\mathcal{M}$ denote the set of masked indices and $\mathcal{V}$ the visible set. Following the masked autoencoder approach (He et al., 2022), the encoder processes only visible tokens $\{\tilde{\mathbf{e}}_i\}_{i\in\mathcal{V}}$, which reduces computation at high mask ratios. A lightweight decoder (width 256, depth 4, 8 heads) receives the encoded visible tokens plus learned mask tokens for $\mathcal{M}$ and predicts the pixel values of masked patches. With 0.70 masking, the encoder sees only $1 - 0.70$ of tokens (e.g., $\approx 30\%$), forcing the model to use long-range context to reconstruct missing content rather than relying on local interpolation.

For downstream tasks, the decoder is discarded and only the encoder is retained. To obtain a fixed-dimensional embedding from patch tokens, attention pooling with 8 learnable latent tokens is used. The latent tokens attend over patch tokens and produce pooled summaries; the pooled summary is projected to two embeddings:

$$\mathbf{z}_{\text{health}} \in \mathbb{R}^{256}, \qquad \mathbf{z}_{\text{nuis}} \in \mathbb{R}^{256}. \tag{10}$$

Table 1. Key pre-training configuration and architecture hyperparameters for VibFM.

Item	Value
Input representation	128×128 log-magnitude STFT, scaled to $[0, 1]$
Window duration (pre-train)	1.5 s
STFT window / hop	12.5 ms / 6.25 ms (Kaiser $\beta = 14.0$)
dB floor / clip	-120 dB / $[-100, 0]$
Conditioning vector	$\mathbf{c} \in \mathbb{R}^4$ (Eq. 6), FiLM-style (Eq. 9)
Patch size / stride	8×8, stride 8
Encoder	8 layers, 8 heads, width 384
Decoder (pre-train only)	4 layers, 8 heads, width 256
Mask ratio	0.70
Pooling	attention pooling with 8 latent tokens
Embeddings	$\mathbf{z}_{\text{health}} \in \mathbb{R}^{256}$, $\mathbf{z}_{\text{nuis}} \in \mathbb{R}^{256}$
Optimizer	AdamW
Batch / Epochs	128 / 93

$\mathbf{z}_{\text{health}}$ is used for downstream diagnosis and transfer. $\mathbf{z}_{\text{nuis}}$ is an auxiliary channel intended to capture dataset- and acquisition-specific variation (useful for analysis and debugging). This factorization is conceptually motivated by the PHM setting: health information should be stable across datasets, while nuisance variability can still be represented. When dataset identities are available during pre-training, regularizers can be added to discourage dataset identity from being encoded in $\mathbf{z}_{\text{health}}$ (e.g., adversarial dataset classification or orthogonality penalties between $\mathbf{z}_{\text{health}}$ and $\mathbf{z}_{\text{nuis}}$). In the current experiments, this is treated as optional and the reconstruction objective is used as the primary learning signal.

Let $x_i \in \mathbb{R}^{64}$ denote the vectorized target pixels for patch i (after the preprocessing pipeline and optional spectrogram normalization), and let $\hat{x}_i$ be the corresponding prediction. Reconstruction is optimized on masked regions only:

$$\mathcal{L}_{\text{mask}} = \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} \|x_i - \hat{x}_i\|_1 . \tag{11}$$

An ℓ_1 loss is used because it is less sensitive to outliers and can preserve sharper structures compared to ℓ_2 in image-like reconstruction.

3.5. Pre-training optimization and reproducibility

VibFM is pre-trained on the unlabeled corpus using AdamW with weight decay 5×10^{-5} and a cosine learning-rate schedule implemented via `LambdaLR`, with 2000 warmup steps. Training uses a global batch size of 128 and runs for 192 594 update steps (93 epochs) on an NVIDIA RTX A6000 GPU, with data augmentation intentionally kept minimal: beyond random window sampling and masking, strong augmentations that could distort physically meaningful spectral patterns are avoided.

3.6. Audits of learned structure

Masked reconstructions provide a direct qualitative check that VibFM learns meaningful time–frequency structure and helps discourage exclusive reliance on dataset-specific textures. Successful reconstructions during pre-training typically recover narrowband ridge continuity, broadband transient energy, and coarse modulation/sideband patterns. This does not prove downstream usefulness, but it is a useful sanity check that the self-supervised objective is being solved in a physically plausible manner.

Attention maps are useful for *auditing* what information is routed into the embedding used downstream, but they must be interpreted with care. In VibFM, the weights of the attention-pooling module (Section 3.4) are visualized because this module directly determines which patches contribute most to the pooled representation. Let N be the number of patch tokens and let $L = 8$ be the number of learnable latent tokens used for pooling. The pooling module produces cross-attention weights

$$A^{(h)} \in \mathbb{R}^{L \times N}, \tag{12}$$

for each head h. A single patch-importance score is obtained by averaging over latent tokens and heads:

$$\bar{a}_j = \frac{1}{H\,L} \sum_{h=1}^{H} \sum_{\ell=1}^{L} A_{\ell j}^{(h)}, \qquad j \in \{1, \ldots, N\}. \tag{13}$$

For visualization, $\bar{\mathbf{a}}$ is reshaped to the 16×16 patch grid and upsampled to 128×128. For display, each attention map is normalized to $[0, 1]$ per example.

Because the maps are computed on a 16×16 patch grid and then upsampled for visualization, they should not be used to count individual over-rolling events or infer an exact impulse repetition frequency. They indicate which coarse time–frequency regions are routed into the pooled embedding, not a high-resolution event detector.

Figure 3 visualizes the pooling module's mean cross-attention for four pre-training examples (a)–(d), showing the input spectrogram (left) and the corresponding normalized attention map (right). Across panels (a)–(c), attention primarily concentrates on persistent narrowband ridges and a few localized transient bursts, patterns that align with common vibration signatures used in classical diagnostics (Randall, 2011). In contrast, panel (d) corresponds to a bearing damage segment, and its attention map shifts markedly toward repeated, broadband transient columns (impact-like structures), indicating that the pooling mechanism routes comparatively more information from these damage-related regions into the learned representation. Importantly, attention is not a causal explanation of model behavior (Jain & Wallace, 2019; Serrano & Smith, 2019): high attention indicates where information is aggregated, but it does not imply necessity or sufficiency for a downstream decision. We therefore use these maps as qualitative sanity checks and to motivate targeted future ablations, for example, by masking or deleting high-attention regions to test sensitivity.

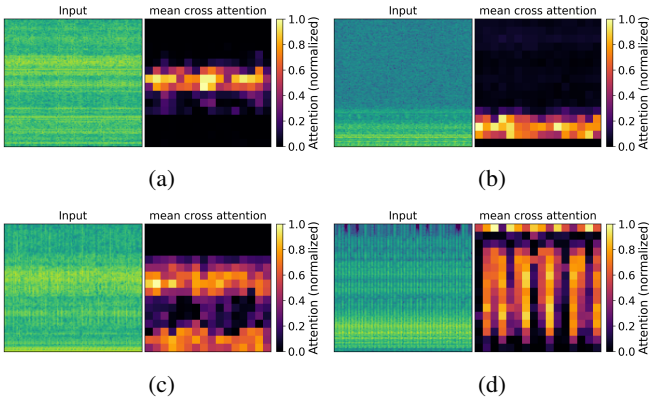


Figure 3. Pooling-module attention maps. Each panel shows the input spectrogram (left) and the normalized mean cross-attention (right), aggregated as in Eq. 13, where brighter regions contribute more to the pooled embedding. Panels (a)–(c) emphasize persistent narrowband ridges and localized bursts, while (d) (bearing damage) shifts attention toward repeated broadband transients. The term "narrowband ridges" is descriptive; no shared fundamental frequency is assumed.

Together, these visual audits complement the quantitative transfer evaluation described next.

4. DOWNSTREAM EVALUATION SETUP

This section specifies the downstream threat model, datasets, and protocols used to evaluate whether VibFM transfers under leakage-resistant unit-level splits.

4.1. Evaluation philosophy and threat model

The downstream evaluation tests a specific claim: *a self-supervised vibration foundation encoder should transfer to a new machine (or benchmark) without relying on unit-identity leakage*. In PHM, a major confound is that multiple measurements are often recorded from the same physical component (e.g., a specific bearing). Random measurement-level splits can therefore leak unit identity into the test set: models may partially memorize component identity or acquisition signatures rather than learning health-relevant structure, inflating test performance. To mitigate this confound, bearing-level splits are adopted, where train and test contain disjoint bearing codes (Hendriks et al., 2022). Because published results on this dataset are highly sensitive to split choice and leakage control, we focus on protocol-matched baselines (scratch, frozen, fine-tuning) to isolate the contribution of self-supervised pre-training rather than claiming state-of-the-art against numbers obtained under incompatible settings.

Two deployment-relevant settings are considered:

Table 2. Operating condition (OC) settings in the Paderborn dataset.

Setting	Speed (rpm)	Torque (Nm)	Radial force (N)	Code
OC1	1500	0.7	1000	N15_M07_F10
OC2	900	0.7	1000	N09_M07_F10
OC3	1500	0.1	1000	N15_M01_F10
OC4	1500	0.7	400	N15_M07_F04

- **Label-scarce diagnosis:** Limited labeled target data is available; frozen-feature transfer and fine-tuning are evaluated.
- **Healthy-only adaptation:** Only healthy target data is assumed for training; reconstruction-error anomaly detection is evaluated.

All baselines use the same input representation (the standardized spectrogram pipeline from Section 3.3) so that performance differences primarily reflect representation quality rather than feature engineering. To further reduce confounds, downstream heads are kept small and consistent across methods (Section 4.5).

4.2. Benchmark dataset: Paderborn University and KAt-DataCenter

Transfer is evaluated on the Paderborn University and KAt-DataCenter bearing dataset introduced by Lessmeier et al. (Lessmeier et al., 2016). The dataset comprises measurements from 32 type-6203 bearing experiments and includes healthy bearings as well as bearings with damaged inner rings and damaged outer rings. The original benchmark also includes bearings with different damage origins (e.g., artificially induced and real damage states) and provides multiple sensing modalities. For each bearing, vibration and motor current signals are recorded synchronously at 64 kHz under four discrete operating conditions. Only the vibration channel is used, consistent with the single-modality scope of VibFM.

Operating conditions are summarized in Table 2, using the dataset's standard condition codes.

Each operating condition contains 20 vibration measurements of 4 seconds per bearing; each 4-second measurement is treated as one sample, a deliberately conservative choice that avoids additional correlation from overlapping sub-windowing and simplifies leakage accounting.

All downstream samples are converted into spectrogram inputs using the same standardized pipeline as in Section 3.3: log-magnitude STFT, resize to 128×128, clip to $[-100, 0]$, and scale to $[0, 1]$. The same acquisition-conditioning vector $\mathbf{c}$ (Equation 6) is attached so the encoder receives consistent physical-scale metadata, and corpus statistics for spectrogram normalization are computed on the pre-training corpus only, never refit on the downstream test set.

4.3. Downstream tasks

Two downstream PHM tasks are evaluated on the held-out Paderborn benchmark. They are included to test two distinct deployment modes rather than to provide an exhaustive benchmark suite. The classification task evaluates label-scarce supervised transfer, while the anomaly-detection task evaluates healthy-only adaptation, which is common when fault examples are unavailable during commissioning or early operation.

Task 1: Label-scarce three-class diagnosis. The established subset and label mapping in the original benchmark are used to define a three-class classification problem:

$$y \in \{\text{Healthy},\ \text{Outer ring fault},\ \text{Inner ring fault}\}. \tag{14}$$

A lightweight classifier head is attached to the pooled embedding $\mathbf{z}_{\text{health}}$ produced by the encoder (Section 3.4).

Transfer modes are compared:

- **Scratch:** same encoder architecture as VibFM, initialized randomly and trained supervised on the training bearings.
- **Frozen:** pretrained VibFM encoder is frozen; only the classifier head is trained.
- **Fine-tuning:** encoder and head are trained end-to-end, using a smaller learning rate for the encoder than the head.

This isolates the contribution of self-supervised pre-training from architectural capacity.

Task 2: Healthy-only adaptation for anomaly detection. A healthy-only adaptation scenario is also evaluated. The pretrained encoder is reused and a decoder is adapted on healthy target data so that normal time–frequency structure reconstructs well. At inference time, reconstruction error is used as an anomaly score; samples that reconstruct poorly are flagged as anomalous (Chandola, Banerjee, & Kumar, 2009).

Concretely, for a given spectrogram X and a random mask $\mathcal{M}$, the masked-patch reconstruction error is defined as

$$s(X; \mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|x_i - \hat{x}_i\|_1 , \tag{15}$$

where x_i and $\hat{x}_i$ denote target and reconstructed pixel vectors of patch i. To reduce variance due to a particular mask draw, the score is averaged over K random masks:

$$s_K(X) = \frac{1}{K} \sum_{k=1}^{K} s(X; \mathcal{M}_k). \tag{16}$$

Table 3. Bearing codes used for the three-class diagnostic task, following the bearing-level protocol (real-damage subset).

Class	Bearing codes
Healthy	K001, K002, K003, K004, K005
Outer ring fault	KA04, KA15, KA16, KA22, KA30
Inner ring fault	KI04, KI14, KI16, KI18, KI21

Anomaly-score behavior is analyzed both by damage state (healthy vs. damaged) and by operating condition (Table 2), matching Figures 5a–5b in the Results section.

4.4. Leakage-resistant splits and evaluation metrics

To prevent unit leakage, we split the data at the bearing-code level (not at the individual measurement level). We follow the bearing-level protocol introduced for the Paderborn (KAt) benchmark in (Lessmeier et al., 2016): for each class, five bearing codes are defined and all $\binom{5}{3} = 10$ disjoint train/test combinations are evaluated (three bearings for training and two bearings for testing per class). Thus, no bearing code appears in both train and test.

Importantly, the three-class diagnostic task in this work uses the healthy bearings and the bearings with real damage produced by accelerated lifetime tests (i.e., physically generated fatigue and related damage mechanisms), as listed in the benchmark paper. The broader dataset also contains artificially induced defects, but these are not part of the main split protocol reported here. Each bearing code represents a fixed health state measured repeatedly under four operating settings, rather than a temporal run-to-failure trajectory.

Because every bearing code is measured under the same set of operating settings with the same number of repetitions, each split is balanced by construction:

- **Training set:** 9 bearings (3 per class),
- **Test set:** 6 bearings (2 per class).

Within each split, a fraction 0.2 of the training bearings' measurements is held out for validation (early stopping and hyperparameter selection). Validation samples are drawn exclusively from training bearings (optionally stratified by class and operating condition), and test bearings remain strictly unseen until final evaluation.

For diagnosis, we report **accuracy** and **macro F1**, averaged over the 10 bearing-level splits. Macro F1 is computed as the unweighted mean of per-class F1-scores:

$$\mathrm{MacroF1} = \frac{1}{3}\sum_{c=1}^{3}\frac{2\,\mathrm{Prec}_c\,\mathrm{Rec}_c}{\mathrm{Prec}_c + \mathrm{Rec}_c}. \tag{17}$$

For anomaly detection, anomaly scores are summarized with receiver operating characteristic area under the curve (ROC-AUC) for (i) healthy vs. damaged samples and (ii) per operating condition. In addition, we visualize grouped score distributions to highlight operating-condition effects and support interpretability of the anomaly-score behavior.

4.5. Downstream model heads and optimization protocol

To reduce confounds, downstream heads are kept small and consistent across methods.

The diagnosis head is an MLP with 2 hidden layers of width 256 with dropout 0.2, followed by a linear layer to 3 logits. GELU activations are used.

AdamW with weight decay 0.01 and batch size 8 is used. Training length is set per transfer mode:

- **Scratch:** trained for up to 100 epochs with learning rate 1×10^{-4} (early stopping).
- **Frozen:** encoder frozen; head trained for 5 epochs with learning rate 1×10^{-4}.
- **Fine-tuning:** trained for 5 epochs, encoder learning rate 1×10^{-5} and head learning rate 1×10^{-4}.

Early stopping is applied based on validation macro F1, so 100 epochs is an upper bound chosen to give scratch training sufficient opportunity to converge.

For anomaly detection, the encoder is frozen and a decoder is trained using only healthy training samples by minimizing the masked reconstruction loss (Equation 11). Test samples are then scored via Equation 15. When a hard decision is needed, a single global threshold is set to the 95th percentile of s_K on the healthy validation subset.

5. RESULTS

This section evaluates whether VibFM provides *transferable* representations under a leakage-resistant protocol (bearing-level splits, Section 4.4). Two adaptation modes that matter in practice are considered: (i) **frozen transfer**, which reduces target-adaptation cost and label requirements by training only a small head, and (ii) **fine-tuning**, which adapts the representation end-to-end.

5.1. Fault diagnosis under bearing-level generalization

Table 4 summarizes three-class diagnosis performance (Healthy vs. Outer-ring fault vs. Inner-ring fault) averaged over all $\binom{5}{3} = 10$ bearing-level splits. Because each split uses disjoint bearing codes for training and testing, the test results cannot be explained by segment-level memorization; they require generalization to previously unseen physical units.

For reference, the original benchmark paper reports up to 98.3% mean accuracy for three-class classification when training feature-based classical ML directly on the real-damage subset using vibration signals and the same bearing-

Table 4. Three-class bearing diagnosis on the Paderborn benchmark (mean over 10 bearing-level train/test combinations).

Method	Accuracy (%)	Macro F1
Scratch (supervised)	48.91	0.481
VibFM (frozen) + head	75.12	0.735
VibFM (fine-tuned)	**85.52**	**0.854**

level split protocol. Our fine-tuned VibFM reaches 85.52% (within $\sim$13 percentage points) after only 5 epochs of end-to-end adaptation following multi-source self-supervised pre-training, and frozen features reach 75.12% with only a lightweight head (Table 4). This comparison highlights the intended trade-off: we do not optimize solely for one benchmark, but aim for a reusable encoder that transfers with minimal task-specific training.

Two patterns stand out from these results and are central to the foundation-model claim:

- **Frozen transfer outperforms training from scratch:** Even with the encoder fixed, a small head achieves higher accuracy, suggesting that pre-training already organizes the embedding around health-relevant time–frequency structure that is readily separable.
- **Fine-tuning further sharpens the representation:** Updating the encoder end-to-end adapts these features to the target labels, typically tightening class clusters and reducing recurring confusions (especially between inner- and outer-ring damage).

To relate diagnostic performance to representation quality, we project the health embedding $\mathbf{z}_{\text{health}}$ to two dimensions using t-SNE with fixed hyperparameters (van der Maaten & Hinton, 2008). Figure 4 contrasts the latent geometry produced by a frozen pretrained encoder (left) with the geometry after end-to-end fine-tuning for diagnosis (right). The frozen encoder already yields meaningful local neighborhoods, with emerging grouping of healthy, outer-ring, and inner-ring conditions, suggesting that pre-training captures health-relevant structure even without Paderborn labels. After fine-tuning, these neighborhoods consolidate into more label-aligned clusters and reduce mixed regions, indicating that supervision primarily sharpens and reorients an already structured embedding to make the downstream decision boundary simpler.

Interpretation is qualitative: t-SNE preserves local neighborhoods but can distort global geometry, so apparent inter-cluster distances and relative cluster sizes should not be over-interpreted. We therefore use the visualization as a sanity check for separability, complemented by quantitative metrics rather than as a causal explanation of which physical features drive the separation.

- **Frozen:** local neighborhoods already show partial grouping by health state, suggesting that pre-training captures transferable structure without access to Paderborn labels.
- **Fine-tuned:** clusters become more compact and more label-aligned, reducing mixed regions and simplifying the downstream decision boundary.
- **Caution:** t-SNE is descriptive rather than mechanistic; it reflects local neighborhood structure but does not identify which physical cues drive separation or preserve global geometry.

Table 5. Diagnosis performance by operating condition.

Method	OC1	OC2	OC3	OC4
Scratch (Acc in %)	52.50	42.08	47.52	53.52
VibFM frozen (Acc in %)	79.58	60.23	80.41	80.27
VibFM fine-tuned (Acc in %)	**86.25**	**93.75**	**76.22**	**85.83**

Accuracy alone can hide systematic failure modes. Bearing faults often share overlapping spectral signatures (e.g., modulation sidebands and broadband impulsive activity), so the remaining inner-vs.-outer confusions should be analyzed when extending VibFM to finer-grained fault taxonomies.

A model that generalizes across unseen bearings can still fail under operating shifts. Table 5 breaks diagnosis performance down by operating point (Table 2). This analysis is important because speed/load changes can shift rotation-related spectral spacing, sideband location, and resonance excitation.

The key diagnostic is *variance across conditions*. If the fine-tuned model improves the average but still underperforms under a specific operating point, that indicates the representation is not fully invariant and suggests: (i) adding operating-point conditioning during fine-tuning, (ii) explicitly balancing per-condition batches, or (iii) using condition-specific calibration (especially relevant for anomaly detection).

5.2. Anomaly detection via transferred reconstruction

Next, the anomaly-detection experiment evaluates whether the pretrained encoder can support a one-class (healthy-only) adaptation for anomaly detection. The desired behavior is that healthy samples reconstruct well (low error / low score), damaged samples reconstruct poorly (high error / high score), and a single threshold does not drift excessively across operating points.

Figure 5a groups anomaly scores by health state. In this experiment, the decision threshold is set to the 95th percentile of scores on *healthy validation data* only, which matches realistic deployment where damaged examples may be unavailable or incomplete. For offline evaluation, damaged labels are used only to compute ROC-AUC and are not used for training or threshold selection.

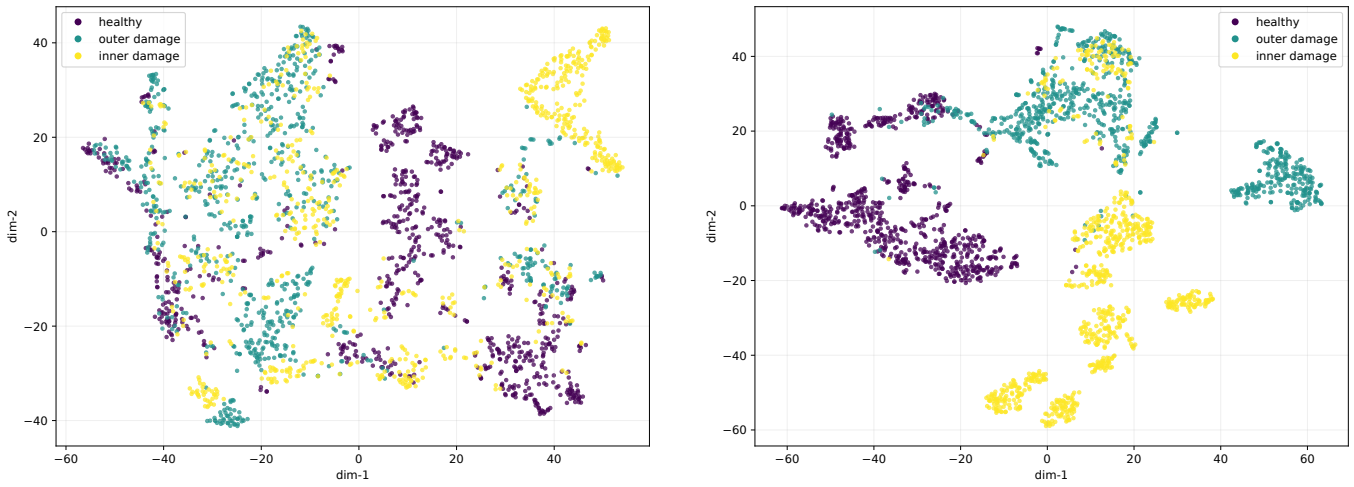


Figure 4. t-SNE of $\mathbf{z}_{\text{health}}$ on the held-out Paderborn benchmark. Left: frozen encoder shows partial class grouping. Right: after fine-tuning, clusters become more compact and label-aligned.

The two score views should be read together: the health-state grouping tests whether damage increases reconstruction error overall, whereas the operating-condition grouping tests whether that separation remains stable enough for threshold calibration.

In addition to these grouped distributions, ROC-AUC summarizes healthy vs. damaged separation. Here, $\text{ROC-AUC}_{\text{overall}} = 0.8928$ pools all samples across operating conditions, while $\text{ROC-AUC}_{\text{macro-OC}} = 0.8653$ is the unweighted mean of ROC-AUC computed separately within each operating condition.

Figure 5b breaks anomaly scores down by operating condition. This is the critical stress-test for reconstruction-based anomaly detection: even if damaged samples score higher on average, a threshold can become unreliable if the healthy score distribution shifts with speed/load.

From a deployment perspective, there are two acceptable outcomes when using reconstruction-based anomaly detection with a transferred VibFM encoder (even though thresholding is not a property of the encoder itself):

- **Single global threshold works:** Healthy scores remain consistently separated from damaged scores across conditions, enabling one calibration step.
- **Condition-aware thresholds are needed:** If the healthy distribution shifts systematically, then calibration should be performed per operating point (or by conditioning the decoder/score on operating variables).

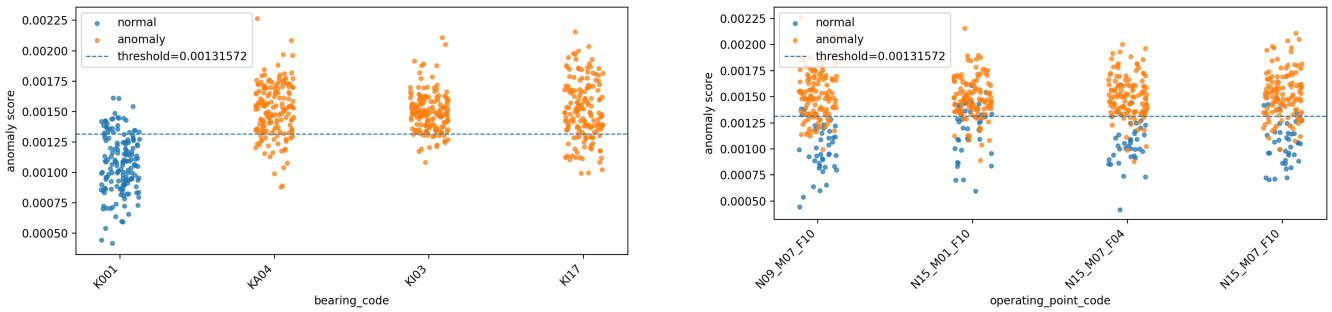
The second outcome is not a failure of the approach; it reflects the fact that operating regime is a dominant factor in vibration spectra. The practical requirement is to avoid false alarms due to non-damage-related regime changes.

6. DISCUSSION

Masked spectrogram modeling encourages VibFM to recover missing patches from broader time–frequency context, including narrowband ridges, modulation sidebands, resonance bands, and impulsive transients (Randall, 2011). The spectrogram is therefore not used merely as an image-like input, but as a common representation on which a reconstruction task can be posed across datasets with different sensors, sampling rates, and operating conditions. This makes the objective useful for pre-training, while leaving conventional diagnostics in their natural role: envelope analysis, Fourier spectra, order tracking, and cyclic spectral methods remain preferable when the goal is interpretable fault-frequency confirmation under known geometry and speed.

The Paderborn results suggest that pre-training provides useful transfer to unseen physical units: frozen features reduce target-label requirements, while fine-tuning sharpens class structure when labeled target data are available (Table 4, Figure 4). This is important because segment-level splits can reward unit identity, whereas bearing-level splits test whether representations transfer across physical units. The condition-wise diagnosis and anomaly-score plots also show why operating regime must be treated as a deployment variable rather than a nuisance detail. If performance or healthy reconstruction baselines drift with speed/load, then balanced adaptation, operating-variable conditioning, or regime-aware thresholds are needed.

From a PHM perspective, the value is the possibility to *train once, adapt many times*: a reusable encoder can support cold-start monitoring, label-efficient transfer, screening of large fleets, and feature extraction before reliable fault labels exist. In an industrial workflow, VibFM should be viewed as a triage and representation layer rather than a standalone di-



(a) (b)

Figure 5. Anomaly scores $s_K(X)$ from healthy-only reconstruction on the held-out Paderborn benchmark. (a) Scores grouped by health state (healthy vs. damaged), with the threshold set to the 95th percentile of healthy validation scores. (b) Scores grouped by operating condition (OC1–OC4; see Table 2).

agnostic authority. Learned scores should be combined with conventional diagnostics, maintenance history, threshold validation, and drift monitoring after sensor, load, speed, or structural changes. Beyond the present three-class diagnosis and healthy-only anomaly detection, the same representation could support severity estimation, health-index construction, remaining useful life modeling, and transfer to gearboxes, motors, pumps, or drivetrains after dedicated validation. The attention maps remain useful qualitative audits of where information is pooled, but they should not be interpreted as causal explanations without perturbation tests.

7. LIMITATIONS

VibFM is evaluated downstream on one held-out bearing benchmark with a coarse healthy/inner/outer taxonomy; broader validation should include additional components, sensing chains, variable-speed regimes, finer severity labels, and prognostics targets. The log-magnitude STFT representation enables efficient multi-source tokenization, but it discards phase, compresses absolute amplitude, and trades physical resolution for a fixed grid. The chosen time-frequency representation is therefore not optimized for resolving every bearing characteristic frequency in the classical spectral sense, and multi-resolution or hybrid waveform/spectrogram tokenization may be preferable for severity or prognostics.

Multi-dataset training reduces dependence on any single test rig, but it does not prove invariance to frequency response functions, sensor mounting, or structural resonances. The current conditioning vector captures acquisition scale but not speed, load, or torque, so regime changes can still shift spectra and reconstruction baselines. Reconstruction error should therefore be interpreted as an out-of-distribution score relative to the healthy adaptation data, not as a fault classifier. Finally, reconstructions and pooling-attention maps are qualitative audits rather than causal explanations; perturbation-based tests are needed to establish which regions drive predictions.

8. CONCLUSION

This paper introduced VibFM, a self-supervised Transformer encoder for vibration-based PHM. By standardizing heterogeneous recordings into conditioned log-magnitude spectrograms and training with masked reconstruction, the model learns a reusable representation that can be transferred to downstream diagnosis and anomaly-detection tasks without training a separate encoder from scratch for every new asset.

On the held-out Paderborn benchmark, evaluated with bearing-level splits, VibFM improves over scratch training as both a frozen feature extractor and a fine-tuned model, while healthy-only reconstruction supports anomaly scoring. These results support the central premise that multi-source self-supervised pre-training can reduce dependence on scarce fault labels and provide a common encoder for new assets.

The practical implication is a shift from repeatedly designing task-specific models toward adapting and calibrating a shared vibration representation. Future work should test broader components and operating regimes, quantify frequency-response and sensor-location robustness, and extend the representation to severity and remaining-useful-life tasks.

BIOGRAPHIES

Giuseppe Mannone studied Autonomous Systems at the University of Stuttgart in Germany and received his M.Sc. in October 2024. Since 2024, he has been working as a research assistant in the field of reliability engineering at the Institute of Machine Components. His research focuses on applying artificial intelligence to reliability engineering.

Paula Fischer studied Mechanical Engineering at the University of Stuttgart in Germany and received her M.Sc. in 2023. Since September 2023, she has been a researcher in the Reliability Engineering Department at the Institute of Machine Components, University of Stuttgart, and is pursuing her PhD studies.

Martin Dazer works as Head of the Reliability & Driveline Department at the Institute of Machine Components. He received his doctoral degree in reliability engineering from the University of Stuttgart, Germany, in 2019 and completed his habilitation in 2025. His fields of interest include reliability testing, lifetime data analysis, and statistical test planning.

CODE AVAILABILITY

The trained model can be requested from the authors. The code for the full project is available at `https://github.com/SFZ-Uni-Stuttgart/VibFM`.

ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) under project number 532373244.

REFERENCES

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 15:1–15:58. doi: 10.1145/1541880.1541882

Chen, S., Liu, Z., He, X., Zou, D., & Zhou, D. (2024). Multi-mode fault diagnosis datasets of gearbox under variable working conditions. *Data in Brief*, *54*, 110453. doi: 10.1016/j.dib.2024.110453

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., … Wei, F. (2023). BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 5178–5193).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt)* (pp. 4171–4186). doi: 10.18653/v1/N19-1423

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.*

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio spectrogram transformer. In *Proceedings of interspeech* (pp. 571–575). doi: 10.21437/Interspeech.2021-698

Gong, Y., Lai, C.-I., Chung, Y.-A., & Glass, J. (2022). SSAST: Self-supervised audio spectrogram transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 10699–10709. doi: 10.1609/aaai.v36i10.21315

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 16000–16009). doi: 10.1109/CVPR52688.2022.01553

Hendriks, J., Dumond, P., & Knox, D. A. (2022). Towards better benchmarking using the cwru bearing fault dataset. *Mechanical Systems and Signal Processing*, *169*, 108732. doi: 10.1016/j.ymssp.2021.108732

Huang, H., & Baddour, N. (2019). *Bearing vibration data under time-varying rotational speed conditions.* doi: 10.17632/v43hmbwxpm.2

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., … Feichtenhofer, C. (2022). Masked autoencoders that listen. In *Advances in neural information processing systems.*

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt)* (pp. 3543–3556). doi: 10.18653/v1/N19-1357

Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, *20*(7), 1483–1510. doi: 10.1016/j.ymssp.2005.09.012

Lai, Z., Yang, C., Lan, S., Wang, L., Shen, W., & Zhu, L. (2024). BearingFM: Towards a foundation model for bearing fault diagnosis by domain knowledge and contrastive learning. *International Journal of Production Economics*, *275*, 109319. doi: 10.1016/j.ijpe.2024.109319

Lee, S., Kim, T., & Kim, T. (2024). *Multi-domain vibration dataset with various bearing types under compound machine fault scenarios: subset 1 (deep groove ball bearing).* doi: 10.17632/53vtnjy6c6.1

Lei, Y., Han, T., Wang, B., Li, N., Yan, T., & Yang, J. (2019). XJTU-SY rolling element bearing accelerated life test datasets: A tutorial. *Journal of Mechanical Engineering*. doi: 10.3901/JME.2019.16.001

Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. *PHM Society European Conference*, *3*(1). doi: 10.36001/phme.2016.v3i1.1577

Li, Y.-F., Wang, H., & Sun, M. (2024). ChatGPT-like large-scale foundation models for prognostics and health management: A survey and roadmaps. *Reliability Engineering & System Safety*, *243*, 109850. doi: 10.1016/j.ress.2023.109850

Liefstingh, M., Taal, C., Echeverri Restrepo, S., & Azarfar, A. (2021). Interpretation of deep learning models in bearing fault diagnosis. In *Annual conference of the*

phm society (Vol. 13). doi: 10.36001/phmconf.2021.v13i1.3047

Luleå University of Technology. (2024). *Vibration data from a gearbox output shaft bearing in a 2.5 MW wind turbine.* Retrieved from https://researchdata.se/en/catalogue/dataset/2024-248 (46-month field degradation dataset)

Lundström, A., & O'Nils, M. (2023). Factory-based vibration data for bearing-fault detection. *Data*, *8*(7), 115. doi: 10.3390/data8070115

NASA Prognostics Data Repository, & IMS Center, University of Cincinnati. (2007). *Ims bearings.* Retrieved from https://catalog.data.gov/dataset/ims-bearings

National Renewable Energy Laboratory. (2014). *Wind turbine gearbox condition monitoring: Vibration analysis benchmarking datasets.* doi: 10.25984/1844194

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In *Proceedings of the ieee international conference on prognostics and health management (phm)* (pp. 1–8). (IEEE PHM 2012 Data Challenge)

Nguyen, T. (2023). *Hust bearing.* doi: 10.17632/cbv7jyx4p9.1

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. doi: 10.1109/TKDE.2009.191

PHM Society, & NASA DASHlink. (2009). *Gearbox fault detection dataset, PHM data challenge 2009.* Retrieved from https://phmsociety.org/public-data-sets/

Randall, R. B. (2011). *Vibration-based condition monitoring: Industrial, aerospace and automotive applications*. John Wiley & Sons.

Schnur, C., Schneider, M., Zhang, Y., Berger, K., Schütze, S., Zou, J., ... Heimes, H. (2025). *A machine learning dataset of artificial inner ring damages on cylindrical roller bearings measured under varying cross-influences.* doi: 10.5281/zenodo.11108503

Sehri, M., & Dumond, P. (2023). *University of ottawa ball-bearing vibration and acoustic fault data under constant load and speed conditions (uods-vafdc).* doi: 10.17632/y2px5tg92h.1

Serrano, S., & Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics (acl)* (pp. 2931–2941). doi: 10.18653/v1/P19-1282

Southeast University (SEU). (2024). *Seu gearbox dataset.* doi: 10.57702/wwr1jf6s

Tandon, N., & Choudhury, A. (1999). A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings. *Tribology International*, *32*(8), 469–480. doi: 10.1016/S0301-679X(99)00077-8

University of New South Wales. (2020). *Bearing run-to-failure datasets of UNSW.* (Mendeley Data) doi: 10.17632/h4df4mgrfb.3

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Wang, H., Liu, Z., Ge, Y., & Peng, D. (2022). Self-supervised signal representation learning for machinery fault diagnosis under limited annotation data. *Knowledge-Based Systems*, *239*, 107978. doi: 10.1016/j.knosys.2021.107978

Xiao, Y., Shao, H., Yan, S., Wang, J., Peng, Y., & Liu, B. (2025). Domain generalization for rotating machinery fault diagnosis: A survey. *Advanced Engineering Informatics*, *64*, 103063. doi: 10.1016/j.aei.2024.103063

Zhao, C., Zio, E., & Shen, W. (2024). Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study. *Reliability Engineering & System Safety*, *245*, 109964. (HUSTGearbox dataset repository: https://github.com/CHAOZHAO-1/HUSTgearbox-dataset) doi: 10.1016/j.ress.2024.109964

Zhao, R., Yan, R., Chen, Z., Mao, K., Liu, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, *115*, 213–237. doi: 10.1016/j.ymssp.2018.05.050

APPENDIX

This appendix lists the open datasets used to build the self-supervised pre-training corpus and summarizes their operating regimes and sampling rates. The downstream diagnostic benchmark in Section 4.2 is excluded from pre-training.

Table 6. Open datasets used to construct the self-supervised pre-training corpus. (D = diagnosis-oriented dataset, P = prognostics or run-to-failure style dataset).

Dataset	Component	Task	Fault origin	Operating regime	Modalities used	f_s	Notes
NASA IMS Bearing (NASA Prognostics Data Repository & IMS Center, 2007)	Bearings	P	Natural (R2F)	Steady	Vibration	∼20 kHz	Long runs
XJTU-SY Bearing (Lei et al., 2019)	Bearings	D/P	Natural (R2F)	3 steady cond.	Vibration	25.6 kHz	Multi-bearings
Ottawa (Const. Speed/Load) (Sehri & Dumond, 2023)	Bearings	P	Natural (R2F)	Steady	Vib., Acoustic, Speed, Load, Temp.	42 kHz	Accel+mic
Ottawa (Variable Speed) (Huang & Baddour, 2019)	Bearings	D	Seeded	Time-varying speed	Vibration, Speed	200 kHz	Variable speed
HUST Bearing (Nguyen, 2023)	Bearings	D	Seeded	Multi-load steady	Vibration	51.2 kHz	5 bearing types
SCA Bearing (Lundström & O'Nils, 2023)	Bearings	D	Field	Variable (field)	Vibration, Speed	var.	Multi-month traces
U. of Seoul Multi-Domain (Lee et al., 2024)	Bearings	D	Seeded	Multi-speed steady	Vibration	8/16 kHz	Compound faults
Saarland Univ. Bearing (Schnur et al., 2025)	Bearings	D	Seeded	Multi-load steady	Vib. (triax.)	20 kHz	LOGOCV-ready
FEMTO-ST / PRONOSTIA (Nectoux et al., 2012)	Bearings	P	Natural (R2F)	3 steady cond.	Vibration, Temperature	25.6 kHz	IEEE PHM 2012
UNSW Bearing (University of New South Wales, 2020)	Bearings	D/P	Natural	4 speeds	Vibration, Speed	var.	4 run-to-failure tests
PHM 2009 Gearbox (PHM Society & NASA DASHlink, 2009)	Gearbox	D	Seeded	Multi-speed/load	Vibration, Tacho	66.7 kHz	Compound faults
SEU Gearbox (Southeast University (SEU), 2024)	Gearbox	D	Seeded	Condition shift	Vib. (triax.), Torque	5.12 kHz	Gearset and Bearingset
NREL Wind Gearbox (National Renewable Energy Laboratory, 2014)	Gearbox	D	Field	Multi-load steady	Vibration, Speed, Torque	40 kHz	Healthy vs. damaged
HUST Gearbox (Zhao et al., 2024)	Gearbox	D	Seeded	Multi-load + variable	Vibration, Speed	25.6 kHz	Single-stage gearbox
MCC5-THU Gearbox (Chen et al., 2024)	Gearbox	D	Artificial	24 steady + 48 transient	Vib. (triax.), Speed, Torque	12.8 kHz	Variable conditions
Luleå Wind Turbine (Luleå University of Technology, 2024)	Gearbox	P	Natural	Variable (field)	Axial Vib., Speed, SCADA	var.	46-month degradation