IndusDiff: A Latent Diffusion Framework for Industrial Audio Generation for Data-Scarce PHM

Muhammad Areeb [1], Jida Huang[2], Sudhir Rajagopalan[3], Merrill Edmonds[4], Miao He[5], and David He[6]

[1,2,6] *University of Illinois Chicago, Chicago, IL, 60607, USA*
mareeb2@uic.edu
jida@uic.edu
davidhe@uic.edu

[3,4,5] *Siemens Corporation, Princeton, NJ, 08540, USA*
sudhir.rajagopalan@siemens.com
merrill.edmonds@siemens.com
miao.he@siemens.com

ABSTRACT

Prognostics and Health Management (PHM) systems increasingly rely on data-driven models for fault diagnosis and anomaly detection, yet their effectiveness is fundamentally constrained by the scarcity and imbalance of labeled fault-condition data. This challenge is particularly acute in industrial audio monitoring, where failure events are rare, hazardous to induce, and highly variable across operating conditions. While recent advances in generative modeling offer a promising pathway for data augmentation, existing audio generation methods either suffer from high computational cost in the waveform-domain diffusion or fail to preserve phase-sensitive characteristics critical for capturing transient fault signatures.

To address these challenges, this paper proposes IndusDiff, a phase-aware latent diffusion framework for high-fidelity industrial audio generation tailored for PHM applications. The proposed approach combines a phase-aware convolutional autoencoder with a latent diffusion model to enable efficient and physically meaningful audio synthesis. The autoencoder is trained using a multi-resolution spectral loss that jointly enforces waveform fidelity, spectral consistency, and temporal phase coherence, thereby preserving diagnostically relevant features such as transients, harmonics, and high-frequency components in the latent representation. Diffusion is then performed in this compressed latent space, significantly reducing computational complexity while maintaining generation quality.

Unlike conventional latent diffusion methods that primarily focus on perceptual realism, the proposed framework explicitly addresses the unique characteristics of industrial audio signals, including non-stationarity, multi-scale temporal dynamics, and phase-sensitive fault signatures. By preserving both magnitude and phase information, the model generates synthetic audio that is not only perceptually realistic but also structurally consistent with real machine signals, making it suitable for downstream PHM tasks.

The framework is evaluated on two complementary datasets: a publicly available industrial motor dataset and an in-house dataset capturing multi-stage assembly operations. Generation quality is assessed using the Frechet Audio Distance (FAD) as the primary quantitative metric, along with waveform and spectrogram analyses for qualitative validation. Experimental results demonstrate that IndusDiff produces statistically consistent, high-fidelity industrial audio while achieving significant improvements in computational efficiency, generating 48 kHz audio samples of several seconds in duration in just seconds on modern GPU hardware.

1. INTRODUCTION

Unplanned failure of industrial assets such as motors, pumps, fans, valves, conveyors, and other rotating or flow-driven machinery can impose severe operational and financial costs on assembly lines. Prognostics and health management (PHM) has emerged as the principal discipline for mitigating this risk by detecting degradation early and informing maintenance decisions before catastrophic failure occurs (Lei et al., 2018; Zhao et al., 2019). In recent years, deep learning has substantially advanced PHM capabilities, supplanting handcrafted feature pipelines with end-to-end representations that

excel at fault classification, remaining useful life (RUL) estimation, and anomaly detection (Li, Ding, & Sun, 2018; Ding, Ma, Ma, Wang, & Lu, 2023). Despite this progress, a persistent gap remains between what these models can achieve in controlled research settings and what is reliably deployable in real industrial environments.

Acoustic sensing has attracted particular interest for PHM because microphones are inexpensive, non-contact, and straightforward to retrofit across diverse machine types and production lines (Purohit et al., 2019; Koizumi et al., 2020). Industrial audio, however, is considerably harder to model than the curated speech or music signals that dominate the audio machine learning literature. Machine sounds combine tonal periodic components (e.g., shaft rotation harmonics) with broadband turbulence, friction noise, and short-duration transient events (e.g., impacts, cavitation, leakage), all of which occur simultaneously over a wide frequency range while being corrupted by environmental reverberation, cross-machine interference, and load-dependent variability (Purohit et al., 2019). Critically, both spectral structure and temporal phase behavior can carry diagnostically relevant information. Subtle transients and high-frequency cues, often under-emphasized by human-centric feature representations, may encode the earliest signatures of mechanical degradation (Koizumi et al., 2020).

A fundamental bottleneck for PHM model development is the chronic scarcity of labeled fault-condition recordings as depicted in Figure 1. Acquiring such data requires deliberately inducing failures in operational equipment which is a costly, hazardous, and rarely sanctioned procedure in production environments. As a result, industrial audio datasets are heavily skewed toward normal operating conditions, with anomalous samples that are genuinely difficult to collect at scale. Public benchmark efforts explicitly acknowledge this deficit: several challenge formulations treat anomalous examples as limited or entirely unavailable during training, relying instead on unsupervised or semi-supervised detection (Koizumi et al., 2020). This scarcity not only complicates supervised learning but also undermines model robustness to the operating-condition variability inevitable in real deployments, where machine configurations, load profiles, and environmental contexts differ from one installation to the next.

Generative modeling provides a principled pathway to address this gap by synthesizing plausible machine sounds across diverse conditions. Early neural audio synthesis was dominated by autoregressive waveform models such as WaveNet (van den Oord et al., 2016), which achieve strong perceptual fidelity through sample-by-sample prediction but scale poorly to the long signal windows required for industrial monitoring. GAN-based generators improved sampling speed substantially through parallel waveform synthesis (Kumar et al., 2019; J. Kong, Kim, & Bae, 2020), yet adversarial training is sensitive to hyperparameter choices and prone to mode collapse, which is a particularly serious drawback for PHM, where failing to reproduce low-probability but high-impact fault signatures is unacceptable. Variational autoencoders offer training stability but often sacrifice the perceptual sharpness of transient events (Kingma & Welling, 2014), precisely the cues most relevant to early fault detection.

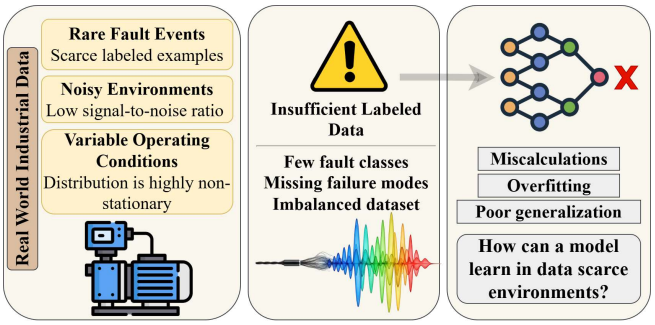

Figure 1. Industrial environments exhibit rare fault events, high noise, and non-stationary conditions, leading to limited and imbalanced labeled data. This constrains diagnostic models, resulting in mis-classification, overfitting, and poor generalization to unseen faults. The figure highlights a key challenge: how can models learn failure patterns that are underrepresented or absent in training data.

Denoising diffusion probabilistic models (DDPMs) have recently emerged as a compelling alternative, providing stable training through a denoising score-matching objective and achieving state-of-the-art generation quality across both images and audio (Ho, Jain, & Abbeel, 2020; Y. Song et al., 2021). In the audio domain, DiffWave (Z. Kong, Ping, Huang, Zhao, & Catanzaro, 2021) and WaveGrad (Chen et al., 2021) demonstrated that iterative diffusion over raw waveforms can match or exceed autoregressive quality while permitting a configurable speed-quality trade-off. A key practical limitation, however, is computational cost: generating high-quality samples requires many sequential denoising evaluations, and the high dimensionality of raw waveforms at industrial sampling rates (e.g., 48 kHz) makes waveform-domain diffusion prohibitively slow for large-scale augmentation workloads. While accelerated samplers (J. Song, Meng, & Ermon, 2021) reduce the required step count substantially, the per-step cost of operating directly in waveform space remains a bottleneck that motivates representational rather than purely algorithmic solutions.

The literature on speech/audio diffusion is mature, but it assumes abundant data and often neglects phase detail. GAN/vocoder works focus on speech conditioning. In contrast, industrial audio requires models that generalize from few examples and respect phase-sensitive anomalies. Additionally, diffusion literature has mostly used waveform-domain and paid less attention to latent compression for audio

(with notable exceptions like AudioLDM (Liu et al., 2023)). To address these limitations we propose a **phase-aware latent diffusion framework for industrial audio generation** that resolves this tension by performing diffusion in a compressed latent space rather than in the raw waveform domain. The central insight is that the computational burden of waveform diffusion can be substantially reduced by first learning a compact, information-preserving representation of the audio signal and then applying diffusion exclusively within that lower-dimensional space, an approach validated by the success of Latent Diffusion Models (LDMs) in high-resolution image synthesis (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022) and general audio generation (Liu et al., 2023). For industrial audio, however, a naive latent space is insufficient: standard autoencoder training objectives that minimize magnitude-spectrum reconstruction discard instantaneous phase information, leading to waveforms with perceptually plausible spectra but corrupted temporal fine structure. This is especially damaging for impulsive fault signatures, where waveform phase coherence directly determines whether a transient event is faithfully rendered.

To address this, we introduce a *phase-aware autoencoder* trained with a multi-resolution spectral loss that jointly supervises magnitude and phase reconstruction across multiple frequency resolutions. This encoder compresses industrial audio into a compact continuous latent code while preserving both spectral content and temporal phase structure. A latent diffusion model then learns to synthesize novel latent codes representative of real machine operating states, which are decoded back to full-resolution waveforms. We validate the framework on two distinct industrial audio datasets, assessing generation quality through Fréchet Audio Distance (FAD) (Kilgour, Zuluaga, Roblek, & Sharifi, 2019) as the primary metric, complemented by qualitative waveform and spectrogram analysis. By confining diffusion to latent space, our system generates 5.0-second audio at 48 kHz in approximately 8.1 seconds on NVIDIA GeForce RTX 5070, an order-of-magnitude reduction in sampling time compared to waveform-domain diffusion baselines while maintaining perceptual and temporal closeness to real signals.

In summary, the main contributions of this work are:

- A **phase-aware autoencoder** with a multi-resolution spectral loss that compresses industrial audio into a compact latent space while faithfully preserving both spectral and temporal phase characteristics essential for PHM-relevant fault signatures.
- A **stable latent diffusion model** for industrial audio synthesis that generates high-fidelity, perceptually realistic machine sounds at 48 kHz with a dramatically reduced sampling cost compared to typical waveform-domain diffusion approaches.
- A **cross-dataset evaluation** on two industrial audio benchmarks using FAD alongside waveform and spectrogram analysis, demonstrating that the proposed framework generalizes across different machine types and operating conditions requiring only dataset-appropriate conditioning and minor capacity adjustments.

The rest of the paper is arranged as follows: Section 2 elaborates the advancements in the field of acoustic signal generation. Section 3 explains the latent diffusion approach we implemented for generating acoustic signals. Section 4 talks about the evaluation methodology for assessing the quality of generation. Next, we describe the datasets and the results from our generation experiment and provide some discussions in Section 5. Lastly, we present our conclusion and highlight future directions of the work in Section 6.

2. Related Work

This section provides a brief review of advancements made in the field of acoustic signal generation. A diagrammatic view of the progression is also presented leading up to our contribution, as can be seen in Figure 2.

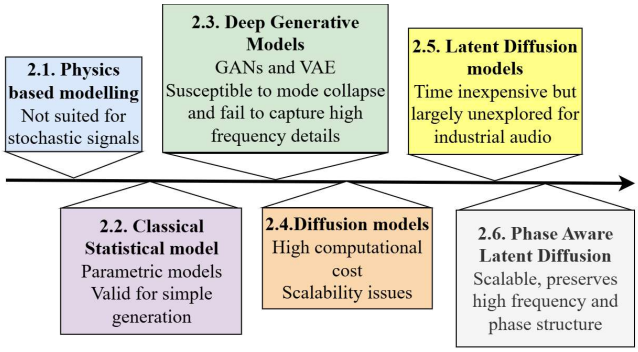

Figure 2. Progression of generative modeling approaches for acoustic signal synthesis, illustrating the evolution from physics-based and classical statistical methods toward deep generative and latent diffusion frameworks. Each paradigm is characterized by its primary limitation relative to industrial audio requirements: physics-based models lack flexibility for stochastic signals; classical statistical models are restricted to simple parametric distributions; deep generative models (GANs and VAEs) suffer from training instability and loss of high-frequency detail; waveform-domain diffusion models incur prohibitive computational cost at industrial sampling rates; and general-purpose latent diffusion models, while efficient, remain largely unadapted to the unique spectral and phase characteristics of industrial machinery sounds. The proposed framework (Section 2.6) addresses these limitations through a phase-aware latent diffusion design that is simultaneously scalable, computationally efficient, and faithful to the high-frequency and phase structure diagnostically relevant in PHM applications.

2.1. Physics-based Modeling Approaches

Prior to the widespread adoption of deep learning, synthetic audio generation in industrial settings relied primarily

on physics-based modeling and procedural signal synthesis. Physics-based approaches attempted to replicate the acoustic behavior of mechanical systems by explicitly modeling vibration dynamics, structural resonance, and acoustic wave propagation, often derived from governing equations of motion and structural dynamics (Lei et al., 2018; Zhao et al., 2019). While these methods provide physically interpretable representations, they require detailed system knowledge and are difficult to generalize across diverse machine types and operating conditions.

Procedural and signal-based synthesis techniques generated audio using handcrafted rules and parametric signal models. Although effective for simple periodic signals, these approaches struggle to capture the non-stationary, multi-scale, and stochastic nature of real industrial audio signals, limiting their applicability for modern PHM systems.

2.2. Statistical Generative Models

Statistical generative models such as Gaussian Mixture Models (GMMs) (Reynolds & Rose, 1995; McLachlan & Peel, 2000) and Hidden Markov Models (HMMs) (Rabiner, 1989) were widely adopted for audio modeling, particularly in speech and environmental sound synthesis. GMMs represent the spectral distribution of an audio signal as a weighted sum of multivariate Gaussians, providing a tractable probabilistic characterization of stationary acoustic features such as speaker identity and phoneme class (Reynolds & Rose, 1995). HMMs extend this framework by coupling GMM emission densities with Markov state-transition dynamics, enabling the modeling of temporal sequences with time-varying acoustic properties, and became the dominant paradigm for speech recognition and synthesis throughout the 1990s (Rabiner, 1989).

However, the reliance of these models on simplifying assumptions such as Gaussian emission distributions and conditional independence between successive frames fundamentally limits their capacity to model high-dimensional, complex signals. Industrial audio, characterized by overlapping harmonics, broadband turbulence, and short-duration transient impulses (Purohit et al., 2019), violates these assumptions directly: individual fault events produce non-Gaussian, impulsive amplitude distributions, while the superposition of multiple mechanical sources creates non-stationary spectral structure that cannot be captured by a fixed mixture model. Consequently, GMM and HMM approaches yield poor representational fidelity for fault-condition industrial audio and provide no mechanism for synthesizing novel, physically plausible machine sounds at arbitrary operating conditions.

2.3. Deep Generative Models for Audio

Deep learning introduced powerful generative frameworks capable of modeling complex data distributions. Variational Autoencoders (VAEs) (Kingma & Welling, 2014) provide stable training and structured latent representations but often produce over-smoothed outputs, failing to preserve high-frequency transients critical for industrial signals.

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) enable high-fidelity audio generation through adversarial learning. Models such as MelGAN (Kumar et al., 2019) and HiFi-GAN (J. Kong et al., 2020) achieve realistic waveform synthesis with efficient parallel generation. However, GANs suffer from training instability and mode collapse, limiting their ability to capture rare but diagnostically important events.

Autoregressive models such as WaveNet (van den Oord et al., 2016) and SampleRNN (Mehri et al., 2017) achieve high-quality synthesis by modeling sample-level dependencies. Despite their fidelity, their sequential nature makes them computationally infeasible for long-duration, high-resolution industrial audio.

2.4. Diffusion Models for Audio Generation

Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) have emerged as a robust alternative for generative modeling, learning to reverse a stochastic noise process. Diffusion models offer stable training and improved mode coverage compared to GANs, making them suitable for diverse data generation. In audio synthesis, DiffWave (Z. Kong et al., 2021) and WaveGrad (Chen et al., 2021) demonstrate that diffusion models can achieve high-fidelity waveform generation. The score-based formulation further generalizes this framework through stochastic differential equations (Y. Song et al., 2021).

Despite these advances, waveform-domain diffusion is computationally expensive due to the high dimensionality of raw audio signals. Even accelerated samplers such as DDIM (Denoising Diffusion Implicit Model) (J. Song et al., 2021) and DPM-Solver (Lu et al., 2022) require significant computational resources, limiting scalability for industrial applications.

2.5. Latent Diffusion for Efficient Audio Synthesis

Latent Diffusion Models (LDMs) (Rombach et al., 2022) address these limitations by performing diffusion in a compressed latent space learned via an autoencoder. This reduces computational complexity while preserving semantic structure. Recent works extend latent diffusion to audio generation. AudioLDM (Liu et al., 2023) demonstrates text-to-audio generation using latent representations, while AudioLDM2 (Liu et al., 2024) improves generalization through self-supervised learning. Stable Audio (Evans et al., 2024) further extends latent diffusion to long-form audio synthesis.

However, these models are designed for general audio do-

mains and do not explicitly address the unique characteristics of industrial signals, such as transient faults, harmonic structures, and operating-condition variability.

2.6. Phase-Aware Latent Diffusion and Relation to Existing Latent Audio Models

Preserving phase alongside spectral magnitude is a central difficulty in industrial audio modeling, since magnitude-only objectives leave phase under-constrained and degrade waveform fidelity (Griffin & Lim, 1984). Complex-valued networks (Choi et al., 2019) and multi-resolution STFT losses (Yamamoto, Song, & Kim, 2020) address this at the level of the training objective, capturing temporal and spectral structure across multiple scales, but their integration into latent diffusion frameworks remains underexplored. This work proposes such an integration: a multi-resolution phase-aware autoencoder coupled to a latent diffusion model.

AudioLDM (Liu et al., 2023) and AudioLDM2 (Liu et al., 2024) share the latent-diffusion formulation adopted here but differ in the construction of the latent space, and the difference matters for industrial audio. Both diffuse over a variational latent of the mel spectrogram, from which instantaneous phase is absent by construction; the waveform is recovered by a separately trained neural vocoder (J. Kong et al., 2020) that *synthesises* a phase field consistent with the predicted magnitudes. Temporal fine structure is therefore governed by a vocoder prior learned on a general-purpose corpus rather than by the machine under analysis. Listeners are largely insensitive to absolute phase, so this is adequate for speech and general audio; it is not adequate for PHM, where onset timing, crest factor and transient sharpness distinguish a mechanical impact from a smeared reconstruction of one. IndusDiff instead diffuses over the latent of a one-dimensional *waveform* autoencoder supervised by a complex-STFT residual, a windowed normalised cross-correlation term and, for impulsive data, a peak-location term. Phase structure is carried in the latent code itself, and decoding is a single deterministic pass with no vocoder and no phase reconstruction. Stable Audio (Evans et al., 2024) also adopts a waveform latent, but learns phase behaviour implicitly through adversarial training rather than through explicit objectives.

Bandwidth imposes a second limitation. AudioLDM and AudioLDM2 operate at 16 kHz, an 8 kHz Nyquist limit, whereas Section 5.2.2 locates mechanical fault energy in the 5–10 kHz band of the broken-engine condition; the upper half of that band is unrepresentable at 16 kHz. IndusDiff operates at the acquisition rates of the source data (44.1 kHz and 48 kHz). A third difference concerns conditioning: general-purpose models rely on natural-language embeddings and pretraining over thousands of hours of captioned audio, neither of which transfers to industrial monitoring, since machine states are not naturally described by text and no captioned corpus of fault recordings exists at that scale. IndusDiff is trained from scratch on the target machine and conditioned on signal-derived features, a pooled mel context for quasi-stationary motor audio, and sequence-level cross-attention with a frame-level event map for temporally structured assembly audio.

The diffusion formulation itself follows established practice. What differs is the latent space, which is designed so that the phase-sensitive quantities carrying fault signatures survive compression rather than being regenerated at decode time.

3. METHODOLOGY

3.1. Framework Overview

Let $x(t) \in R^T$ denote a real-valued industrial audio waveform signal sampled at high resolution. Direct generative modeling in waveform space is computationally prohibitive due to the high dimensionality of $x(t)$ and the iterative nature of diffusion models.

To address this, we adopt a *latent diffusion framework*, where generation is performed in a compressed latent space learned via an autoencoder. The overall pipeline is defined as:

$$x(t) \xrightarrow{E_\phi} \mathbf{z}_0 \xrightarrow{\text{Diffusion}} \hat{\mathbf{z}}_0 \xrightarrow{D_\psi} \hat{x}(t), \tag{1}$$

where E_ϕ and D_ψ denote the encoder and decoder, respectively, and $\mathbf{z}_0$ is the latent representation. Diffusion is performed entirely in latent space, while waveform space is used only for reconstruction and auxiliary supervision.

This decoupling significantly reduces computational complexity while preserving perceptual fidelity.

We employ a fully convolutional 1D autoencoder operating directly on raw waveforms $x \in R^{1\times T}$, T=132,096 samples (≈3 s at 44.1 kHz) for the engine data and T=240,000 samples (≈5 s at 48 kHz) for the assembly data.

3.1.1. Encoder

The encoder E_ϕ begins with a kernel-7 convolutional stem (Conv–GN–SiLU) and is followed by L hierarchical downsampling blocks. Each block applies strided convolution with residual learning:

$$h^{(l)} = \mathrm{Conv}_{k=4}^{s=2}\big(\mathrm{SiLU}\,\mathrm{GN}\,RB(h^{(l-1)})\big), \tag{2}$$

where $RB(\cdot)$ is a residual block of two Conv–GN–SiLU layers with a skip connection:

$$RB(h) = h + f_2(f_1(h)), \qquad f_i(\cdot) = \mathrm{SiLU}\,\mathrm{GN}\big(\mathrm{Conv}_{k=3}(\cdot)\big). \tag{3}$$

Each stride-2 block halves the temporal resolution, so L blocks give a compression ratio $R = 2^L$, and a final 1×1

convolution projects to $C = 16$ latent channels:

$$R = 2^L, \qquad z_0 = E_\phi(x) \in R^{16\times T_z}, \quad T_z = T/R. \tag{4}$$

The encoder depth is set per dataset to match the temporal structure of each signal regime. Motor audio is quasi-stationary, its diagnostic content lies in a slowly varying spectral envelope and periodic harmonics, so the waveform is temporally redundant and well captured by a coarse latent grid; we use $L = 6$ ($R = 64$) with channel progression $[128, 192, 256, 384, 512, 512, 512]$, giving $T_z = 2064$ for $T = 132{,}096$ while minimizing the latent length and downstream diffusion cost. Assembly audio is the opposite regime: it is dominated by sparse, impulsive transients whose information is concentrated in sub-millisecond onsets with a high crest factor (peak-to-RMS ratio), localized to a small fraction of the clip. The compression ratio bounds both the latent grid spacing (f_s/R frames per second) and the resolution of the frame-level event map $E \in R^{T_z\times 4}$ used for conditioning; at 48 kHz, $R = 64$ places one latent frame every 1.33 ms, whereas $R = 32$ halves this to 0.67 ms. A coarser grid smears the onset across fewer frames and caps the recoverable peak amplitude, making the transient region the binding term in the reconstruction error. We therefore use $L = 5$ ($R = 32$) with channel progression $[128, 192, 256, 384, 512, 512]$ for the assembly model, doubling the latent resolution to $T_z = 7500$ ($T = 240{,}000$) to preserve onset sharpness and peak structure at the cost of a longer latent sequence.

3.1.2. Decoder

The decoder D_ψ mirrors the encoder with L upsampling blocks ($L = 6$ for the motor model, $L = 5$ for the assembly model). Each block applies a learned transposed convolution, a refinement convolution, and a residual block R', with Snake activation:

$$\hat{\mathbf{h}}^{(l)} = R'\Big(\mathrm{Snake}\big(\mathrm{GN}\big(\mathrm{Conv}_{k=3}\big(\mathrm{ConvT}^{s=2}_{k=4}(\hat{\mathbf{h}}^{(l-1)})\big)\big)\big)\Big). \tag{5}$$

The decoder uses the Snake activation (Ziyin, Hartwig, & Ueda, 2020):

$$\mathrm{Snake}(x;\alpha) = x + \frac{\sin^2(\alpha x)}{\alpha}, \tag{6}$$

where α is learned; this periodic inductive bias improves reconstruction fidelity for oscillatory signals and enhances zero-crossing rate (ZCR) preservation. For the motor model, R' is a multi-dilated residual block (MDResBlock) that aggregates multi-scale temporal features through parallel dilated convolutions with $d \in \{1, 3, 9\}$:

$$\begin{aligned}\mathrm{MDResBlock}(\mathbf{h}) &= \mathrm{Snake}\left(\mathbf{h} + \sum_{d\in\{1,3,9\}} g_d(\mathbf{h})\right),\\ g_d(\mathbf{h}) &= \mathrm{GN}\big(\mathrm{Conv}_{k=3,d}\big(\mathrm{Snake}\big(\\ &\qquad \mathrm{GN}\big(\mathrm{Conv}_{k=3,d}(\mathbf{h})\big)\big)\big)\big).\end{aligned} \tag{7}$$

The dilated branches widen the receptive field to reconstruct the extended quasi-periodic and harmonic structure of motor audio. For the assembly model, whose content is a localized impulsive transient rather than sustained periodicity, R' reduces to a single non-dilated residual block ($d = 1$), which suffices for the shorter effective support and lowers decoder capacity. A final reconstruction head (Conv–GN–SiLU–Conv–Tanh) produces the output waveform:

$$\hat{x} = D_\psi(\mathbf{z}_0) \in R^{1\times T}. \tag{8}$$

The detailed architecture of the encoder and decoder employed for the motor (engine) data is shown in Figure 3. For the assembly data we use 5 down- and upsampling blocks instead of 6 to better preserve transient and onset detail; the encoder structure is otherwise identical, while the decoder additionally replaces the multi-dilated residual block with a single non-dilated residual block.

3.1.3. Phase-Aware Reconstruction Loss

The autoencoder is trained with a composite loss that combines established spectral objectives from neural vocoding with domain-specific terms tailored to industrial audio. The two datasets share a common pool of component losses but use different instantiations. The motor model uses the full phase-aware set, including a complex-STFT phase term and two high-frequency emphasis terms:

$$\begin{aligned}\mathcal{L}_{\mathrm{AE}}^{\mathrm{motor}} =& \lambda_{\ell_1}\mathcal{L}_{\ell_1} + \lambda_{\mathrm{MR}}\mathcal{L}_{\mathrm{MR}} + \lambda_{\mathrm{corr}}\mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{rms}}\mathcal{L}_{\mathrm{rms}}\\ &+ \lambda_{\mathrm{hf}}\mathcal{L}_{\mathrm{hf}} + \lambda_{\mathrm{hf}^+}\mathcal{L}_{\mathrm{hf}^+} + \lambda_{\mathrm{dc}}\mathcal{L}_{\mathrm{dc}} + \lambda_{\mathrm{sc}}\mathcal{L}_{\mathrm{sc}}.\end{aligned} \tag{9}$$

For the assembly model, whose recordings are impulsive with near-silent backgrounds, the complex-STFT term ($\lambda_C = 0$) and the high-frequency terms are removed, global phase and high-band objectives are otherwise dominated by reconstruction of the silent noise floor and a peak-location term is added to directly preserve onset sharpness and crest factor:

$$\begin{aligned}\mathcal{L}_{\mathrm{AE}}^{\mathrm{asm}} =& \lambda_{\ell_1}\mathcal{L}_{\ell_1} + \lambda_{\mathrm{MR}}\mathcal{L}_{\mathrm{MR}} + \lambda_{\mathrm{corr}}\mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{rms}}\mathcal{L}_{\mathrm{rms}}\\ &+ \lambda_{\mathrm{peak}}\mathcal{L}_{\mathrm{peak}} + \lambda_{\mathrm{dc}}\mathcal{L}_{\mathrm{dc}} + \lambda_{\mathrm{sc}}\mathcal{L}_{\mathrm{sc}}.\end{aligned} \tag{10}$$

The waveform, multi-resolution STFT, complex-STFT, correlation, RMS, and high-frequency terms are computed on DC-centered signals, $\tilde{x} = x - \bar{x}$ and $\tilde{\hat{x}} = \hat{x} - \bar{\hat{x}}$, whereas the DC, peak-location, and spectral-centroid terms operate on the original signals (the DC term explicitly penalizes the residual

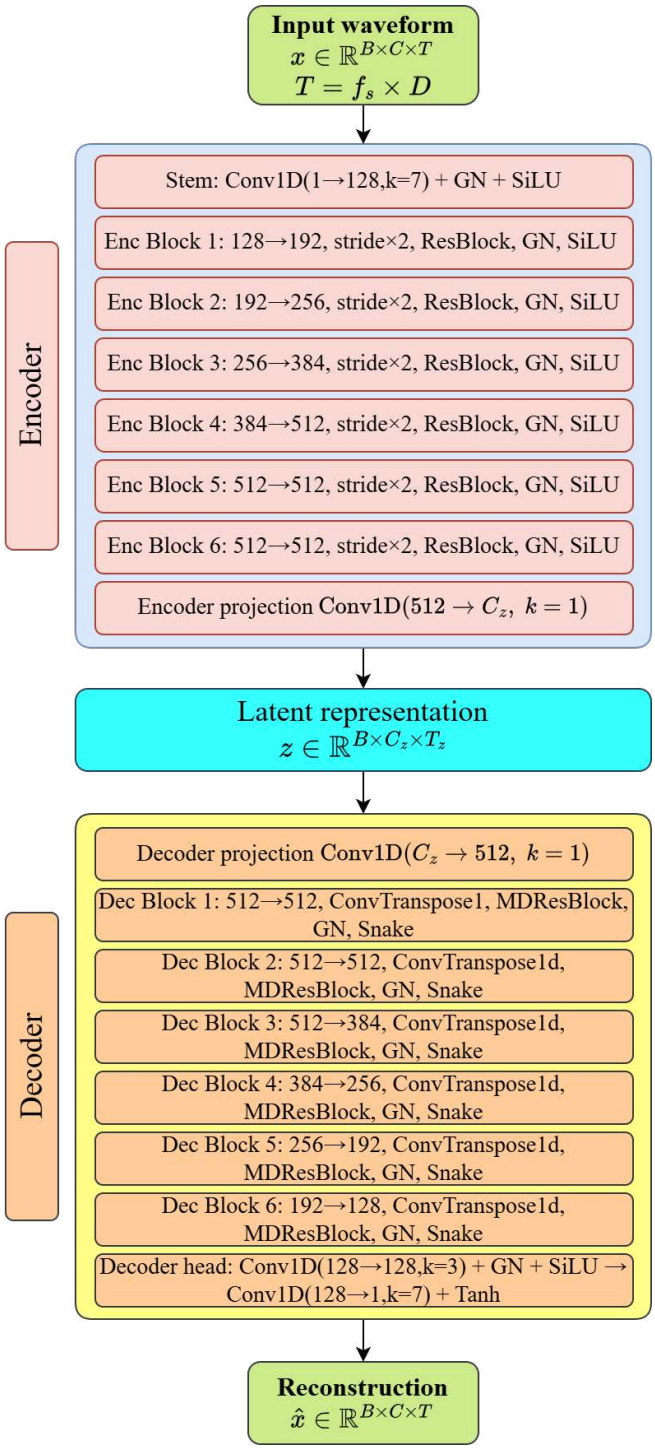

Figure 3. Proposed 1D convolutional phase-aware autoencoder, shown for the motor model ($L = 6$, $R = 64$); the assembly model uses $L = 5$ ($R = 32$) with single-dilation decoder residual blocks.

offset). Each component is described below, and a brief summary is given in Table 1, along with the weights used to train the model. We used a grid search method to optimize the loss function.

Waveform Loss (both). A time-domain ℓ_1 term enforces sample-level fidelity and amplitude alignment:

$$\mathcal{L}_{\ell_1} = \tfrac{1}{T}\|\tilde{x} - \tilde{\hat{x}}\|_1. \tag{11}$$

Multi-Resolution STFT Loss (both). Following Parallel WaveGAN (Yamamoto et al., 2020), the MR-STFT loss enforces spectral consistency across multiple time–frequency scales, combining a spectral-convergence term (first term) with a log-magnitude term (second term) over a set Ω of STFT configurations:

$$\mathcal{L}_{\mathrm{MR}} = \frac{1}{|\Omega|}\sum_{n\in\Omega}\left[\frac{\big\|\,|S_n(x)| - |S_n(\hat{x})|\,\big\|_F}{\big\||S_n(x)|\big\|_F} + \big\|\log|S_n(x)| - \log|S_n(\hat{x})|\big\|_1\right] + \lambda_C\,\mathcal{L}_C. \tag{12}$$

The complex-domain residual term $\mathcal{L}_C$ injects phase information by penalizing the real and imaginary STFT components at the highest resolution, in the spirit of complex-spectrum modeling (Choi et al., 2019; Takaki, Nakashika, Wang, & Yamagishi, 2019); magnitude-only losses are known to leave phase under-constrained (Griffin & Lim, 1984). It is active for the motor model only ($\lambda_C = 0.1$) and disabled for the assembly model, where $\mathcal{L}_{\mathrm{MR}}$ reduces to a magnitude objective:

$$\begin{aligned}\mathcal{L}_C &= \|\mathrm{Re}(\Delta S)\|_1 + \|\mathrm{Im}(\Delta S)\|_1,\\ \Delta S &= S_{2048}(\tilde{x}) - S_{2048}(\tilde{\hat{x}}).\end{aligned} \tag{13}$$

Correlation Loss (both). A windowed normalized cross-correlation term penalizes local waveform decorrelation, complementing the spectral losses with an explicit time-domain phase-alignment objective over N_w non-overlapping windows (window length 100 ms for the motor model, 10 ms for the assembly model):

$$\mathcal{L}_{\mathrm{corr}} = 1 - \frac{1}{N_w}\sum_{i=1}^{N_w}\frac{\langle\tilde{x}_i, \tilde{\hat{x}}_i\rangle}{\|\tilde{x}_i\|\,\|\tilde{\hat{x}}_i\| + \epsilon}. \tag{14}$$

RMS Loss (both). An energy-consistency term matches the global amplitude envelope, which encodes load-state information relevant to downstream PHM:

$$\mathcal{L}_{\mathrm{rms}} = \big|\,\mathrm{rms}(\tilde{x}) - \mathrm{rms}(\tilde{\hat{x}})\,\big|. \tag{15}$$

High-Frequency Loss (motor only). Inspired by pre-emphasis weighting, a high-pass ℓ_1 term at 2 kHz and 8 kHz (H_{f_c} denotes a high-pass at cutoff f_c) emphasizes bright harmonic and fault-related content in the quasi-stationary motor

signal:

$$\mathcal{L}_{\text{hf}} = \|\mathrm{H}_{2k}(\tilde{x}) - \mathrm{H}_{2k}(\tilde{\hat{x}})\|_1, \quad \mathcal{L}_{\text{hf}^+} = \|\mathrm{H}_{8k}(\tilde{x}) - \mathrm{H}_{8k}(\tilde{\hat{x}})\|_1. \tag{16}$$

Peak-Location Loss (assembly only). To preserve the amplitude and timing of the dominant transient, we introduce a term that evaluates the reconstruction at the exact sample of the true peak, $t_b^\star = \arg\max_t |x_b(t)|$, so it cannot be satisfied by high-amplitude content placed elsewhere in the clip:

$$\mathcal{L}_{\text{peak}} = \frac{1}{B} \sum_{b=1}^{B} \big|\, |\hat{x}_b|(t_b^\star) - |x_b|(t_b^\star) \,\big|. \tag{17}$$

DC Loss (both). A DC-offset term suppresses residual bias introduced by the decoder:

$$\mathcal{L}_{\text{dc}} = |\bar{\tilde{x}} - \bar{x}| + |\bar{\tilde{\hat{x}}}|. \tag{18}$$

Spectral Centroid Loss (both). We additionally introduce a spectral-centroid term that matches the frequency-weighted spectral mean, regularizing the global spectral balance (brightness) of the reconstruction:

$$\mathcal{L}_{\text{sc}} = \frac{1}{B} \sum_{b} \frac{|c(x_b) - c(\hat{x}_b)|}{c(x_b) + \epsilon}. \tag{19}$$

3.2. Latent Diffusion Model

3.2.1. Forward Process

Diffusion is performed in latent space. The forward process gradually corrupts $\mathbf{z}_0$ with Gaussian noise:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\, \mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t}\, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \tag{20}$$

where the noise schedule $\{\bar{\alpha}_t\}$ follows a cosine schedule (Nichol & Dhariwal, 2021) over $T = 1000$ timesteps.

3.2.2. Reverse Process and v-Prediction

The denoiser is trained with the velocity ($\mathbf{v}$) parameterization (Salimans & Ho, 2022), which improves stability over noise prediction at low step counts:

$$\mathbf{v}_t = \sqrt{\bar{\alpha}_t}\, \boldsymbol{\epsilon} - \sqrt{1 - \bar{\alpha}_t}\, \mathbf{z}_0. \tag{21}$$

The denoising network f_θ is trained to predict $\mathbf{v}_t$:

$$\mathcal{L}_{\text{diff}} = E\big[\|f_\theta(\mathbf{z}_t, t, \mathbf{c}) - \mathbf{v}_t\|_2^2\big], \tag{22}$$

and the clean latent is recovered at inference via $\hat{\mathbf{z}}_0 = \sqrt{\bar{\alpha}_t}\, \mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t}\, \hat{\mathbf{v}}_t$. For the assembly model, whose signals are temporally sparse, the per-element error is weighted by an event mask m derived from the event map, up-weighting transient frames by a factor ω_{imp}:

$$\mathcal{L}_{\text{diff}}^{\text{asm}} = E\Big[\big\|m \odot \big(f_\theta(\mathbf{z}_t, t, \mathbf{c}, \mathbf{e}, \mathbf{E}) - \mathbf{v}_t\big)\big\|_2^2\Big]. \tag{23}$$

3.2.3. Conditioning

Both models inject the diffusion timestep through feature-wise linear modulation (FiLM) (Perez, Strub, de Vries, Dumoulin, & Courville, 2018) at every residual block, where $[\gamma; \beta]$ are predicted from the time embedding. The two datasets differ in how the acoustic conditioning is applied.

Motor model (global mel-FiLM). The mel spectrogram is mean-pooled into a single global context vector $\bar{\mathbf{c}}$, which drives an additional FiLM modulation at every residual block:

$$[\boldsymbol{\gamma}; \boldsymbol{\beta}] = \mathrm{MLP}(\bar{\mathbf{c}}), \qquad \mathrm{FiLM}(\mathbf{h}) = \boldsymbol{\gamma} \odot \mathbf{h} + \boldsymbol{\beta}, \qquad f_\theta(\mathbf{z}_t, t, \mathbf{c}). \tag{24}$$

No attention or event conditioning is used.

Assembly model (cross-attention + Block-FiLM). Because assembly audio is governed by the temporal position of mechanical events, the mel conditioning is applied as a sequence rather than a pooled vector. The mel context and a prepended event token $\mathbf{e}$ form the keys and values of a cross-attention (Rombach et al., 2022) applied at the coarser U-Net scales. In parallel, the frame-level event map $\mathbf{E}$ drives a Block-FiLM modulation in *every* residual block, providing per-position control aligned to the latent temporal axis, following the temporal-event conditioning of T-Foley (Chung, Lee, & Nam, 2024):

$$\begin{aligned} [\boldsymbol{\gamma}; \boldsymbol{\beta}] &= \mathrm{Conv}_{1\times1}(\mathbf{E}), \\ \mathrm{BlockFiLM}(\mathbf{h}) &= (1 + \boldsymbol{\gamma}) \odot \mathbf{h} + \boldsymbol{\beta}. \end{aligned} \tag{25}$$

$$f_\theta(\mathbf{z}_t, t, \mathbf{c}, \mathbf{e}, \mathbf{E}). \tag{26}$$

The assembly model does not use the pooled global mel-FiLM of the motor model; the mel information enters only through cross-attention.

3.2.4. Classifier-Free Guidance

To enable controllable conditioning strength, classifier-free guidance (Ho & Salimans, 2022) is used: the conditioning is randomly dropped with probability p_{drop} during training, and the conditional and unconditional predictions are combined at inference with a guidance scale γ:

$$\hat{\mathbf{v}}_t^{(\gamma)} = \hat{\mathbf{v}}_t + \gamma\, (\hat{\mathbf{v}}_t^{\mathbf{c}} - \hat{\mathbf{v}}_t). \tag{27}$$

3.2.5. Auxiliary Waveform Losses

Although diffusion operates in latent space, additional supervision is applied in waveform space at low-noise timesteps

Table 1. Reconstruction loss components for the motor and assembly autoencoders. "–" denotes a term disabled for that model. λ_C is the complex-STFT weight inside $\mathcal{L}_{\mathrm{MR}}$.

Term	Role	Motor λ	Asm. λ	Source
Waveform ℓ_1	Sample-level fidelity	2.0	2.0	Standard
MR-STFT	Multi-scale spectral consistency	0.1	1.0	Yamamoto et al. (Yamamoto et al., 2020)
Complex STFT (λ_C)	Phase consistency (Re/Im)	0.1	–	Choi et al. (Choi et al., 2019); Takaki et al. (Takaki et al., 2019)
Correlation	Local time-domain alignment	1.0	0.3	Custom
RMS	Energy / envelope consistency	0.5	1.0	Custom
HF (2 kHz)	Mid-frequency detail	1.0	–	Pre-emphasis (custom)
HF (8 kHz)	High-frequency detail	0.8	–	Pre-emphasis (custom)
Peak-location	Transient peak amplitude/timing	–	2.0	Introduced here
DC	Offset consistency	0.5	0.2	Custom
Centroid	Spectral balance (brightness)	0.8	0.5	Introduced here

($t \leq \tau$), where the predicted clean latent is decoded:

$$\hat{x}_0(t) = D_\psi(\hat{\mathbf{z}}_0). \tag{28}$$

The auxiliary composition is dataset-specific. The motor model uses spectral objectives—log-mel, MR-STFT, complex-STFT, and instantaneous-frequency losses:

$$\mathcal{L}_{\mathrm{aux}}^{\mathrm{motor}} = \lambda_{\mathrm{mel}}\mathcal{L}_{\mathrm{mel}} + \lambda_{\mathrm{MR}}\mathcal{L}_{\mathrm{MR}} + \lambda_{\mathrm{cx}}\mathcal{L}_{\mathrm{cx}} + \lambda_{\mathrm{IF}}\mathcal{L}_{\mathrm{IF}}. \tag{29}$$

The assembly model instead uses temporal objectives tailored to impulsive audio—an envelope-alignment term, a sparsity term penalizing energy outside the event window, and a crest-factor term encouraging sharp transients:

$$\mathcal{L}_{\mathrm{aux}}^{\mathrm{asm}} = \lambda_{\mathrm{env}}\mathcal{L}_{\mathrm{env}} + \lambda_{\mathrm{sp}}\mathcal{L}_{\mathrm{sp}} + \lambda_{\mathrm{cr}}\mathcal{L}_{\mathrm{cr}}. \tag{30}$$

In both cases the final objective combines the latent and auxiliary terms through an epoch-dependent warmup weight $w(e)$:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{diff}} + w(t)\,\mathcal{L}_{\mathrm{aux}}. \tag{31}$$

3.2.6. Inference

Generation begins from Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, I)$ and proceeds by iterative DDIM denoising (J. Song et al., 2021):

$$\mathbf{z}_T \to \mathbf{z}_{T-1} \to \cdots \to \mathbf{z}_0, \tag{32}$$

followed by decoding to a waveform $\hat{x}(t) = D_\psi(\mathbf{z}_0)$. The motor model uses this single-stage procedure directly, while the assembly model employs an extended, event-guided sampling procedure detailed in Section 5.5.2. This latent-space formulation enables efficient generation of high-resolution industrial audio while maintaining perceptual realism.

The detailed architecture of the latent diffusion framework is presented in Figure 4

4. Generation Quality Evaluation

Evaluating the quality of synthetically generated industrial audio requires a multi-faceted assessment strategy. A single metric is insufficient: distributional similarity scores may fail to detect local waveform artifacts, while signal-level measures may not capture perceptual or spectral realism. We therefore structure our evaluation across two complementary axes: qualitative inspection and quantitative measurement, covering temporal fidelity, spectral structure, event-onset localization, and statistical distributional similarity to real machine recordings.

4.1. Qualitative Assessment

Qualitative evaluation provides an intuitive, interpretable view of generation quality that complements numerical metrics by revealing structured artifacts (e.g., spectral holes, envelope distortions, or missing harmonic components) that aggregate scores may obscure.

4.1.1. Waveform-Level Temporal Analysis

We directly superimpose real and generated waveforms in the time domain to qualitatively assess temporal amplitude range, envelope characteristics, and overall signal structure. Let $x^{(r)}(t)$ and $x^{(g)}(t)$ denote real and generated waveforms, respectively, defined over a common temporal support $t \in [0, T_s]$, where T_s represents the clip duration in seconds.

Both signals are plotted on a shared time axis to enable visual comparison of large-scale temporal characteristics, including amplitude modulation patterns, transient behavior, periodicity, and dynamic range consistency. This full-duration visualization facilitates assessment of whether the generative model captures the global structure of the signal distribution.

It is important to note that point-wise alignment between real and generated waveforms is neither expected nor required in a generative modeling framework. Instead, the comparison is inherently distributional rather than sample-wise, emphasizing structural and statistical similarity over exact temporal correspondence. For impulsive assembly recordings we additionally examine the temporal localization of the dominant event, since onset timing unlike fine-grained sample alignment is diagnostically meaningful.

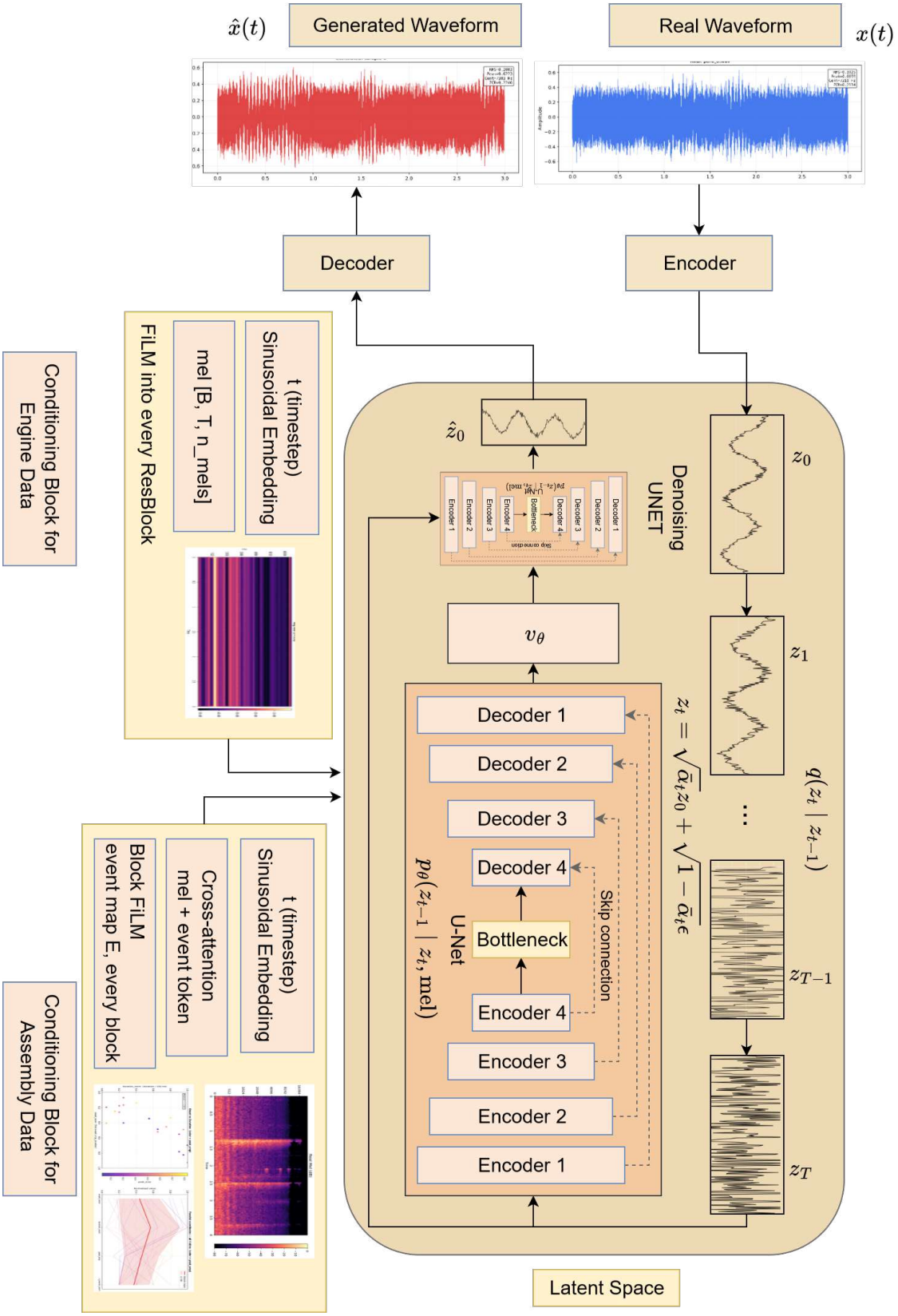

Figure 4. Overview of the proposed phase-aware latent diffusion framework for high-fidelity industrial audio generation.

4.1.2. Average Mel Spectrogram and Difference Map

To analyze spectral fidelity at the population level, we compute mel-scale log-power spectrograms for both real and generated datasets and visualize their averages independently. Let $\{x_i\}_{i=1}^{N}$ denote a set of waveform samples. The average mel spectrogram $\bar{M} \in R^{N_{\text{mel}} \times F}$ is defined as:

$$\bar{M}_{k,f} = \frac{1}{N} \sum_{i=1}^{N} \log \left(\sum_{n=0}^{N_{\text{FFT}}/2} H_k[n] \; |S_i[n,f]|^2 + \epsilon \right), \quad (33)$$

where $S_i[n,f]$ denotes the short-time Fourier transform (STFT) coefficient at frequency bin n and time frame f for sample i, $H_k[n]$ represents the k-th mel filterbank weight, and ϵ is a small constant introduced to ensure numerical stability.

In this study, we employ $N_{\text{mel}} = 128$ mel bands spanning from $f_{\min}$ up to the Nyquist frequency of each dataset (22.05 kHz for the motor data at 44.1 kHz and 24 kHz for the assembly data at 48 kHz).

To further highlight discrepancies between real and generated distributions, we compute a pixel-wise difference map:

$$\Delta M_{k,f} = \bar{M}^{(r)}_{k,f} - \bar{M}^{(g)}_{k,f}, \quad (34)$$

where $\bar{M}^{(r)}$ and $\bar{M}^{(g)}$ denote the average mel spectrograms of the real and generated datasets, respectively.

This difference map reveals systematic spectral biases, identifying frequency–time regions where the generative model consistently overestimates or underestimates energy. Such population-level analysis is particularly critical for industrial audio, where harmonic structures and broadband noise characteristics remain stable across operating conditions and must be reproduced consistently across generated samples.

4.2. Quantitative Assessment

Quantitative metrics provide reproducible and objective measures for evaluating generative performance and enable comparison across models and datasets.

4.2.1. RMS Amplitude

Root-mean-square (RMS) amplitude quantifies the average energy content of a signal:

$$A_{\text{rms}} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} x_t^2}, \quad (35)$$

where T denotes the number of temporal samples. We report the mean and standard deviation of A_{rms} over both real and generated datasets, along with the ratio $A^{\text{gen}}_{\text{rms}}/A^{\text{real}}_{\text{rms}}$.

RMS consistency is particularly important in industrial condition monitoring, where signal energy encodes load-state information that downstream prognostics and health management (PHM) models rely upon. A ratio close to unity indicates preservation of amplitude scale.

4.2.2. Peak Amplitude

The peak amplitude measures the maximum instantaneous magnitude of a waveform:

$$A_{\text{peak}} = \max_{1 \le t \le T} |x_t|. \quad (36)$$

We report the mean peak amplitude across both real and generated sets. Significant deviations indicate distortion of dynamic range, which may adversely affect PHM models that rely on absolute amplitude thresholds.

4.2.3. Spectral Centroid

The spectral centroid characterizes the frequency-weighted mean of the signal spectrum:

$$f_{\text{sc}} = \frac{\sum_{n=0}^{N/2} f_n \, |S[n]|}{\sum_{n=0}^{N/2} |S[n]|}, \quad (37)$$

where $S[n]$ denotes the Fourier transform coefficient at frequency bin n, and $f_n = \frac{n f_s}{N}$ is the corresponding physical frequency, with f_s representing the sampling rate. We report the time-averaged spectral centroid $\bar{f}_{\text{sc}}$ over all short-time Fourier transform (STFT) frames.

In industrial audio, the spectral centroid reflects dominant mechanical frequency components and is sensitive to operational conditions such as rotational speed and load variation. Agreement between real and generated centroids indicates accurate modeling of global spectral balance.

4.2.4. Onset Timing (Event Localization)

For impulsive assembly audio, the temporal position of the dominant mechanical event is a primary quality criterion. We detect the first-event onset t_{on} as the earliest frame at which the short-time RMS envelope e_{RMS} (10 ms frames) exceeds a fraction θ of its peak:

$$t_{\text{on}} = \frac{H}{f_s} \min \Big\{ k : e_{\text{RMS}}[k] \ge \theta \max_j e_{\text{RMS}}[j] \Big\}, \quad (38)$$

where H is the hop length. We report the onset-time distributions of the real and generated pools and their mean deviation $\Delta t_{\text{on}} = | \, \bar{t}^{(g)}_{\text{on}} - \bar{t}^{(r)}_{\text{on}} \, |$; smaller values indicate that generated events occur at the correct time.

4.2.5. Fréchet Audio Distance (FAD)

To evaluate distributional similarity between real and generated audio, we compute the Fréchet Audio Distance (FAD) (Kilgour et al., 2019), which measures the Wasserstein-2 distance between Gaussian distributions fitted to feature embeddings:

$$\mathrm{FAD} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right), \quad (39)$$

where (μ_r, Σ_r) and (μ_g, Σ_g) denote the mean vectors and covariance matrices of embeddings extracted from real and generated datasets, respectively, and $\mathrm{Tr}(\cdot)$ denotes the trace operator. Lower FAD values indicate closer alignment between distributions.

While VGGish (Hershey et al., 2017) was originally used for FAD computation, its performance on non-speech and non-music domains is limited. Tailleur et al. (Tailleur et al., 2024) evaluated multiple embedding models on the DCASE 2023 Task 7 dataset and reported that PANNs CNN14 Wavegram-LogMel achieved the highest correlation with human perceptual scores ($\rho > 0.5$), whereas VGGish exhibited weak correlation ($\rho < 0.1$). These findings have been adopted in the DCASE 2024 Sound Scene Synthesis challenge, which standardizes PANN-based embeddings for evaluation.

Our application focuses on industrial and environmental audio, including motor recordings and assembly processes, which differ significantly from the speech-centric data used to train VGGish. PANNs (Q. Kong et al., 2020), trained on the large-scale AudioSet corpus (5,800 hours, 527 classes), provide representations better aligned with mechanical and environmental sound characteristics. Accordingly, we adopt the CNN14 Wavegram-LogMel model for all FAD computations, implemented using the `fadtk` toolkit (Gui, Gamper, Braun, & Emmanouilidou, 2024).

5. Dataset, Results and Discussion

In this section, we first explain our datasets which we used to train the models. Next, we describe the model parameters we set to obtain results. Next, we explain the data evaluation from the two datasets. Finally, we discuss insights and limitations.

5.1. Dataset Details

To evaluate the proposed latent diffusion framework, experiments were conducted on two complementary industrial audio datasets: (i) a publicly available benchmark of machine operational sounds under documented fault and load conditions, and (ii) a custom dataset collected in a controlled laboratory environment capturing a multi-stage manual assembly process. This dual-dataset strategy enables validation across acoustically distinct scenarios, encompassing both steady-state rotating machinery and temporally structured human-machine interactions, two signal regimes that impose fundamentally different demands on a generative model.

5.1.1. Public Dataset: Sounding Industry

The first dataset is drawn from the Sounding Industry benchmark introduced by Grollmisch (Grollmisch, Abeßer, Liebetrau, & Lukashevich, 2019), a publicly available collection of industrial motor recordings acquired under three documented operating conditions: normal operation (good condition), heavy-load operation, and broken states. The dataset captures realistic variations in machine dynamics, background noise, and operational context, and has been adopted as a standard benchmark for industrial sound analysis and unsupervised anomaly detection (Koizumi et al., 2020; Grollmisch & Abeßer, 2021).

From a generative modelling perspective, this dataset is characterised by quasi-stationary acoustic structure: dominant periodic components associated with shaft rotation harmonics are superimposed on a broadband noise floor whose energy distribution shifts predictably across operating conditions. This property makes the dataset well-suited for assessing whether the proposed framework can capture both the tonal machinery signature and the condition-dependent spectral envelope, which together constitute the diagnostically relevant information targeted by PHM systems.

5.1.2. In-House Dataset: Manual Assembly Process

The second dataset was developed in-house to capture structured assembly operations in a controlled environment. Audio was recorded during the sequential assembly of a consumer-grade computing component, a process comprising multiple mechanically distinct stages, including component placement, alignment, and enclosure that give rise to temporally evolving, non-stationary acoustic signatures.

To introduce natural execution variability, multiple operators participated in the data collection process, producing observed differences in motion dynamics, force application, and inter-event timing. Varying background noise conditions were additionally incorporated to reflect realistic deployment environments. All recordings were acquired using an iPhone 16 Pro Max and subsequently processed to extract high-quality mono audio signals, resampled to a uniform rate of 48 kHz to ensure cross-sample consistency (Salamon & Bello, 2017).

Unlike the Sounding Industry benchmark, this dataset is dominated by transient, impulsive events—brief high-energy bursts corresponding to discrete mechanical interactions such as component seating and fastener tightening—interspersed with near-silence intervals. This signal character presents a

substantially harder generative challenge, as faithful synthesis requires the model to reproduce not only the spectral content of individual events but also their temporal sparsity and sharp onset structure.

5.1.3. Dataset Summary

Table 2 consolidates the key properties of both datasets. Together, they provide complementary evaluation settings: the Sounding Industry benchmark represents standardised, steady-state machine operation with class-annotated condition variability, while the in-house dataset captures complex, operator-influenced assembly processes with rich transient structure. This pairing enables a comprehensive assessment of the framework's capacity to generate realistic industrial audio across the diverse acoustic regimes encountered in real-world PHM deployments. These datasets are visualized in Figure 5.

5.2. Architecture Details

5.2.1. Phase-Aware Autoencoder

The first stage learns a compact latent representation using the phase-aware convolutional autoencoder of Section 3 (Encoder, Eqs. (2)–(4); Decoder, Eqs. (5)–(8)). The architecture is instantiated per dataset to match each signal regime; the two configurations are summarized in Table 3. The encoder applies a kernel-7 stem followed by L strided residual stages, giving a compression ratio $R = 2^L$ and a latent $\mathbf{z} = E_\phi(\mathbf{x}) \in R^{16\times T_z}$ with $T_z = T/R$. The motor model uses $L = 6$ ($R = 64$, $T_z = 2064$ for $T = 132{,}096$ at 44.1 kHz); the assembly model uses $L = 5$ ($R = 32$, $T_z = 7500$ for $T = 240{,}000$ at 48 kHz), trading a longer latent sequence for finer temporal resolution to preserve impulsive onsets. The decoder mirrors the encoder with transposed-convolution upsampling; the motor decoder uses multi-dilated residual blocks (dilations $\{1, 3, 9\}$), while the assembly decoder uses a single non-dilated residual block.

Reconstruction objective. Both models are trained with the composite phase-aware loss (components and weights in Table 1). The motor model uses the full set, including the complex-STFT phase term and the 2 and 8 kHz high-frequency terms; the assembly model disables the complex-STFT ($\lambda_C = 0$) and high-frequency terms and adds the peak-location term, reflecting its sparse, impulsive structure.

Training details. Both autoencoders are optimized with AdamW under a linear warmup followed by cosine decay to $\eta_{\min} = 10^{-6}$; per-model settings are given in Table 4. The motor model uses weight decay 0.01 and gradient clipping at $\|\mathbf{g}\|_2 \leq 1.0$.

The configurations in Tables 3 and 4 were selected through a constrained pilot study informed by established audio autoencoder architectures, the characteristics of each signal regime, and the available computational resources. For motor recordings, the quasi-stationary harmonic structure motivated a temporal downsampling factor of 64 together with a multi-dilated decoder; for assembly recordings, a lower factor of 32 retained the finer latent temporal resolution needed to represent short-duration mechanical events. Batch sizes were constrained by GPU memory, while learning rates and loss weights were refined across pilot runs and assessed by reconstruction performance and training stability, with candidate configurations compared using standard waveform and spectral reconstruction metrics.

5.2.2. Autoencoder Reconstruction Quality

Prior to evaluating generation quality, we assess the reconstruction fidelity of the Autoencoder (AE) in isolation by encoding and decoding audio recordings from all four conditions. This establishes condition-specific upper bounds on generation quality: any distributional error introduced by the AE decoder propagates through the full pipeline irrespective of diffusion model performance.

Reconstruction quality is evaluated using signal-to-noise ratio (SNR), waveform ℓ_1 distance, log-Mel spectral distance (d_{Mel}), RMS amplitude error, and spectral centroid error. Table 5 reports mean and standard deviation across all conditions. Reconstructed signals of engine dataset and assembly dataset are visualized in Figure 6 and Figure 7 respectively. A complementary quantitative view for the engine conditions is given in Figure 13, comparing per-file amplitude distributions, mean-amplitude correlation, and spectral centroid between original and reconstructed signals.

Heavy load. The heavy-load condition achieves the highest reconstruction fidelity, with mean SNR of 26.99 ± 0.34 dB and waveform $\ell_1 = 0.010$. These recordings exhibit sustained, quasi-periodic structure with high RMS amplitude, making them highly compressible in latent space. RMS amplitude is preserved within 0.09%, and centroid error is 2.15 Hz, indicating accurate reconstruction of both amplitude envelope and spectral balance.

Good engine. The good engine condition achieves moderate reconstruction fidelity with SNR 16.64 ± 0.95 dB. Increased variance compared to the heavy-load case reflects greater diversity in operating cycles. The centroid error (5.15 Hz) is the largest among engine conditions, indicating mild harmonic instability.

Broken engine. The broken engine condition yields SNR 13.03 ± 0.10 dB, the lowest among engine datasets but with

Table 2. Summary of experimental datasets used for evaluation.

Property	Sounding Industry (Grollmisch et al., 2019)	In-House Assembly
Availability	Public	Private (in-house)
Source	Industrial motors	Manual assembly process
No. samples	356 (105+127+124)	103
Clip duration	3.0 s	5.0 s
Sampling rate	44.1 kHz	48 kHz (resampled)
Audio format	Mono	Stereo converted to Mono
Operating conditions	Normal, heavy-load, fault	Multi-stage assembly, multiple operators

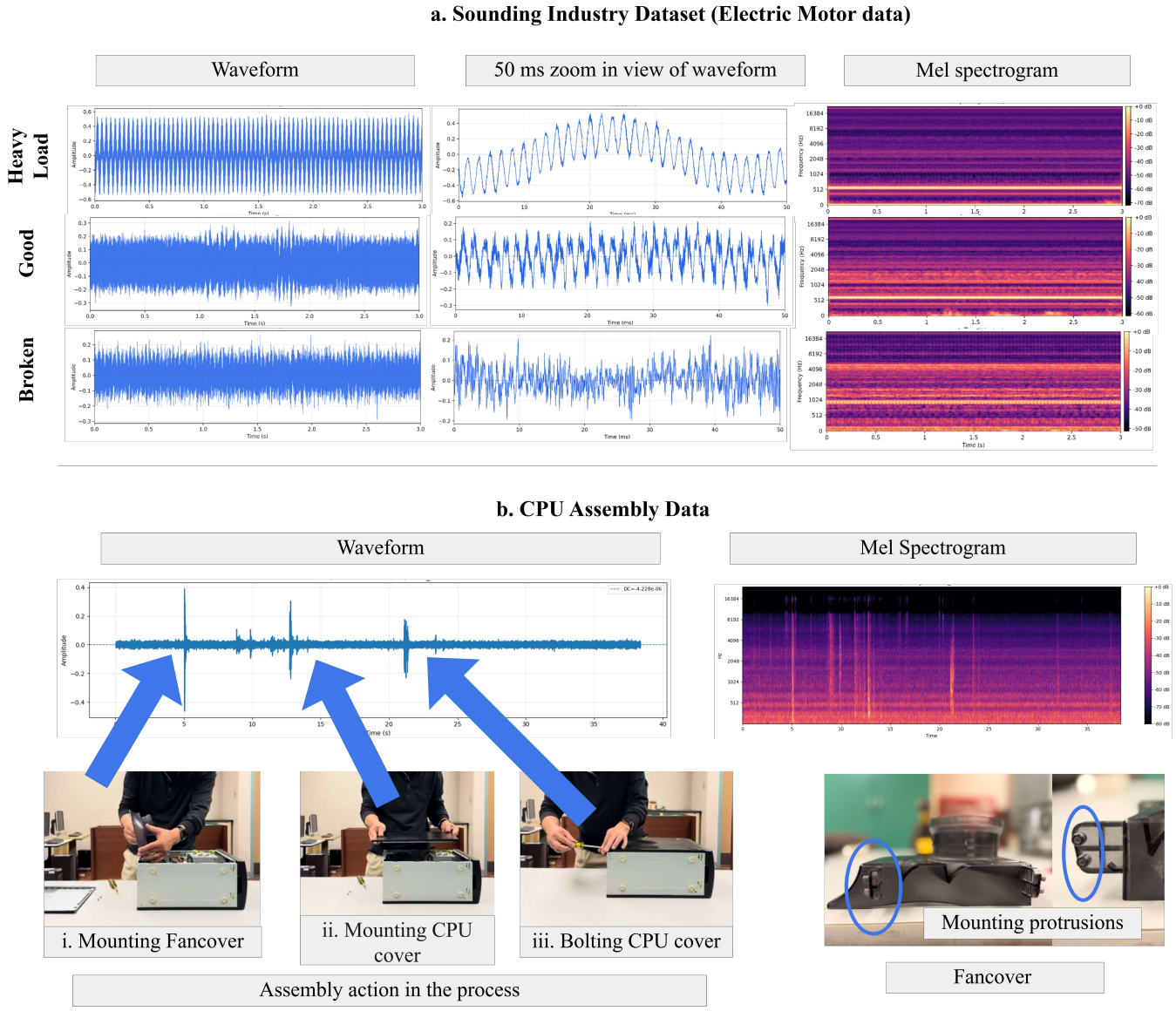

Figure 5. Representative visualisation of industrial audio signals across operating conditions and assembly stages. *Top row:* Waveform overlays, 50 ms zoomed views, and mel-spectrograms for three machine states (heavy-load, good, and broken condition). *Bottom row:* Temporal waveform and mel-spectrogram of a representative assembly recording, aligned with discrete mechanical events including fan-cover mounting, CPU-cover placement, and bolting.

minimal variance, indicating consistent yet lossy reconstruc- tion.

Table 3. Autoencoder configuration for the two datasets.

	Motor	**Assembly**
Sample rate	44.1 kHz	48 kHz
Clip duration	3.0 s	5.0 s
Input length T	132,096	240,000
Stages L / ratio R	6 / 64	5 / 32
Latent $C_z \times T_z$	16×2064	16×7500
Decoder res. block	dilated $\{1, 3, 9\}$	single (non-dilated)
Reconstruction loss	full phase-aware	reduced + peak

Table 4. Autoencoder training settings.

	Motor	**Assembly**
Optimizer	AdamW	AdamW
Learning rate	5×10^{-5}	2×10^{-4}
Batch size	4	2
Epochs	1000	250
Warmup epochs	25	50

Error is concentrated in two regions: (i) the 320–640 Hz band, corresponding to fault-affected harmonic content, and (ii) the 5–10 kHz band, where positive SNR ($+3.55$ dB) reflects real high-frequency energy from mechanical faults (e.g., bearing impacts). This energy is partially attenuated by the latent bottleneck, establishing a ceiling on reconstructable fault detail.

Assembly (lab dataset). The assembly dataset presents the most challenging case, with mean SNR 8.26 ± 4.90 dB and high variance due to sparse impulse-dominated structure.

Segment-wise analysis yields impulse SNR of 9.86 dB and silence-region SNR of 13.16 dB. The reported RMS error (11.39%) is dominated by amplitude imbalance between silence and transient segments and does not reflect reconstruction fidelity alone.

Centroid error (18.35 Hz) and log-Mel distance (0.359) indicate that while broadband characteristics are captured, the model struggles to reproduce precise transient sharpness and spectral detail, particularly above 4 kHz.

Reconstruction ceilings. These results establish condition-dependent bottlenecks imposed by the autoencoder. For steady-state engine signals, reconstruction is highly accurate and generation performance is primarily diffusion-limited.

For fault conditions, spectral attenuation in high-frequency bands limits achievable fidelity. For impulsive assembly data, the inability to reconstruct sharp transients constitutes the dominant limitation.

These constraints directly influence generation performance and must be considered when interpreting downstream FAD.

5.3. Autoencoder Ablation Study

To evaluate the importance of the proposed reconstruction objective, an ablation study was performed by training the autoencoder with and without the composite loss formulation. The study is conducted on both the motor and assembly datasets, since the two signal regimes place different demands on the objective and, as shown below, fail in qualitatively different ways when the composite terms are removed.

5.3.1. Motor Data

Table 6 reports the ablation for the heavy-load and broken engine conditions. In both cases the autoencoder trained with the proposed loss achieves substantially better reconstruction across every metric, while removing the composite terms causes a collapse rather than a degradation. Under the heavy-load condition the SNR falls from 26.99 dB to -2.12 dB and the centroid error rises from 2.15 Hz to 6462.8 Hz; under the broken condition the SNR falls from 13.03 dB to -5.44 dB and the centroid error rises from 1.49 Hz to 4566.05 Hz. A negative SNR indicates that the reconstruction error exceeds the energy of the original signal, so the ablated model does not produce a degraded version of the input but an essentially unrelated one.

The RMS error is equally revealing: it rises from 0.09% to 16.71% under heavy load and from 0.13% to 58.28% under the broken condition, indicating that without the composite objective even the global amplitude envelope that encodes load state for downstream PHM is not recovered. The low standard deviations of the ablated model across all files of the broken condition (± 0.023 dB SNR, $\pm 0.541\%$ RMS) show that this is a systematic failure of the training objective rather than instability on a subset of recordings.

The waveform and mel-spectrogram comparisons in Figures 8 and 9 further confirm that the ablated model preserves neither temporal structure nor spectral distribution. Since the autoencoder without the proposed loss already produced poor latent reconstructions, diffusion training was not pursued for this ablated setting, as the downstream latent diffusion model would inherit these reconstruction errors from the latent space.

5.3.2. Assembly Data

Impulsive assembly audio poses a different question. Because the signal consists of sparse transients embedded in near-silence, a purely sample-wise objective can attain low error without reproducing the acoustic structure that matters. We therefore repeat the ablation on the assembly autoencoder, comparing the full composite objective of Eq. (10) against a model trained with the waveform ℓ_1 term alone, with all architectural and optimisation settings held fixed.

Table 7 reports the outcome, and the two configurations sepa-

Table 5. Autoencoder reconstruction quality across all conditions.

Condition	n	SNR (dB)	ℓ_1	d_{Mel}	RMS (%)	Centroid (Hz)
Engine1 (good)	105	16.64 ± 0.95	0.021	0.247	0.11	5.15
Engine2 (broken)	124	13.03 ± 0.10	0.026	0.252	0.13	1.49
Engine3 (heavy)	127	26.99 ± 0.34	0.010	0.308	0.09	2.15
Assembly (lab)[†]	103	8.26 ± 4.90	0.009	0.359	11.39	18.35

[†] Assembly impulse SNR: 9.86 dB; silence-region SNR: 13.16 dB. Overall RMS error is inflated due to amplitude variation between silence and impulse segments.

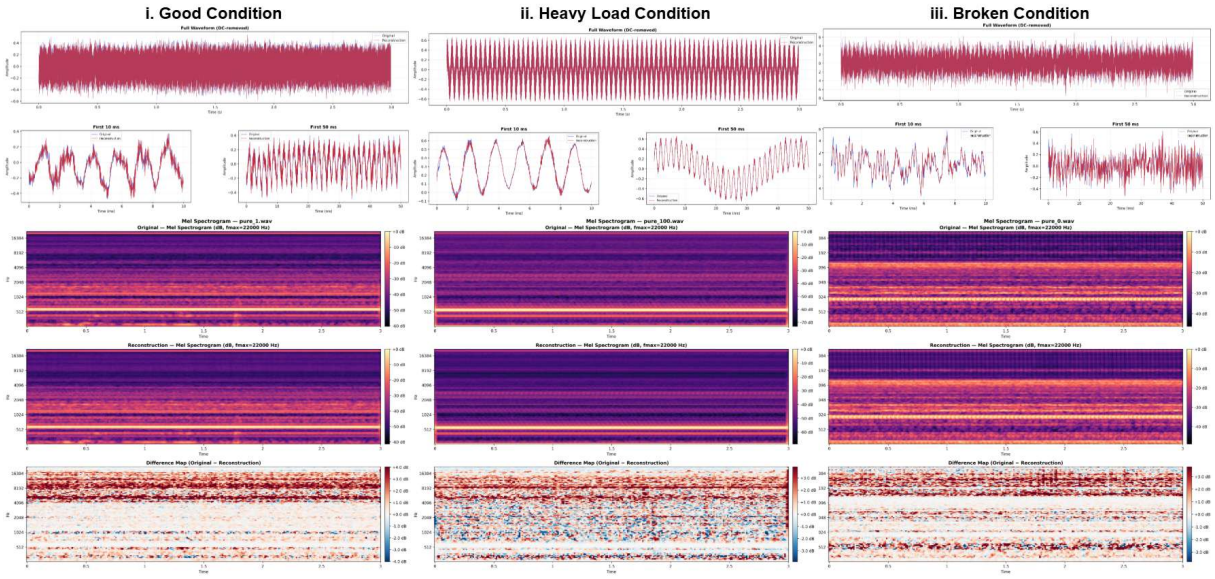

Figure 6. Qualitative evaluation of autoencoder reconstruction fidelity across three representative audio samples. Each column corresponds to a different signal type. From top to bottom: (i) full waveform comparison (DC-removed), (ii) zoomed-in temporal views (first 10 ms and 50 ms) highlighting phase and amplitude alignment, (iii) mel spectrograms of the original signals, (iv) mel spectrograms of the reconstructed signals, and (v) spectrogram difference maps (original minus reconstruction).

Table 6. Autoencoder ablation on the motor dataset, for the heavy-load and broken engine conditions, with and without the proposed composite reconstruction loss. Negative SNR indicates that the reconstruction error exceeds the original signal energy.

Condition	Loss	SNR (dB) ↑	ℓ_1 ↓	RMS (%) ↓	Centroid (Hz) ↓
Heavy load	With	**26.99**	**0.010**	**0.09**	**2.15**
	Without	−2.12	0.306	16.71	6462.80
Broken	With	**13.03**	**0.026**	**0.13**	**1.49**
	Without	−5.44	0.241	58.28	4566.05

Table 7. Autoencoder ablation on the assembly dataset. The ℓ_1-only model optimises precisely the quantity that SNR and ℓ_1 error measure and consequently scores better on both, while the composite objective reduces the mean spectral-centroid error roughly fifty-fold.

Metric	ℓ_1 only	Composite
Mean SNR (dB) ↑	**15.73**	8.26
Mean ℓ_1 waveform ↓	**0.00304**	0.009
Mean centroid error (Hz) ↓	939.90	**18.35**

rate along the axis that distinguishes sample-wise from spectral fidelity. The ℓ_1-only model achieves the higher SNR (15.73 vs. 8.26 dB) and the lower waveform error (0.00304 vs. 0.009). This is the expected result rather than a favourable one: SNR and ℓ_1 error are monotone functions of the quantity that model was trained to minimise, so the comparison on these two metrics is not informative about reconstruction quality in any broader sense. It is, however, informative about how misleading those metrics are for this signal regime. Because assembly recordings are dominated in sample count by near-silent background—the transient occupies a small fraction of each 5 s clip—a model can drive the pointwise error down by reconstructing the background accurately while smoothing the impulsive onsets, and the aggregate SNR will improve rather than degrade.

The spectral-centroid comparison exposes this directly. The ℓ_1-only reconstruction deviates from the true centroid by 939.90 Hz on average, roughly half the mean centroid of the assembly recordings, indicating that the frequency balance of

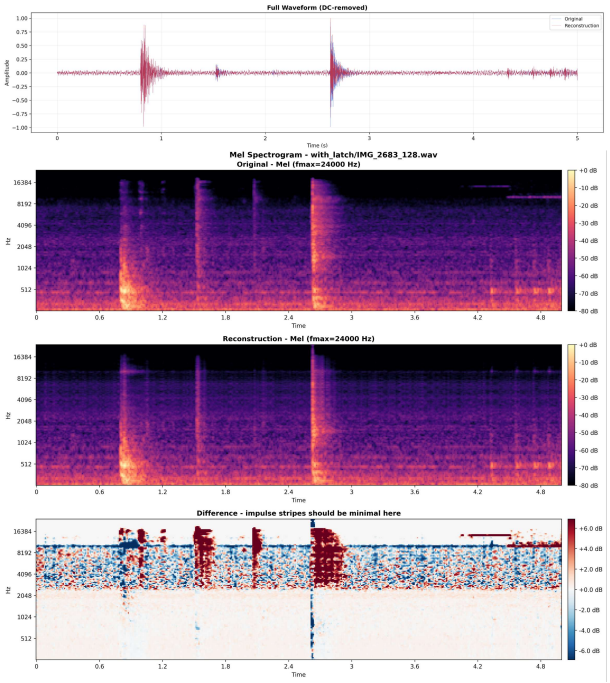

Figure 7. Qualitative reconstruction analysis of an impulse-dominated audio sample. Top: waveform comparison (DC-removed) between original and autoencoder reconstruction, showing accurate temporal alignment and preservation of the dominant transient event. Middle panels: mel spectrograms of the original and reconstructed signals, illustrating that the autoencoder retains the primary spectral structure. Bottom: spectrogram difference map (original minus reconstruction), where low-energy regions remain largely consistent, while deviations are primarily concentrated in higher-frequency bands and weaker background components. The minimal discrepancy around the main impulse indicates strong preservation of transient timing and energy, with residual errors mainly affecting fine spectral details rather than global structure.

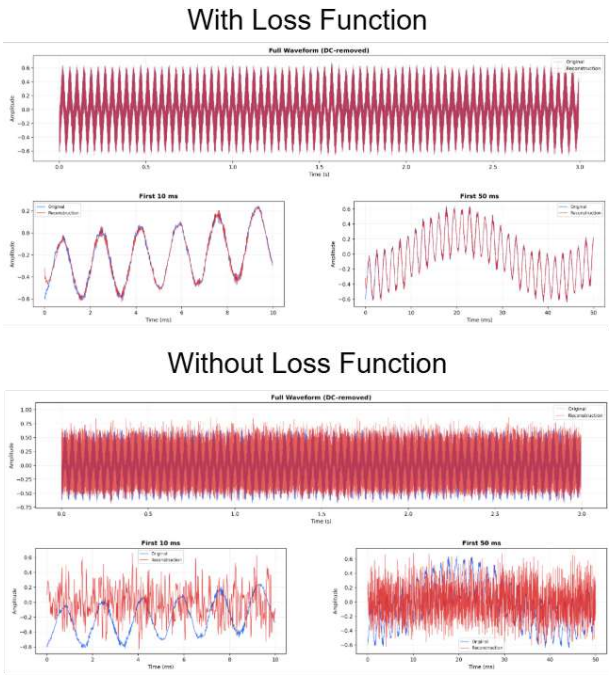

Figure 8. Waveform-level autoencoder ablation comparison with and without the proposed composite reconstruction loss. The model trained with the proposed loss closely follows the original waveform at both full-duration and zoomed temporal scales, whereas the ablated model produces noisy and poorly aligned reconstructions.

the reconstructed signal bears little resemblance to the original despite the favourable waveform metrics. The composite objective reduces this to 18.35 Hz, a fifty-fold improvement. Since the diagnostic content of assembly audio is concentrated in the high-frequency components of the transient onsets, a reconstruction that misplaces the spectral centre of mass by this margin cannot support a downstream latent diffusion model, irrespective of its sample-wise accuracy.

5.4. Latent Diffusion Model

Generation is performed in the latent space using the diffusion model of Section 3 (forward process, Eq. (20); v-prediction, Eqs. (21)–(22); classifier-free guidance). The frozen encoder produces the clean latent $\mathbf{z}_0 = E_\phi(\mathbf{x})$, which is corrupted under a cosine noise schedule (Nichol & Dhariwal, 2021) over $T_{\text{diff}} = 1000$ steps and denoised with the velocity parameterization (Salimans & Ho, 2022).

5.4.1. Denoising architecture.

The denoiser $f_\theta(\mathbf{z}_t, t, \mathbf{c})$ is a four-level 1D U-Net with a bottleneck, skip connections, and two residual blocks per stage, hidden width 512 throughout ($[512, 512, 512, 512]$). The timestep is encoded by a sinusoidal embedding followed by a two-layer MLP, $\mathbf{e}_t \in R^{256}$, and injected through FiLM

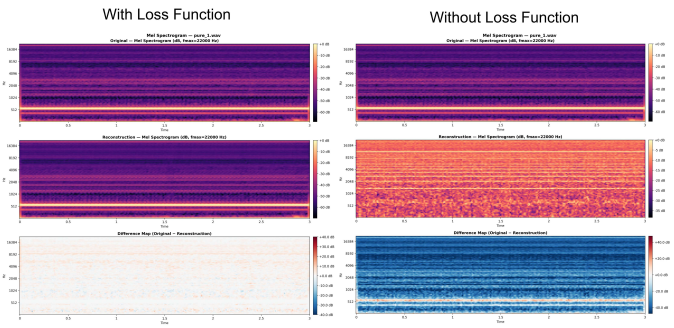

Figure 9. Mel-spectrogram comparison for the autoencoder ablation study. With the proposed loss, the reconstructed spectrogram preserves the dominant harmonic structure and produces a low-error difference map. Without the proposed loss, the reconstruction introduces strong broadband spectral distortion, confirming that the proposed loss is essential for learning a useful latent representation.

at every residual block in both models.

5.4.2. Conditioning.

The conditioning differs per dataset.

Motor (Sounding Industry) — global mel-FiLM. At 44.1 kHz ($N_{\text{samples}} = 132{,}096$) the mel spectrogram yields $\mathbf{c}_{\text{SI}} \in R^{259\times128}$, which is projected to $D_{\text{ctx}} = 256$ and mean-pooled to a global vector $\bar{\mathbf{c}} \in R^{256}$ driving FiLM at every block:

$$[\boldsymbol{\gamma};\boldsymbol{\beta}] = \text{MLP}_{\text{FiLM}}(\bar{\mathbf{c}}), \quad \text{FiLM}(\mathbf{h}) = \boldsymbol{\gamma} \odot \mathbf{h} + \boldsymbol{\beta}. \tag{40}$$

No attention or event conditioning is used.

Assembly — cross-attention + event conditioning. At 48 kHz ($N_{\text{samples}} = 240{,}000$) the mel spectrogram yields $\mathbf{M}_{\text{Asm}} \in R^{469\times128}$, used as cross-attention (Rombach et al., 2022) keys/values, interpolated to the U-Net scales

$$L \in \{7500,\ 3750,\ 1875,\ 937,\ 468\}, \tag{41}$$

with head count

$$N_{\text{heads}} = \begin{cases} 4, & L \geq 1000, \\ 8, & L < 1000. \end{cases} \tag{42}$$

Attention is applied at the coarser scales only; the finest scale (7500) and the largest decoder scale use no attention. A 6-dimensional event token $\mathbf{e} \in R^6$ (four event descriptors plus a two-way condition one-hot) is prepended to the mel context as an additional token, and a frame-level event map $\mathbf{E} \in R^{7500\times4}$ drives Block-FiLM in every residual block, following the temporal-event conditioning of T-Foley (Chung et al., 2024). Self-attention is disabled; only cross-attention to the conditioning context is used.

Table 8. Latent diffusion training settings.

	Motor	**Assembly**
Learning rate	5×10^{-5}	1×10^{-5}
Effective batch size	8	2 (1×accum 2)
Epochs	1800	2590
Weight decay	0.01	1×10^{-4}
Gradient clip	1.0	0.5
Cond. dropout p_{drop}	0.1	0.20
Aux. threshold τ	100	500
EMA decay	—	0.9999

5.4.3. Training objective.

Both models use the velocity objective of Eq. (22) with classifier-free guidance (Ho & Salimans, 2022) (conditioning dropped with probability p_{drop}; Table 8). For the assembly model the latent error is event-weighted, up-weighting transient frames by $\omega_{\text{imp}} = 4$. At low-noise steps $t \leq \tau$ the predicted clean latent is decoded and supervised in waveform space; the auxiliary set is dataset-specific:

$$\mathcal{L}_{\text{aux}}^{\text{motor}} = \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \lambda_{\text{MR}}\mathcal{L}_{\text{MR}} + \lambda_{\text{cx}}\mathcal{L}_{\text{cx}} + \lambda_{\text{IF}}\mathcal{L}_{\text{IF}}, \tag{43}$$

$$\mathcal{L}_{\text{aux}}^{\text{asm}} = \lambda_{\text{env}}\mathcal{L}_{\text{env}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} + \lambda_{\text{cr}}\mathcal{L}_{\text{cr}}, \tag{44}$$

with weights $(\lambda_{\text{mel}}, \lambda_{\text{MR}}, \lambda_{\text{cx}}, \lambda_{\text{IF}}) = (0.05, 0.10, 0.05, 0.02)$ for the motor model and $(\lambda_{\text{env}}, \lambda_{\text{sp}}, \lambda_{\text{cr}}) = (0.25, 0.45, 0.12)$ for the assembly model. The total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + w(e)\mathcal{L}_{\text{aux}}$, where $w(e)$ ramps linearly to 1 over the auxiliary-warmup epochs (50 for the motor model, 30 for the assembly model). For the assembly model the decode step is non-differentiable, so the temporal terms act through a surrogate-gradient regularizer on the latent.

5.5. Inference

5.5.1. Sounding Industry Dataset

For the Sounding Industry benchmark, inference follows the standard single-stage latent diffusion procedure described in Section 3. Sampling begins from Gaussian noise $z_T \sim \mathcal{N}(0, I)$ in the latent space $R^{16\times2{,}064}$ and proceeds using DDIM over $N_{\text{DDIM}} = 50$ reverse steps:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{z}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\,\hat{\varepsilon}_t, \tag{45}$$

conditioned on the mel-spectrogram features $\mathbf{c}$ of a reference recording from the target operating condition. The denoised latent $\hat{z}_0$ is decoded by the frozen autoencoder decoder D_ψ to recover a full-resolution waveform at 44.1 kHz. A total of 300 samples were generated per operating condition (normal, heavy-load, and fault) for evaluation.

5.5.2. In-House Assembly Dataset: Event-Guided Inference

Assembly recordings are characterised by sparse transient events interleaved with near-silence, so their structure is dominated by the temporal position of each mechanical event rather than a stationary spectral envelope. Inference is therefore conditioned on the event signals and uses a latent-domain background-suppression step to sharpen the transient.

Conditioning representation. In addition to the mel context $\mathbf{c}$, the assembly model accepts two event-derived signals extracted from each reference recording: an event token $\mathbf{e} \in R^6$ (four temporal descriptors plus a two-way condition one-hot) and a frame-level event map $\mathbf{E} \in R^{7500\times 4}$ over the latent temporal axis. The event token is prepended to the mel context as an additional cross-attention token (encoding the global onset), while the event map drives Block-FiLM at every residual block (encoding the precise latent-frame location). Conditioning tensors were precomputed from the 103 reference recordings.

Sampling. For each conditioning file, latents are generated by a single DDIM pass over $N_{\text{DDIM}} = 50$ reverse steps using the EMA (Exponential Moving Average) model weights, conditioned on $(\mathbf{c}, \mathbf{e}, \mathbf{E})$. A low classifier-free guidance scale is used ($\gamma \in \{1.0, 1.5\}$), reflecting that the event map already supplies strong temporal structure.

Latent background suppression. Before decoding, the predicted clean latent is sharpened by attenuating non-event regions. Outside the detected event window, each latent channel is blended toward its training mean $\bar{\mathbf{z}}_c$ with a residual floor,

$$\mathbf{z}'_c = \bar{\mathbf{z}}_c + g\,(\mathbf{z}_c - \bar{\mathbf{z}}_c), \qquad g = \mathbf{m} + (1-\mathbf{m})\,f_{\text{floor}}, \quad (46)$$

where $\mathbf{m}$ is the cosine-faded event mask ($g = 1$ inside the event region) and $f_{\text{floor}} = 0.15$ retains 15% of the background energy outside it. Pulling toward the mean rather than zeroing keeps the decoder in-distribution while protecting the impulse. The suppressed latent is decoded by D_ψ and the waveform is loudness-normalised to -14 LUFS (Loudness Units relative to Full Scale) for evaluation.

Output scale. Each of the 103 conditioning files was processed with $N_{\text{seeds}} = 3$ seeds yielding 309 generated clips at 48 kHz. FAD was computed on these outputs.

The Sounding Industry pipeline comprises 28.6M parameters in the phase-aware autoencoder (8.3M encoder, 20.3M decoder) and 55.2M parameters in the FiLM-only conditioning U-Net, giving a total of **83.8M trainable parameters**. The assembly pipeline uses a separate, lighter autoencoder (13.1M; 5.7M encoder, 7.4M decoder), reflecting its shallower five-stage design and single-dilation decoder residual blocks together with a 52.5M conditioning U-Net that adds cross-attention and Block-FiLM event conditioning, giving a total of **65.6M trainable parameters**. In the motor autoencoder the decoder dominates (20.3M of 28.6M), reflecting the multi-dilated residual blocks with Snake activation used for high-fidelity reconstruction; the assembly decoder is substantially lighter, as it replaces these with single residual blocks. The time required to complete each stage of the full pipeline is listed in Table 9.

5.6. Generation Results

Generation quality was evaluated separately for each operating condition of the Sounding Industry dataset—normal, heavy-load, and fault (broken)—and for the in-house assembly dataset. For the Sounding Industry benchmark, a dedicated diffusion model was trained per condition on the corresponding real recordings, and 300 samples were generated per condition for evaluation. Signal-level statistics (RMS amplitude, peak amplitude, and spectral centroid) were computed over the full real and generated pools, and the Fréchet Audio Distance (FAD) was computed using PANNs CNN14 Wavegram-LogMel embeddings via the `fadtk` toolkit (Gui et al., 2024), with all audio normalised to -14 LUFS prior to feature extraction. Qualitative assessment was performed via full-duration waveform overlays and population-averaged mel spectrogram comparisons with pixel-wise difference maps.

5.6.1. Sounding Industry: Quantitative Results

Table 10 consolidates the quantitative generation metrics across all three operating conditions of the Sounding Industry dataset and the in-house assembly dataset.

Normal condition. The model achieves close RMS amplitude agreement (ratio $= 1.011$), indicating that the $\mathcal{L}_{\text{rms}}$ term in the autoencoder loss propagates correctly to the generated distribution. The spectral centroid of generated audio is 98.5 Hz higher than real recordings, a relative shift of $\approx 1.4\%$ relative to the 7,269 Hz dataset mean, consistent with a modest high-frequency overestimation in the 1–4 kHz region.

Heavy-load condition. The heavy-load condition yields the strongest signal-level agreement across all metrics. The RMS ratio of 0.998 and peak amplitude agreement (0.662 vs. 0.665) are near-perfect, reflecting the model's ability to capture the elevated broadband energy characteristic of high-load motor operation. The spectral centroid deviation of only -38.3 Hz confirms that the dominant low-frequency energy distribution—centred around 4,455 Hz compared to 7,269 Hz under normal operation, reflecting the increased contribution of sub-1 kHz harmonics under mechanical loading—is repro-

Table 9. Per-stage computational cost for both datasets. Sounding Industry: A100 GPU (Google Colab Pro+). Assembly: NVIDIA GeForce RTX 5070 Laptop (8 GB VRAM). SI inference covers 3 conditions × 300 clips each. Assembly figures aggregate the with- and without-fancover conditions; the assembly diffusion time (†) is wall-clock including training pauses/resumes.

Stage	Description	SI Dataset (A100)	Assembly (RTX 5070)
Preprocessing	Video → mono WAV	—	<1 min
Conditioning extraction	Mel spectrogram, event token, event map	<1 min (127 files)	≈26 s (103 files)
AE training	Phase-aware autoencoder	≈3.6 hr (1,000 epochs)	≈1 hr 55 min (250 epochs)
Latent encoding	Encode all files	<5 s (127 files)	< 2 min
Diffusion training	U-Net denoiser	≈30 min (1,800 epochs)	≈25 hr† (2,590 epochs)
Inference	50 DDIM steps	0.1193 s / clip (≈2 min total)	≈8.1 s / clip (≈41.6 min, 309 clips)
Total pipeline	**All stages combined**	**≈4.1 hr**	**≈27.7 hr†**

Table 10. Generation quality metrics across all evaluated conditions. FAD is computed using PANNs CNN14 Wavegram-LogMel embeddings (Q. Kong et al., 2020); lower is better. RMS and peak amplitude are omitted for the assembly dataset, where the sparse impulse-in-silence structure combined with −14 LUFS normalisation prior to evaluation causes pooled RMS and peak to reflect the loudness normalisation rather than generation fidelity.

Condition	n_{real}	n_{gen}	RMS Real	RMS Gen	Peak Real	Peak Gen	Centroid Real (Hz)	Centroid Gen (Hz)	FAD
SI – Normal	105	300	0.190 ± 0.003	0.192 ± 0.003	0.606	0.570	7269.37	7367.82	5.66
SI – Heavy Load	127	300	0.297 ± 0.006	0.297 ± 0.006	0.662	0.665	4455.06	4416.72	6.01
SI – Broken	124	300	0.159 ± 0.001	0.172 ± 0.001	0.723	0.699	6525.26	6746.53	5.18
Assembly	103	309	-	-	-	-	1860.6	1769.5	54.12

duced faithfully. The PANNs FAD of 6.01 is moderately higher than the normal-condition score, which is attributable to the greater within-class variability of heavy-load recordings making the embedding-space covariance harder to match with the available sample count.

Broken condition. The broken condition presents the most challenging generation target. The RMS ratio of 1.082 indicates that generated audio carries approximately 8.2% more energy than real fault recordings, and the spectral centroid shift of $+221.27$ Hz is the largest across all conditions. This reflects the inherent difficulty of the fault condition: real fault recordings exhibit highly irregular, low-amplitude broadband bursts superimposed on a suppressed harmonic structure, a pattern that is substantially different from the dominant periodic signals seen under normal and heavy-load conditions.

The full per-condition distributions of spectral centroid and peak amplitude for the real and generated pools are shown in Figure 15.

5.6.2. Sounding Industry: Qualitative Results

Figure 10 presents the population-averaged mel spectrogram comparisons for all three Sounding Industry conditions. Across all panels, the generated spectrograms qualitatively reproduce the characteristic stationary harmonic structure of industrial motor audio: the dominant energy band at approximately 512 Hz is present in all generated pools, and the progressive roll-off toward higher frequencies follows the same slope as the corresponding real recordings.

For the normal condition, the difference map (Real − Generated) shows a predominantly positive residual above 4 kHz (up to +3 dB), consistent with the high-frequency attenuation arising from the 64× latent compression. A narrow negative band near 1 kHz indicates a minor energy overestimation in generated audio at the fundamental mechanical frequency.

For the heavy-load condition, the difference map is noticeably more uniform, with residuals remaining within ±1 dB across the diagnostically relevant 512–2,048 Hz range. A slightly elevated positive residual persists above 2 kHz. Notably, the sub-512 Hz region shows a small negative residual, indicating that the diffusion model slightly overestimates the lowest-frequency content.

For the fault condition, the difference map is dominated by a positive residual across most of the spectrum, with the largest discrepancies above 2 kHz (up to +4 dB). The 1 kHz band shows a pronounced white region indicating near-zero difference surrounded by opposing residuals, suggesting that the model has learned the approximate position of the fault-related spectral peak.

For waveform level comparison between real and generated waveforms, refer to Figure 14 in Appendix.

5.6.3. In-House Assembly Dataset

The PANNs FAD of **54.12** is higher than the Sounding Industry scores (5.18–6.01), which is expected given the fundamentally different nature of the assembly signal and the

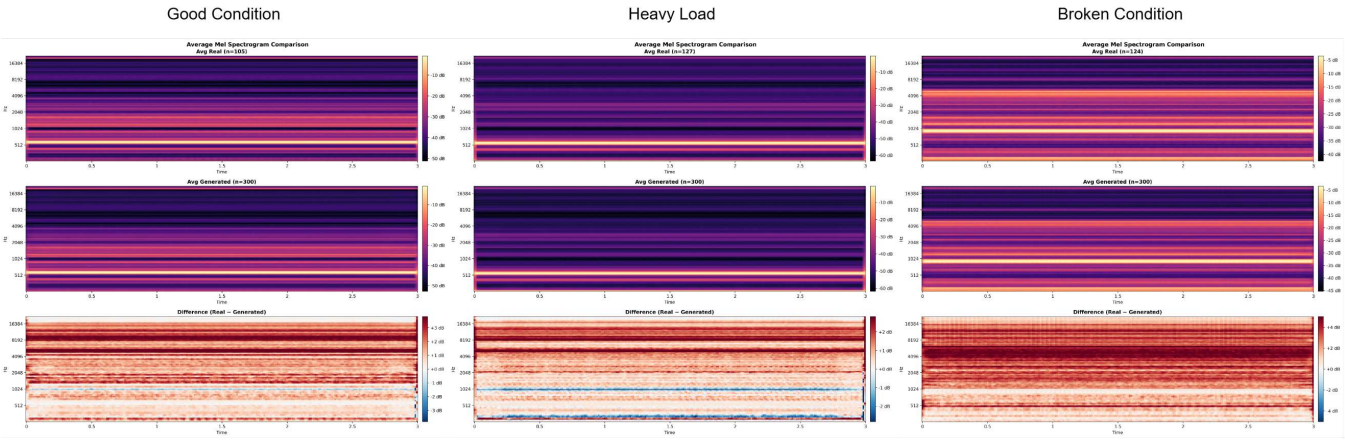

Figure 10. Population-averaged mel spectrogram comparisons for the three Sounding Industry operating conditions (Good, heavy-load, and broken): *top panel* shows the average mel spectrogram of the real pool; *middle panel* shows the average of the generated pool ($n_{\text{gen}} = 300$); *bottom panel* shows the pixel-wise difference map (Real − Generated).

severely constrained dataset size. To contextualise this result, we compare against T-Foley (Chung et al., 2024), a closely related temporal-event-guided Foley sound synthesis model that also employs Block-FiLM conditioning and PANNs-based FAD evaluation. T-Foley was trained on the DCASE 2023 Task 7 dataset comprising approximately 5,000 samples totalling 5.4 hours of audio across 7 sound classes, and achieves a PANNs FAD of 41.59. Our assembly model was trained on 103 clips of 5 seconds each, a total of 8.5 minutes of audio, representing roughly $40\times$ less training data than T-Foley. Despite this extreme data scarcity, our PANNs FAD of 54.12 is within 30.1% of T-Foley's score, suggesting that the pipeline and event-aware conditioning effectively compensate for the limited training set in capturing the temporal structure of impulsive assembly events. The remaining gap is attributable to the combination of dataset size, the non-stationary impulsive signal character of assembly audio, and the known finite-sample bias of FAD scaling as $\mathcal{O}(1/N)$ with $n_{\text{real}} = 103$. The spectral centroid deviation of -91.1 Hz (real: 1,860.6 Hz; generated: 1,769.5 Hz), a relative shift of 4.9%, indicates that the framework successfully captures the broadband spectral character of assembly events, with residual discrepancy concentrated in high-frequency detail above 4 kHz, consistent with the autoencoder's $32\times$ latent compression ceiling established in Section 5.2.2. Figure 11 confirms that the inference pipeline preserves event onset timing: the mean onset of generated clips (1.107 s) deviates by only 104 ms from the real mean (1.003 s), validating the effectiveness of the event token conditioning.

Figure 12 provides sample-level qualitative corroboration. Waveform comparisons confirm that generated clips reproduce the sparse impulsive structure of real assembly recordings, with event-guided sampling correctly positioning the transient burst at the conditioned onset time. Mel spectrogram difference maps show that residual energy is concen-

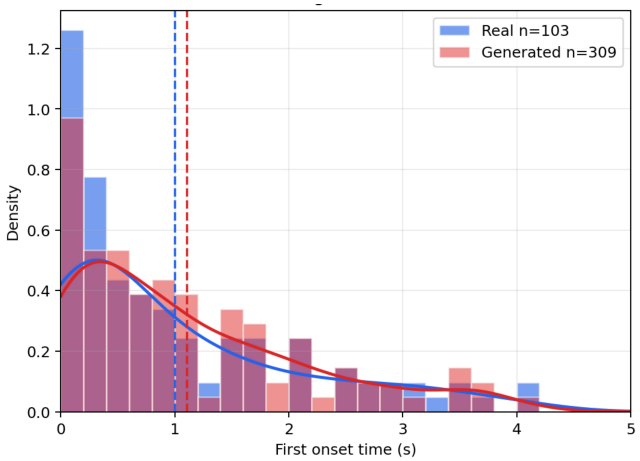

Figure 11. Onset time distribution of real ($n = 103$) and generated ($n = 309$) assembly audio clips. Histograms and kernel density estimates show the distribution of the first detected mechanical event onset across the 5.0 s clip duration. The generated distribution reproduces the real mean onset time closely (real: 1.003 s; generated: 1.107 s, a difference of 104 ms).

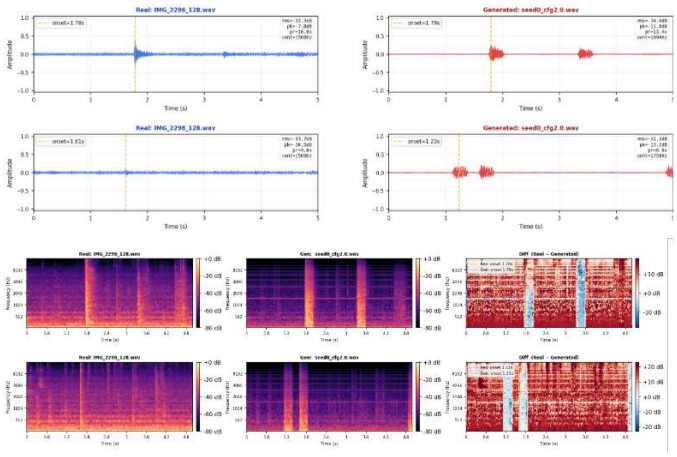

Figure 12. Qualitative evaluation of assembly dataset generation. *Top:* Full-duration waveform comparisons (real, blue; generated, red) for two representative conditioning files. Orange dashed lines mark the detected first event onset. Generated waveforms reproduce the characteristic sparse structure of real assembly recordings, a near-silent background punctuated by a single high-energy transient burst, with onset timing closely matching the real recordings. *Bottom:* Mel spectrogram comparisons (real, generated, and pixel-wise difference map) for the same two pairs.

trated at the impulse onset and in high-frequency bands above 4 kHz, while background regions exhibit near-zero discrepancy.

5.7. Comparison with Waveform-Domain Diffusion Baselines

To position the proposed latent-diffusion framework against waveform-domain alternatives, we benchmark IndusDiff on the heavy-load condition against two baselines trained on the same recordings and evaluated under an identical protocol (PANNs CNN14 Wavegram-LogMel FAD with audio normalised to -14 LUFS): a 1D waveform-domain DDPM (Ho et al., 2020) operating directly on the 44.1 kHz waveform, and DiffWave (Z. Kong et al., 2021), a strong fast-sampling waveform diffusion model. Table 11 reports the results against the real heavy-load reference (RMS $= 0.297$, peak $= 0.662$, centroid $= 4455.06$ Hz).

IndusDiff attains the lowest FAD (6.01), approximately 42% below DiffWave (10.31) and 70% below the 1D waveform diffusion (20.12). The advantage is most pronounced in spectral fidelity: IndusDiff reproduces the real spectral centroid to within 38 Hz (-0.9%), whereas both waveform baselines underestimate it by 840–920 Hz ($\approx$19–21%), reflecting a systematic loss of high-frequency energy and spectral brightness. IndusDiff also matches the real amplitude statistics almost exactly (RMS 0.297, peak 0.665), while DiffWave underestimates both signal energy (RMS 0.241) and dynamic range (peak 0.550). The 1D waveform DDPM preserves coarse amplitude statistics but yields the highest FAD and the largest centroid error, indicating that diffusing directly in the high-dimensional waveform domain on this limited dataset degrades distributional fidelity. Overall, confining diffusion to the phase-aware latent space improves generation fidelity over waveform-domain diffusion rather than trading quality for efficiency.

5.8. Discussion

1. **Generative Performance and Representation Learning.** The proposed latent diffusion framework demonstrates strong capability in generating high-fidelity industrial audio signals that are both perceptually realistic and statistically consistent with the training data distribution. By performing diffusion in a compressed latent space rather than directly in the waveform domain, the model achieves a favorable balance between computational efficiency and generative quality. The phase-aware autoencoder plays a critical role in this pipeline, as it preserves both spectral structure and temporal fine-scale characteristics, which are essential for industrial acoustic signals where subtle deviations often carry diagnostically relevant signatures. The integration of multi-resolution spectral losses and phase-sensitive objectives ensures that the learned latent space retains physically meaningful acoustic features, enabling the downstream diffusion model to operate on a structured and information-rich representation.

2. **Impact of Phase-Aware Loss Design.** A key contribution of this work lies in the formulation of the phase-aware loss function, which combines waveform-domain, spectral, correlation, and high-frequency components. This design addresses a well-known limitation in audio generative models, where magnitude-based optimization alone can result in phase inconsistencies and perceptual artifacts. The inclusion of correlation-based and complex STFT losses improves temporal alignment and phase coherence, which is particularly important for transient-rich industrial signals such as impacts, frictional contacts, and mechanical interactions. Empirically, this results in improved waveform reconstruction and more stable diffusion training, as the latent space encodes both amplitude and phase information more effectively.

3. **Sampling Cost and Fidelity.** Operating in a compressed latent space (16×2064 versus 132,096 waveform samples) lowers the per-step computational cost, which translates directly into improved sampling efficiency. IndusDiff generates one clip in 0.12 s, compared with 0.24 s for the DiffWave fast sampler and 2.20 s for full-step 1D waveform diffusion. IndusDiff is therefore the fastest of the three methods while also achieving the lowest FAD, reducing FAD by approximately 42% relative to DiffWave and 70% relative to the 1D waveform diffusion

Table 11. Benchmark comparison on the heavy-load condition. For RMS, peak, and centroid, values are reported as absolute differences from the real reference, i.e., $|\Delta| = |\text{Generated} - \text{Real}|$; only the magnitude of the deviation is reported, so lower is better. The real reference values are RMS = 0.297, peak = 0.662, and centroid = 4455.06 Hz. For FAD and time, lower values are also better. Runtime values are reported as observed per-clip sampling costs, evaluated on an NVIDIA A100-SXM4-40GB. * denotes CFG (Classifier-Free Guidance) = 1.0, batch size 8, and 50 DDIM steps, while DiffWave used batch size 8 and 6 sampling steps per clip.

Metric	1D Waveform	DiffWave	IndusDiff
Δ RMS	0.0099	0.0558	**0.0000**
Δ Peak	0.0183	0.1118	**0.0030**
Δ Centroid (Hz)	837.06	923.48	**38.34**
FAD $\downarrow$	20.12	10.31	**6.01**
Time (s/clip) $\downarrow$	2.20	0.2355	**0.1193***

baseline. In addition, it provides the most accurate spectral centroid among the compared methods. These results show that the proposed latent diffusion formulation improves efficiency and fidelity simultaneously, rather than trading one for the other. In practice, this makes the approach well suited for large-scale augmentation, generating multi-second, 44.1 kHz clips at approximately 0.12 s per clip while achieving substantially higher fidelity than waveform-domain baselines at lower computational cost.

4. **Generalization Across Heterogeneous Datasets.** The use of two distinct datasets: one publicly available motor dataset and one in-house assembly dataset, provides insight into the generalizability of the framework across different acoustic domains. The model demonstrates the ability to learn both steady-state machine sounds and transient, event-driven acoustic patterns. This suggests that the latent representation captures a broad spectrum of acoustic behaviors. However, generative performance is influenced by dataset size and variability, with more structured datasets yielding more stable and diverse generations.

5.9. Limitations

1. **Limited Dataset Size and Diversity.** A primary limitation of the proposed framework is the relatively small size of the available datasets, particularly the in-house assembly dataset. Diffusion models are inherently data-driven and benefit from large and diverse training distributions. With limited samples, there is a risk of overfitting or reduced diversity in generated outputs. Future work should focus on scaling the dataset size and incorporating more operating conditions and environments.
2. **Dependence on Autoencoder Quality.** The performance of the latent diffusion model is fundamentally constrained by the quality of the learned latent representation. Any loss of information during encoding, particularly in high-frequency components or phase structure, directly impacts the fidelity of generated samples. While the phase-aware loss mitigates this issue, the autoencoder still introduces an information bottleneck.
3. **Mismatch Between Conditioning and Amplitude Dynamics.** Despite mel-spectrogram conditioning, maintaining accurate amplitude scaling and energy distribution remains challenging. This can lead to discrepancies in RMS energy and peak amplitudes between generated and real signals. Incorporating explicit amplitude constraints could improve physical realism.
4. **Computational Cost of Diffusion Training.** Although latent diffusion reduces dimensionality, training still requires many timesteps and epochs. The iterative nature of diffusion increases computational cost, especially for hyperparameter tuning and ablation studies. More efficient diffusion variants remain an important direction.
5. **Evaluation Limitations.** Metrics such as Fréchet Audio Distance (FAD) provide distribution-level assessment but do not fully capture perceptual quality or temporal coherence. Qualitative evaluations are subjective and dataset-dependent. Task-driven evaluation remains necessary for comprehensive validation.
6. **Lack of Explicit Physical Modeling.** The framework is entirely data-driven and does not incorporate physical models of sound generation. While this enables flexibility, it limits interpretability and physical consistency. Hybrid physics-informed approaches could improve robustness.

6. Conclusion and Future Work

This work presents a latent diffusion framework for industrial audio generation, designed to address the challenges of data scarcity and variability in prognostics and health management (PHM) applications. By integrating a phase-aware autoencoder with a latent diffusion model, the proposed approach enables generative learning in a compressed representation that preserves both spectral structure and temporal phase information. The use of multi-resolution and phase-sensitive loss functions ensures that the learned latent space captures diagnostically relevant acoustic features, while the diffusion model leverages this representation to synthesize high-fidelity

audio signals.

Experimental evaluation on two distinct datasets: a publicly available industrial motor dataset and an in-house assembly dataset, demonstrates the effectiveness and versatility of the framework across different acoustic domains. Quantitative results based on Fréchet Audio Distance (FAD), along with qualitative waveform and spectrogram analysis, indicate that the generated samples are both perceptually realistic and statistically consistent with real data. Furthermore, performing diffusion in latent space significantly reduces computational complexity, enabling efficient generation of high-resolution audio suitable for practical deployment.

Overall, the proposed framework establishes a scalable and robust foundation for synthetic industrial audio generation, with direct applicability to data augmentation and downstream PHM tasks.

Several directions can be explored to further enhance the proposed framework. First, scaling the training datasets to include a broader range of operating conditions, machine types, and environmental variations would improve the robustness and diversity of generated samples. Incorporating multisensor data, such as vibration or thermal signals, could also enable multimodal generative modeling for more comprehensive PHM applications.

Second, improvements in latent representation learning could be investigated. While the current autoencoder captures key spectral and temporal characteristics, hierarchical or multiscale latent representations may further enhance the model's ability to represent complex acoustic structures. Additionally, exploring alternative architectures, such as variational or diffusion-based autoencoders, could improve reconstruction fidelity and latent smoothness.

Third, conditioning mechanisms can be extended beyond mel-spectrograms to include explicit physical or operational parameters, such as load, speed, or fault type. This would enable more controllable and interpretable generation, allowing the model to synthesize audio corresponding to specific machine states or failure modes.

Fourth, reducing the computational cost of diffusion remains an important direction. Techniques such as accelerated sampling, consistency models, or reduced-step diffusion processes could significantly improve training and inference efficiency without compromising generation quality.

Fifth, evaluating the downstream benefits of synthetic audio is an important next step toward assessing IndusDiff's practical utility. The present study focuses on generation quality, providing a foundation for subsequent evaluation in fault classification and anomaly detection. Ongoing work examines an operating-condition classifier trained on progressively smaller subsets of real data, with and without IndusDiff augmentation, measured against a held-out real test set.

Taken together, these directions point toward the broader role of diffusion-based models as scalable data augmentation tools for PHM. By synthesizing diverse, realistic machine sounds across operating conditions, the framework offers a route to improved model robustness in data-limited environments. This work bridges a critical gap between modern generative modeling techniques and the practical requirements of industrial PHM systems, establishing a foundation for future research in physics-aware generative modeling and multimodal data augmentation for intelligent maintenance.

ACKNOWLEDGMENT

The authors would like to thank Siemens Corporation for their generous support in funding this project.

REFERENCES

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., & Chan, W. (2021). WaveGrad: Estimating gradients for waveform generation. In *Proceedings of the international conference on learning representations (iclr).*

Choi, H.-S., Kim, J.-H., Huh, J., Kim, A., Ha, J.-W., & Lee, K. (2019). Phase-aware speech enhancement with deep complex U-Net. In *Proceedings of the international conference on learning representations (iclr).*

Chung, Y., Lee, J., & Nam, J. (2024). T-FOLEY: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis. In *Proceedings of the ieee international conference on acoustics, speech and signal processing (icassp).* doi: 10.1109/ICASSP48485.2024.10447898

Ding, Y., Ma, L., Ma, J., Wang, M., & Lu, C. (2023). A survey on deep learning for machinery fault diagnosis and prognosis. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–20.

Evans, Z., Parker, J. D., Simon, C. J., Carr, C. J., Zukowski, Z., & Abela, J. (2024). Stable audio open. *arXiv preprint arXiv:2407.14358*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems (neurips)* (Vol. 27, pp. 2672–2680).

Griffin, D. W., & Lim, J. S. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236–243. doi: 10.1109/TASSP.1984.1164317

Grollmisch, S., & Abeßer, J. (2021). Audio-based machine fault diagnosis with WaveNet and data augmentation. In *Proceedings of the 9th european workshop on structural health monitoring (ewshm).*

Grollmisch, S., Abeßer, J., Liebetrau, J., & Lukashevich, H.

(2019). Sounding industry: Challenges and datasets for industrial sound analysis. In *Proceedings of the 27th european signal processing conference (eusipco)* (pp. 1–5). doi: 10.23919/EUSIPCO.2019.8902941

Gui, A., Gamper, H., Braun, S., & Emmanouilidou, D. (2024). Adapting Fréchet audio distance for generative music evaluation. In *Proceedings of the ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1331–1335). doi: 10.1109/ICASSP48485.2024.10446663

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., … Wilson, K. (2017). CNN architectures for large-scale audio classification. In *Proceedings of the ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131–135). doi: 10.1109/ICASSP.2017.7952132

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in neural information processing systems (neurips)* (Vol. 33, pp. 6840–6851).

Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Kilgour, K., Zuluaga, M., Roblek, D., & Sharifi, M. (2019). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of interspeech 2019* (pp. 2350–2354). doi: 10.21437/Interspeech.2019-2219

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the international conference on learning representations (iclr).*

Koizumi, Y., Kawaguchi, Y., Imoto, K., Nakamura, T., Niitsuma, Y., Tanabe, R., … Harada, N. (2020). Description and discussion on DCASE 2020 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring. In *Proceedings of the detection and classification of acoustic scenes and events workshop (dcase)* (pp. 1–6).

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in neural information processing systems (neurips)* (Vol. 33, pp. 17022–17033).

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2880–2894. doi: 10.1109/TASLP.2020.3030497

Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2021). DiffWave: A versatile diffusion model for audio synthesis. In *Proceedings of the international conference on learning representations (iclr).*

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., … Courville, A. (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis. In *Advances in neural information processing systems (neurips)* (Vol. 32).

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, *104*, 799–834.

Li, X., Ding, Q., & Sun, J.-Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, *172*, 1–11.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., … Plumbley, M. D. (2023). AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the international conference on machine learning (icml)* (pp. 21450–21474).

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., … Plumbley, M. D. (2024). AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 2871–2883.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., & Zhu, J. (2022). DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in neural information processing systems (neurips)* (Vol. 35, pp. 5775–5787).

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., … Bengio, Y. (2017). SampleRNN: An unconditional end-to-end neural audio generation model. In *Proceedings of the international conference on learning representations (iclr).*

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International conference on machine learning (icml)* (pp. 8162–8171).

Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual reasoning with a general conditioning layer. In *Aaai conference on artificial intelligence.*

Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., & Kawaguchi, Y. (2019). MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In *Proceedings of the detection and classification of acoustic scenes and events workshop (dcase)* (pp. 1–6).

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi: 10.1109/5.18626

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*(1), 72–83. doi:

10.1109/89.365379

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 10684–10695).

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, *24*(3), 279–283. doi: 10.1109/LSP.2017.2657381

Salimans, T., & Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. In *International conference on learning representations (iclr)*.

Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. In *Proceedings of the international conference on learning representations (iclr)*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *Proceedings of the international conference on learning representations (iclr)*.

Tailleur, M., Lee, J., Lagrange, M., Choi, K., Heller, L. M., Imoto, K., & Okamoto, Y. (2024). Correlation of Fréchet audio distance with human perception of environmental audio is embedding dependent. In *Proceedings of the 32nd european signal processing conference (EUSIPCO)*. Lyon, France. doi: 10.48550/arXiv.2403.17508

Takaki, S., Nakashika, T., Wang, X., & Yamagishi, J. (2019). STFT spectral loss for training a neural speech waveform model. In *Ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7065–7069).

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., … Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Yamamoto, R., Song, E., & Kim, J.-M. (2020). Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. of icassp* (pp. 6199–6203).

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, *115*, 213–237.

Ziyin, L., Hartwig, T., & Ueda, M. (2020). Neural networks fail to learn periodic functions and how to fix it. In *Advances in neural information processing systems (neurips)* (Vol. 33, pp. 1583–1594).

BIOGRAPHIES

Muhammad Areeb received the Bachelors of engineering in Mechanical Engineering from NED University of Engineering and Technology, Karachi, Pakistan, in 2020, and the M.S. degree in Mechanical Engineering from the University of Illinois Chicago (UIC) in 2025. He is currently pursuing a Ph.D. in Industrial Engineering and Operations Research at UIC. He is a Graduate Research Assistant and Siemens Fellow in the Industrial AI and PHM Lab at UIC, where his research focuses on generative modeling for industrial audio and multimodal fault detection.

Jida Huang is an Assistant Professor in the Department of Mechanical and Industrial Engineering at the University of Illinois Chicago, where he directs the Design Reasoning and Evolution for Advanced Manufacturing (DREAM) lab. DREAM lab focuses on geometric data analytics and design informatics for advanced materials and structures to elucidate materials architecture-process-property relation spanning the nano-, micro-, and macro-scales. DREAM lab develops computational approaches for design and fabrication problems occurring in the entire product lifecycle, from early-stage conceptual design to design modeling, simulation, realization, and final fabrication, with the mission of tackling societal challenges in healthcare, energy, and climate change. Dr. Huang received his Ph.D. in Industrial and Systems Engineering in 2019 from the University at Buffalo, State University of New York, where he received presidential fellowship. He also received NSF CAREER Award and the best paper award at ASME IDETC/CIE 2021.

Miao He received a Ph.D. in Industrial Engineering and Operations Research from University of Illinois at Chicago, a B.E. degree in Vehicle Engineering from University of Science and Technology Beijing, Beijing, China and a M.S. degree in Mechanical Engineering from Purdue University, Hammond, IN. She is currently a senior research scientist at Siemens Corporation. Her research interests include digital signal processing, machinery health monitoring and fault diagnosis, prognostics, artificial neural network applications, and edge computing on factory automation.

David He received a Ph.D. in Industrial Engineering from The University of Iowa. Dr. He is a Professor and Director of the Industrial AI and PHM Laboratory in the Department of Mechanical and Industrial Engineering at The University of Illinois-Chicago. Dr. He is also a Fellow of the Prognostics and Health Management (PHM) Society. Dr. He's research areas include PHM, Industrial AI, smart manufacturing systems modeling and analysis, quality and relia-

bility engineering.

APPENDIX

The appendix includes additional results showing the AE reconstruction quality in Figure 13, Comparison of generated and real audio signal waveform form of the engine data in Figure 14 and latent diffusion generation quality elucidating spectral centroid distribution and peak amplitude distribution in Figure 15.

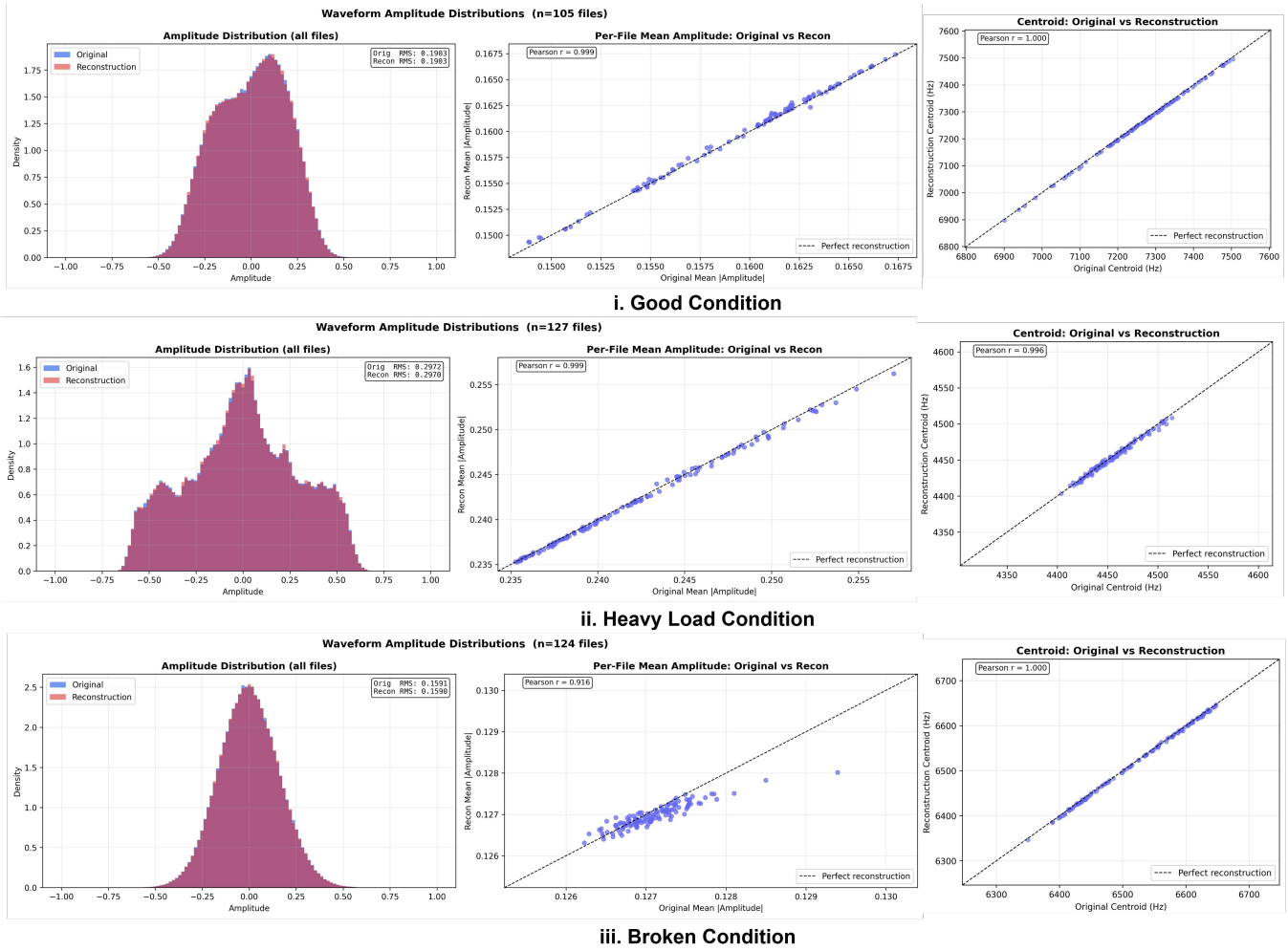

Figure 13. Quantitative evaluation of autoencoder reconstruction fidelity across three operating conditions (rows). Left: waveform amplitude distributions for original and reconstructed signals, showing near-complete overlap and preservation of global amplitude statistics (RMS). Middle: per-file mean amplitude, indicating strong linear agreement between the two. Right: per-file spectral centroid, confirming that spectral balance is preserved across the full range of each condition.

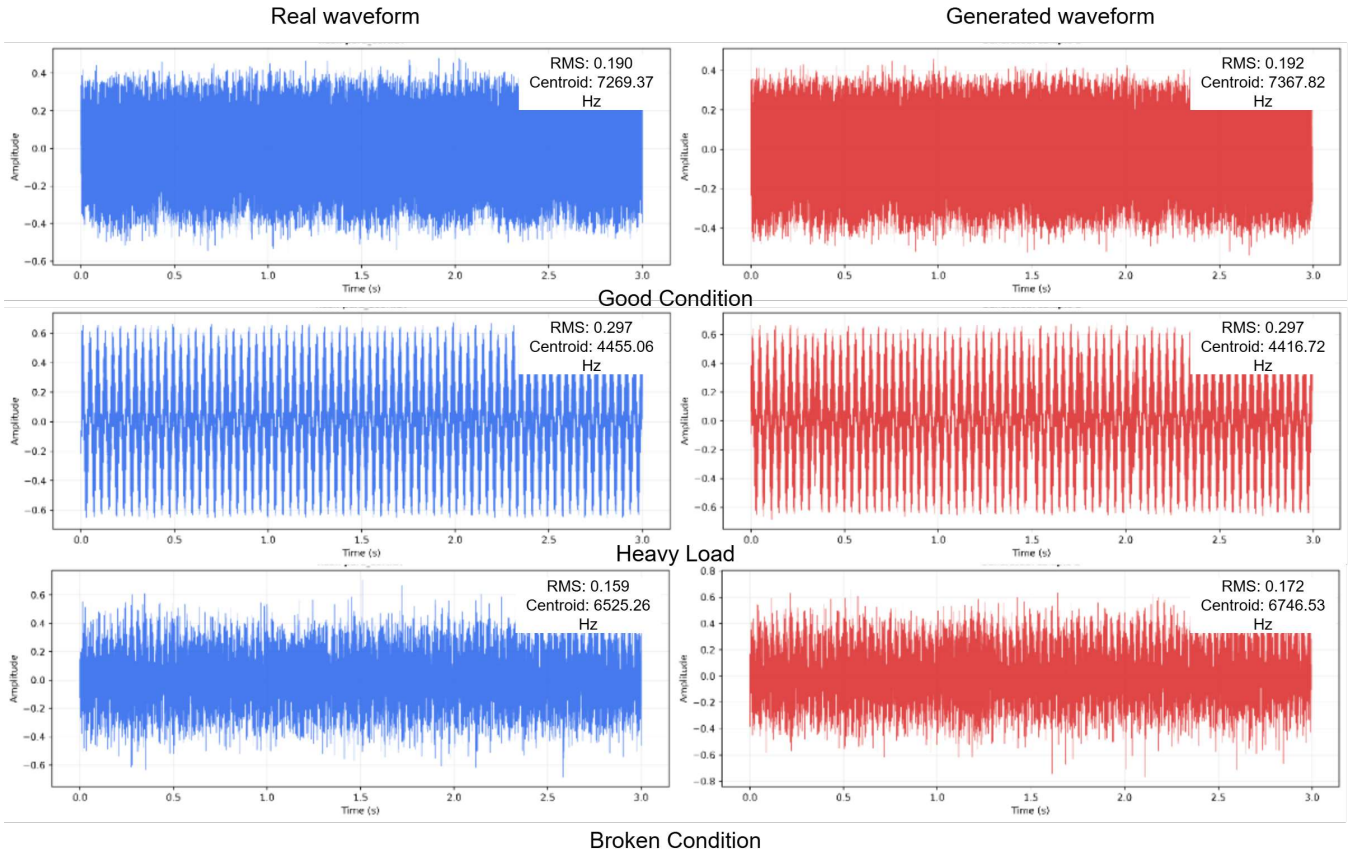

Figure 14. A comparison of real and generated waveform for the three different operating condition of SI dataset.

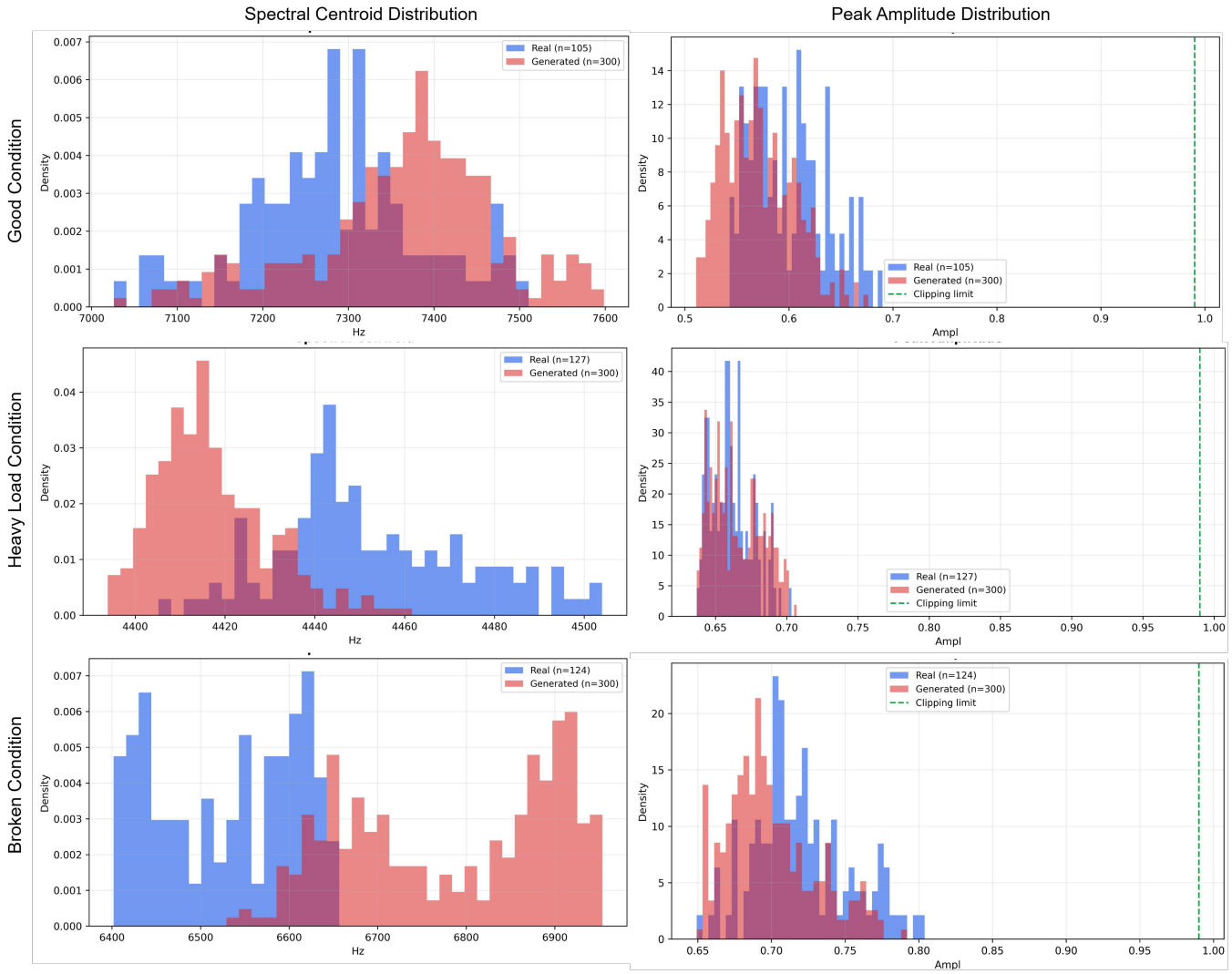

Figure 15. Distributional comparison between real and generated audio samples across three datasets. For each operating condition (rows), the left column shows the distribution of spectral centroid values (Hz), while the right column presents peak amplitude distributions. Blue histograms correspond to real data and red to generated samples.