WaveFM 🌊: A Wavelet-Native Pretrained Model for Industrial Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Time-series foundation models have largely converged on a recipe of fixed-length time-domain patches and parameter counts in the hundreds of millions. On industrial telemetry—non-stationary and dominated by transients and harmonics—patching averages together the high-frequency structure that drives operational decisions. We introduce **WaveFM**, a 10 M-parameter pretrained industrial forecasting model that replaces time-domain patching with tokens over a 5-level discrete wavelet decomposition, exposing six dyadic frequency bands to band-specific processing and output heads. A latent bottleneck shortens the cross-channel sequence from $97C$ to $25C$ tokens ($\approx 15\times$ fewer token-pair interactions). On 15 industrial and public datasets, zero-shot WaveFM achieves accuracy comparable to Chronos-2-Full (0.301 vs. 0.302 MSE; geometric-mean relative MSE 0.995, 95% CI $[0.983, 1.008]$), including on the seven public datasets alone (0.323 vs. 0.322), with one-twelfth the parameters and $13.8\times$ faster multivariate inference (2.5 ms per window). Same-corpus, parameter-matched controls show that pretraining data alone does not reproduce this result: PatchTST trained on the same corpus reaches 0.353. On high-band-energy windows and around detected transients, WaveFM's error is 6.2% and 7.6% lower than Chronos-2-Full's, with the largest gains in the finest detail bands. WaveFM trails Chronos-2-Full on a GIFT-Eval tier outside its pretraining domain (0.319 vs. 0.302) and at longer horizons.

1 Introduction

Time-series foundation models have rapidly converged on a common recipe: large transformer backbones, fixed-length time-domain patches, and pre-training corpora of 10^{10}–10^{11} observations (Das et al., 2024; Ansari et al., 2025; Liu et al., 2025a; Shi et al., 2025; Goswami et al., 2024). On standard academic benchmarks, scaling this recipe has produced steady gains. Industrial telemetry is different: historians in process plants, power grids, and rotating-equipment fleets record sub-second signals that are non-stationary and dominated by transient and harmonic content (Randall, 2011). On data of this kind, patched models tend toward smooth, mean-reverting forecasts that miss the events operators care about. This paper asks whether a small model with a physically motivated tokenizer can match these multi-100 M baselines on industrial workloads, and at what cost.

The patching paradox. Time-domain patching has become the de-facto tokenization for long-context transformers (Nie et al., 2023; Liu et al., 2025b): it widens the local receptive field while reducing sequence length. Averaging over fixed-length windows, however, acts as an implicit low-pass filter that attenuates the sharp, high-frequency transients—bearing impulses, arc flashes, pressure spikes—that are the earliest precursors of mechanical degradation. Standard point-wise objectives (MSE) compound the problem by rewarding mean-reverting forecasts that score well numerically while obscuring the events the model is deployed to detect.

Why wavelets, not Fourier. Reliable industrial forecasting demands joint time–frequency localization. In rotating machinery, imbalance manifests at the running frequency ($1\times$), misalignment at twice ($2\times$), and bearing defects at geometry-dependent characteristic frequencies (Randall, 2011). Time-domain patching can mix these scales within a single token, and the Fourier transform trades temporal resolution for spectral resolution, smearing transients across the analysis window. The discrete wavelet transform (DWT) instead yields a multi-resolution representation that keeps these scales

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

separately addressable and localizes *when* a high-frequency event occurs relative to the long-term trend. WaveFM applies a level-5 biorthogonal (bior2.2) DWT to both context and target, yielding six dyadic bands from slow trend ($\sim f_s/64$) to fine detail ($\sim f_s/2$). The tokenizer is therefore physically motivated; WaveFM does not incorporate governing equations or explicit physical constraints.

Contributions.

1. **Wavelet-native tokenization with conditioned band heads.** We replace time-domain patches with a 5-level DWT vocabulary of six frequency bands, processed through band-specific output heads with a learned per-sample gate that blends a linear baseline with the transformer's residual correction, trained end-to-end through a differentiable inverse DWT.
2. **Latent-bottleneck cross-variate attention.** Compressing each channel's 96 band tokens into $Q{=}24$ latent queries (plus CLS) shortens the cross-channel sequence from $97C$ to $25C$ tokens, giving $\approx 15\times$ fewer token-pair interactions. This is a practical reduction: cross-channel attention remains quadratic in the number of channels C.
3. **A controlled evaluation of accuracy and cost.** We report public-first per-dataset results with scale-robust paired statistics, same-corpus parameter-matched controls, a domain-adapted Chronos-2 comparison, and direct measurements of error on high-band-energy windows, around transient events, and per wavelet band.

WaveFM achieves industrial forecasting accuracy comparable to Chronos-2-Full zero-shot with one-twelfth the parameters and $13.8\times$ faster multivariate inference on a single workstation (§5). An optional retrieval component, the Wavelet Profile Bank, is described and evaluated separately in App. B; it is not used in any headline result.

2 Related Work

Foundation models for time series. A first generation of time-series foundation models has converged on large-scale pre-training over billions of observations with transformer or hybrid backbones. Google's TimesFM (Das et al., 2024) and Salesforce's Moirai (Woo et al., 2024; Liu et al., 2025a) use patched decoders; Amazon's Chronos and Chronos-2 (Ansari et al., 2024; 2025) cast forecasting as token-level language modeling; MOMENT (Goswami et al., 2024) and OneFitsAll (Zhou et al., 2023) repurpose pretrained language backbones; TiRex (Auer et al., 2025) and xLSTM-based variants (Beck et al., 2024) replace attention with recurrent state-tracking; and Time-MoE (Shi et al., 2025) scales to billions of parameters with sparse experts. Moirai 2.0 (Liu et al., 2025a) shows that compact models (roughly 11 M parameters) can be competitive; we do not benchmark it here. Recent industrial deployments—Datadog's Toto (Cohen et al., 2025) and Cisco TSM (Gou et al., 2025)—explicitly target observability and telemetry. WaveFM is positioned in the *small* regime ($\sim$10 M parameters) and trades domain breadth for a physically motivated wavelet tokenizer.

Patching and its inductive bias. PatchTST (Nie et al., 2023) established fixed-length time-domain patches as the dominant tokenization for long-context forecasting, and Timer-XL (Liu et al., 2025b) extended this to multi-thousand-step contexts. Patching reduces sequence length and widens the effective receptive field, but a stride-P patch also mixes the highest $\log_2 P$ octaves of the spectrum into a single token. Multi-resolution input streams (Gou et al., 2025) mitigate but do not remove this bias, and TSMixer (Chen et al., 2023) shows that linear time-domain mixers can match patched transformers on smooth benchmarks.

Frequency- and wavelet-domain representations. Frequency-aware architectures predate foundation models: FEDformer (Zhou et al., 2022) attends in the Fourier domain, and TimesNet (Wu et al., 2023) reshapes sequences into 2D maps along dominant periods; we train both on our corpus as controls. The closest precedent to our tokenizer is WaveToken (Masserano et al., 2025), which quantizes wavelet coefficients into a discrete vocabulary for language-model consumption. WaveFM differs in three respects: (i) a 5-level decomposition processed as six explicit band streams; (ii) continuous-valued DWT token embeddings, avoiding discretization error; and (iii) frequency-aware rotary embeddings whose base period scales with the band.

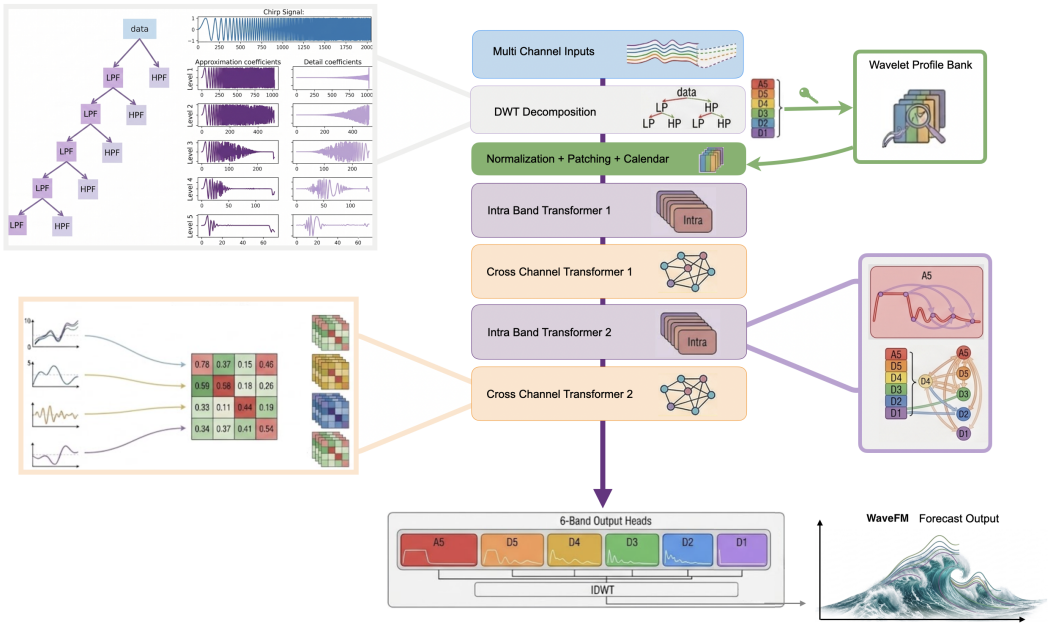


Figure 1: WaveFM architecture overview. A multi-level DWT decomposes each channel into six frequency bands, which are tokenized and processed by dual-phase attention with a latent bottleneck. Predicted wavelet coefficients are reconstructed to the time domain via a differentiable inverse DWT.

Cross-variate modeling. Channel-mixing strategies range from fully channel-independent (PatchTST) to fully cross-attentive (iTransformer (Liu et al., 2024)). Cross-variate attention scales as $\mathcal{O}(C^2)$ in the channel dimension. Chronos-2's group attention (Ansari et al., 2025) and Moirai-2's variate sub-sampling (Liu et al., 2025a) truncate the variate context. WaveFM instead compresses each channel to a fixed 25-token summary before cross-channel mixing, decoupling cross-variate cost from the per-band token count; the cost remains quadratic in C.

Retrieval and in-context adaptation. RAFT (Han et al., 2025), TimeRAF (Zhang et al., 2024), and NV-Tesseract (Gautam et al., 2025) retrieve similar historical windows and condition predictions on them; TimesFM-ICF (Faw et al., 2025) places related series in the prompt, and ChatTime (Wang et al., 2024) fuses text and series. WaveFM's optional Profile Bank (App. B) follows the same retrieve-then-attend skeleton with wavelet-domain CLS keys; we have not compared these keys against time-domain retrieval. App. D.1 (Table 9) summarises the design space.

3 WaveFM Model

WaveFM tokenizes each channel's signal directly in the wavelet domain via a 5-level DWT, then processes the resulting six-band representation with interleaved intra-/cross-channel attention. Unlike patch-based tokenizers (Ansari et al., 2025; Das et al., 2024), this gives each attention head access to physically distinct frequency scales. Figure 1 gives an overview.

3.1 Wavelet-Based Tokenization

Given a multivariate input $\mathbf{X} \in \mathbb{R}^{C \times T}$ with C channels and context length $T{=}1536$, WaveFM first applies a biorthogonal wavelet decomposition (bior2.2, level $L{=}5$) independently to each channel. This yields six coefficient sequences per channel:

$$\{c_A^{(5)},\ c_D^{(5)},\ c_D^{(4)},\ c_D^{(3)},\ c_D^{(2)},\ c_D^{(1)}\} = \mathrm{DWT}_{L=5}(\mathbf{x}_c), \tag{1}$$

where $c_A^{(5)}$ contains the coarsest trend and each $c_D^{(l)}$ captures progressively finer oscillations in the dyadic frequency band $[f_s/2^{l+1},\ f_s/2^l]$ relative to the sampling rate f_s. Each band b is mapped to

a fixed-size token sequence by a per-band Conv1d layer with band-specific kernel size and stride, producing exactly $P{=}16$ tokens regardless of the coefficient length:

$$\mathbf{T}_b = \mathrm{Conv1d}(\mathbf{c}_b;\ \mathbf{W}_b) + \mathbf{e}^{(b)}_{\mathrm{band}} + \mathbf{e}_{\mathrm{pos}}, \qquad \mathbf{T}_b \in \mathbb{R}^{16\times d}, \tag{2}$$

where $\mathbf{e}^{(b)}_{\mathrm{band}} \in \mathbb{R}^d$ identifies the frequency scale and $\mathbf{e}_{\mathrm{pos}} \in \mathbb{R}^{16\times d}$ encodes position within the band. A learnable `[CLS]` token $\mathbf{t}_{\mathrm{cls}} \in \mathbb{R}^d$ is prepended, giving a per-channel sequence of $6{\times}16 + 1 = 97$ tokens at dimension $d{=}192$. Per-band coefficient counts, frequency ranges, and design choices are listed in App. A.3 (Table 6).

3.2 Frequency-Aware Rotary Position Embeddings

Standard RoPE applies a single base frequency θ_{base} to all tokens. In WaveFM the 97-token sequence mixes six bands whose tokens represent vastly different time scales: 16 tokens spanning the entire trend (A5) versus 16 tokens capturing the finest detail (D1). We therefore introduce *frequency-aware RoPE*, which scales the base frequency by a per-band multiplier m_b:

$$\theta_{b,i} = \frac{\mathrm{pos}}{(\theta_{\mathrm{base}} \cdot m_b)^{\,2i/d_k}}, \qquad m_b \in \{1, 1, 2, 4, 8, 16\}. \tag{3}$$

Larger m_b slows the rotation of the position signal, reducing positional distinction for high-frequency bands where token content is more informative than token position. These multipliers are applied only in the intra-channel phases; cross-channel attention uses no positional encoding.

3.3 Dual-Phase Attention with Latent Bottleneck

Naïvely applying self-attention across all channels and frequency tokens would yield a sequence of length $97C$. WaveFM instead uses an *interleaved dual-phase* design: two rounds of intra-channel attention alternated with two rounds of cross-channel attention, connected by latent bottleneck compression.

Intra-channel phase. Each phase consists of $N_{\mathrm{intra}}{=}4$ pre-norm transformer blocks operating independently per channel on the 97-token sequence, with RMSNorm, multi-head attention (4 heads, $d_k{=}48$) using frequency-aware RoPE, and a $d \to 4d \to d$ MLP (App. A.1).

Latent bottleneck. $Q{=}24$ learnable latent queries $\mathbf{Z} \in \mathbb{R}^{Q\times d}$ cross-attend to the 96 band tokens (CLS excluded):

$$\mathbf{Z}' = \mathrm{softmax}\Big(\frac{\mathbf{Z}\mathbf{W}_Q\,(\mathbf{T}_{\mathrm{bands}}\mathbf{W}_K)^\top}{\sqrt{d_k}}\Big)\,\mathbf{T}_{\mathrm{bands}}\mathbf{W}_V, \tag{4}$$

followed by a residual MLP. Prepending an optional calendar-gated CLS (App. A.2) produces a compressed summary $\mathbf{S} \in \mathbb{R}^{25\times d}$ per channel (24 latents plus CLS). This shortens the cross-channel sequence from $97C$ to $25C$ tokens, so cross-channel attention costs $\mathcal{O}((25C)^2 d)$ rather than $\mathcal{O}((97C)^2 d)$: $(97/25)^2 \approx 15.1\times$ fewer token-pair interactions. This is a practical reduction, not a different complexity class; the cost remains quadratic in C.

Cross-channel mixing. A learnable channel-identity embedding $\mathbf{e}_{\mathrm{ch}}(c)$ is added to each channel's summary. All channels' summaries are concatenated into a single sequence of length $25C$ and processed by $N_{\mathrm{cross}}{=}3$ transformer blocks without positional encoding, enabling the model to learn multivariate dependencies (e.g., current $\leftrightarrow$ torque $\leftrightarrow$ vibration).

Interleaving. The phases are interleaved as Intra $\to$ Cross $\to$ Intra $\to$ Cross. After the first cross-channel pass we re-inject only the updated `[CLS]` into the original 97-token sequence; the 96 fine-grained band tokens bypass the bottleneck via a skip connection, so high-frequency detail is never compressed.

3.4 Conditioned Output Heads

Each of the six frequency bands has a dedicated output head. The head input concatenates the final CLS token with the mean-pooled band tokens from the second intra-channel phase: $\mathbf{h}_b =$

$[\mathbf{t}_{\text{cls}}^{\text{final}};\, \overline{\mathbf{T}}_b] \in \mathbb{R}^{2d}$; exact positional detail is carried by the baseline path of Eq. 5. Every head follows a *baseline-plus-correction* paradigm. A direct linear projection of the input wavelet coefficients provides a baseline prediction that bypasses the transformer entirely:

$$\hat{\mathbf{y}}_b^{\text{base}} = \mathbf{W}_b^{\text{base}}\, \mathbf{c}_b^{\text{ctx}}, \qquad \mathbf{W}_b^{\text{base}} \in \mathbb{R}^{L_b^{\text{tgt}} \times L_b^{\text{ctx}}}. \tag{5}$$

The transformer path computes a learned correction $\boldsymbol{\delta}_b$ through an MLP applied to $\mathbf{h}_b$. For the two coarsest bands (A5, D5) we use a deeper three-layer MLP with an additive skip path, $\boldsymbol{\delta}_b = \text{MLP}_{\text{deep}}(\mathbf{h}_b) + \mathbf{W}_s \mathbf{h}_b$; the remaining bands use a two-layer MLP. A *conditioned gate* controls the per-sample blend between baseline and correction:

$$g_b = \sigma\big(\text{MLP}_{64}(\mathbf{h}_b)\big), \qquad \hat{\mathbf{y}}_b^{\text{param}} = \hat{\mathbf{y}}_b^{\text{base}} + g_b \cdot \boldsymbol{\delta}_b, \tag{6}$$

where $g_b \in (0, 1)$ is a scalar gate. When the input is well-predicted by linear extrapolation the gate learns to stay near zero; for complex non-linear dynamics it opens to incorporate the transformer's correction. Predicted band coefficients are reassembled into a time-domain forecast through precomputed bior2.2 synthesis matrices, and the model is trained with a weighted per-band log-cosh loss plus a time-domain reconstruction term (App. A.4). Because the head dimensions depend on the target coefficient lengths, the heads are specific to the forecast horizon.

Optional retrieval. WaveFM can optionally be paired with a *Wavelet Profile Bank* that retrieves training-window precedents at inference time and fuses their target coefficients into the heads through a separate retrieval gate. The bank is not used in any headline result (Tables 1–5); its construction, gate composition, leakage controls, and a controlled evaluation are given in App. B.

4 Experiments

4.1 Setup

Pre-training corpus. WaveFM is pre-trained on a 53-dataset industrial time-series corpus (App. C) that combines proprietary high-fidelity telemetry, used with explicit customer consent, with open-access corpora including Time-300B (Maple728, 2024; Shi et al., 2025) and the Monash archive (Godahewa et al., 2021). The mix spans $\sim 10^8$ windows covering power, transportation, renewables, manufacturing, process industries, and rotating machinery.

Contamination controls. ETT, Solar-AL, all GIFT-Eval subsets, and all held-out industrial and proprietary evaluation assets are excluded from the pre-training corpus. The Monash archive contributes other series to pre-training; its overlap with our Commercial-Bldg Load and Residential Demand evaluation sets is at the archive level only, and the exact evaluation series and time ranges were removed before window generation.

Evaluation benchmarks. The industrial tier comprises 15 datasets. Seven are public: four Electricity Transformer datasets (ETTh1, ETTh2, ETTm1, ETTm2; 7 channels, hourly/15-min) (Zhou et al., 2021), Solar-AL (11 channels, 10-min) (Lai et al., 2018), and two load benchmarks derived from the Monash *Australian Electricity Demand* archive (16 per-region meters stacked into a multivariate frame, hourly) (Godahewa et al., 2021), reported as *Commercial-Bldg Load* and *Residential Demand*. Eight are proprietary: two *High-Frequency Vibration Buses* (HF-VB; 9 and 12 acceleration channels at 10 Hz) from a centrifugal pump skid and a reciprocating compressor; two 10 Hz vehicle-telemetry buses (16 channels each), *Drivetrain Telem.* and *Chassis Motion*; and four process-industry assets, *Distillation Col.*, *Heat Exchanger*, *Reactor Loop*, and *Compressor Train*. A second tier averages 5 high-frequency subsets of GIFT-Eval (Aksu et al., 2024), reported as a single tier-level row. All datasets are split chronologically 70/15/15 into train/validation/test; values are z-score normalised using training-split statistics, and all reported numbers are on the test split.

Baselines. We compare against five publicly available foundation models, all evaluated zero-shot—**Chronos-2-Small** (28 M) (Ansari et al., 2025), **Chronos-2-Full** (120 M), **TimesFM 2.5** (200 M) (Das et al., 2024), **Toto-Open-Base** (151 M) (Cohen et al., 2025), and **TiRex** (35 M) (Auer et al., 2025).[1]

[1] Chronos-2-S/F: tokenized multivariate, 8,192-step context. TimesFM 2.5: patched univariate, 2,048-step context. Toto-Open-Base: observability-oriented transformer with full $\mathcal{O}(C^2)$ cross-channel attention. TiRex: xLSTM-based recurrent forecaster.

To separate architecture from pre-training data we add *same-corpus controls*: PatchTST (Nie et al., 2023) (10 M), TimesNet (Wu et al., 2023) (11 M), and FEDformer (Zhou et al., 2022) (25 M), trained on WaveFM's 53-dataset corpus with the same context length, horizon, chronological splits, and evaluation protocol; and *Chronos-2-Full fine-tuned* on the same corpus with AutoGluon's documented Chronos-2 workflow (learning rate 10^{-4}, 2,000 steps, batch size 32).

Protocol. All models forecast $H{=}96$ steps in a single forward pass using rolling test windows with stride 96. WaveFM uses a context of $T{=}1{,}536$; baselines use their maximum context or $T{=}1{,}536$, whichever is smaller. WaveFM and the pretrained baselines are evaluated zero-shot: no Profile Bank, no fine-tuning, and no test-time fitting. We report MSE and MAE in z-score space, averaged over all windows and channels per dataset. The industrial result is the simple mean over all 15 datasets; the overall result is the unweighted mean of the industrial-tier and GIFT-Eval-tier averages. Latency is the median of 32 rolling-window forecasts after 3 warm-up runs on an Apple M5 Max workstation (40-core GPU, 128 GB unified memory).

5 Results

5.1 Main Results

Table 1 reports zero-shot MSE on all 15 industrial datasets, public datasets first. WaveFM (0.301) and Chronos-2-Full (0.302) achieve comparable industrial-tier accuracy. On the seven public datasets the two are a near-tie (0.323 vs. 0.322), so the public evidence supports the efficiency claim more strongly than any accuracy advantage; on the eight proprietary datasets WaveFM is 0.282 vs. 0.284. WaveFM has the lowest MSE on 9 of the 15 datasets; Chronos-2-Full is lowest on four (ETTm1, Commercial-Bldg Load, Drivetrain Telem., and Reactor Loop), TimesFM 2.5 on ETTh1, and Chronos-2-Small on HF-VB Pump. The remaining zero-shot baselines have higher industrial means (TimesFM 2.5 0.310, Chronos-2-Small 0.314, TiRex 0.326, Toto-Open 0.338); the WaveFM–Chronos-2-Full difference is within noise (§5.2).

On the overall average WaveFM is second (0.310 vs. 0.302, $+2.6\%$), and the gap comes entirely from the GIFT-Eval tier (0.319 vs. 0.302), where WaveFM does not dominate every subset categorized as high-frequency. WaveFM's pre-training corpus is deliberately industrial and contains essentially no retail-demand, finance, or web-traffic series; we hypothesize that this domain mismatch, rather than the tokenizer, explains the gap, but we have not tested this directly.

5.2 Statistical Comparison

Because neighbouring rolling windows overlap heavily, we treat datasets, not windows, as paired units. For each dataset i we compute $r_i = \log(m_i^{\mathrm{WF}}/m_i^{\mathrm{B}})$ from the MSE of WaveFM and baseline B, and report the geometric-mean relative MSE $R = \exp(\frac{1}{n}\sum_i r_i)$; $R < 1$ favours WaveFM (Table 2). WaveFM and Chronos-2-Full zero-shot are not significantly different, on all 15 datasets or on the seven public ones; fine-tuned Chronos-2-Full has a small, non-significant numerical advantage. We do not interpret non-significance as formal equivalence. Against the same-corpus PatchTST control, WaveFM has 14.4% lower geometric-mean MSE and wins on 14 of 15 datasets; the public-only comparison is favourable but not significant. An absolute-difference paired test is reported in App. D.5.

5.3 Controlling for the Pre-training Corpus

WaveFM is pre-trained on industrial data while the foundation-model baselines are zero-shot, so part of its advantage could come from the corpus rather than the architecture. Table 3 separates the two. Same-corpus models at comparable parameter budgets do not reproduce WaveFM's accuracy: PatchTST, TimesNet, and FEDformer are 17.3%, 18.9%, and 21.9% above WaveFM on the industrial mean, while WaveFM runs 4.4–$8.4\times$ faster. Domain adaptation also matters: fine-tuning Chronos-2-Full on the same corpus improves it from 0.302 to 0.298, slightly ahead of WaveFM, while retaining a 120 M backbone and 36 ms latency ($12\times$ larger and $14.4\times$ slower than WaveFM). On public data WaveFM remains ahead of fine-tuned Chronos-2-Full (0.323 vs. 0.326). We therefore do not attribute WaveFM's result to the tokenizer alone.

Table 1: Main results: zero-shot MSE in z-score space (H=96, T=1,536, stride 96; no Profile Bank). **Bold**: best per row; underline: second best. Group means use unrounded values.

Dataset (C, f_s)	WaveFM	Chronos-2-S	Chronos-2-F	TimesFM 2.5	Toto-Open	TiRex
Public datasets (7)						
ETTh1 (7, 1h)	0.378	0.385	0.371	**0.368**	0.418	0.395
ETTh2 (7, 1h)	**0.272**	0.290	0.275	0.278	0.325	0.319
ETTm1 (7, 15min)	0.305	0.312	**0.295**	0.300	0.358	0.346
ETTm2 (7, 15min)	**0.158**	0.172	0.160	0.165	0.192	0.183
Solar-AL (11, 10min)	**0.205**	0.208	0.213	0.220	0.247	0.235
Commercial-Bldg Load (16, 1h)	0.452	0.470	**0.445**	0.465	0.495	0.478
Residential Demand (16, 1h)	**0.488**	0.510	0.495	0.492	0.528	0.515
Proprietary datasets (8)						
HF-VB Pump (9, 10Hz)	0.292	**0.288**	0.296	0.302	0.320	0.310
HF-VB Compressor (12, 10Hz)	**0.312**	0.330	0.318	0.328	0.345	0.335
Drivetrain Telem. (16, 10Hz)	0.252	0.258	**0.245**	0.265	0.280	0.270
Chassis Motion (16, 10Hz)	**0.355**	0.378	0.362	0.372	0.395	0.380
Distillation Col. (24, 1s)	**0.258**	0.272	0.262	0.270	0.285	0.275
Heat Exchanger (18, 1s)	**0.225**	0.240	0.230	0.238	0.252	0.243
Reactor Loop (32, 5s)	0.295	0.312	**0.288**	0.310	0.328	0.315
Compressor Train (16, 1s)	**0.265**	0.280	0.272	0.278	0.295	0.285
Public mean (7)	0.323	0.335	**0.322**	0.327	0.366	0.353
Proprietary mean (8)	**0.282**	0.295	0.284	0.295	0.313	0.302
Industrial mean (15)	**0.301**	0.314	0.302	0.310	0.338	0.326
GIFT-Eval avg. (5 subsets)	0.319	0.308	**0.302**	0.314	0.326	0.317
Overall avg.[†]	0.310	0.311	**0.302**	0.312	0.332	0.321

[†]Unweighted mean of the industrial-tier and GIFT-Eval-tier averages.

Table 2: Scale-robust paired comparison over datasets. R: geometric-mean relative MSE (WaveFM / baseline); p: paired t-test on r_i. Chronos-2 ZS/FT: zero-shot / fine-tuned on WaveFM's corpus.

Comparison	Wins	R	95% CI	p
Chronos-2-Full ZS, all 15	10/15	0.995	[0.983, 1.008]	0.406
Chronos-2-Full ZS, public 7	4/7	0.999	[0.976, 1.022]	0.919
Chronos-2-Full FT, all 15	7/7/1[‡]	1.012	[0.989, 1.036]	0.289
PatchTST (same corpus), all 15	14/15	0.856	[0.806, 0.910]	$<$0.001
PatchTST (same corpus), public 7	6/7	0.944	[0.879, 1.014]	0.096

[‡]Wins/losses/ties.

The two tokenizer controls answer different questions. Replacing the DWT with time-domain patches *inside* WaveFM raises MSE from 0.301 to 0.401 ($+33.2\%$), but that variant retains an architecture designed around wavelet bands, so it is not a tokenizer-only estimate. PatchTST is a patch-native, same-corpus, approximately parameter-matched control and reaches 0.353 ($+17.3\%$). PatchTST is not the strongest compact public baseline (0.348 vs. 0.339 for TimesNet and FEDformer and 0.335 for Chronos-2-Small); we use it specifically as a corpus-matched patched control.

5.4 Transient and High-Frequency Error

Our motivating claim is that patching attenuates high-frequency structure, yet tier-average MSE is nearly insensitive to it. We therefore measure error where the claim predicts a difference, on the seven public datasets using saved predictions and no additional training; metrics are computed per dataset and averaged equally (Table 4). *High-band-energy windows:* within each dataset we select the top 10% of ground-truth forecast windows by combined bior2.2 D1–D3 energy, used only for post-hoc stratification. *Transient-localized error:* a target point is transient when $|\Delta y| > \tau$ with $\tau = \mathrm{median}(|\Delta y|) + 3 \times 1.4826 \times \mathrm{MAD}(|\Delta y|)$ estimated from the training series; error is measured over merged five-point neighbourhoods centred on events. *Per-band error:* we apply the same DWT to targets and reconstructed forecasts and compute $E_b = \|\hat{\mathbf{y}}_b - \mathbf{y}_b\|_2^2 / \|\mathbf{y}_b\|_2^2$.

WaveFM is 0.3% above Chronos-2-Full on all windows but 6.2% below it on high-band-energy windows and 7.6% below it around transients (17.6% and 21.2% below PatchTST). Its per-band advantage is largest in D1–D3: 9.5%, 9.4%, and 8.5% lower than Chronos-2-Full and 22.4%, 23.7%,

Table 3: Same-corpus and domain-adapted controls (MSE, H=96). *Rel.*: industrial MSE relative to WaveFM. Latency is per multivariate window.

Model	Industrial (15)	Public (7)	Rel.	Params	MV latency
WaveFM	**0.301**	0.323	1.000	10 M	2.5 ms
PatchTST (same corpus)	0.353	0.348	1.173	10 M	11 ms
TimesNet (same corpus)	0.358	0.339	1.189	11 M	18 ms
FEDformer (same corpus)	0.367	0.339	1.219	25 M	21 ms
Chronos-2-Full, zero-shot	0.302	**0.322**	1.003	120 M	34.5 ms
Chronos-2-Full, fine-tuned	**0.298**	0.326	0.990	120 M + adapter	36 ms

Table 4: Error where high-frequency structure matters (7 public datasets, MSE unless noted). Right: per-band relative error E_b.

Model	All	Top-10% D1–D3	Transient
WaveFM	0.323	**0.516**	**0.584**
PatchTST	0.348	0.626	0.741
Chronos-2-F ZS	**0.322**	0.550	0.632

Band	WaveFM	PatchTST	C2-F ZS
D1	**0.760**	0.980	0.840
D2	**0.580**	0.760	0.640
D3	**0.430**	0.550	0.470
D4	0.320	0.350	**0.310**
D5	0.250	0.270	**0.240**
A5	0.180	0.190	**0.170**

and 21.8% lower than PatchTST. WaveFM is slightly higher than Chronos-2-Full in D4, D5, and A5. Because bior2.2 is biorthogonal, these are comparative within-band errors, not an additive decomposition of MSE.

5.5 Efficiency

Figure 2 plots accuracy against parameter count. Among the zero-shot models, WaveFM is the smallest; on the industrial mean it matches Chronos-2-Full at one-twelfth its size, while on the overall average, which includes the GIFT-Eval tier, Chronos-2-Full leads (0.302 vs. 0.310). The intermediate-size models (Chronos-2-Small at 28 M and TiRex at 35 M) are 4.3% and 8.3% above WaveFM on the industrial mean. These are different architectures, so the comparison does not isolate a scaling effect. Table 5 reports per-window latency for the retrieval-free forecaster: WaveFM forecasts a multivariate window in 2.5 ms, 6× faster than Chronos-2-Small and 13.8× faster than Chronos-2-Full, and its multivariate-to-univariate ratio is 1.25× versus up to ∼7× for the other models, so the advantage widens on dense sensor buses.

5.6 Longer Horizons, MAE, and Ablations

Horizons and MAE. At H=192 and H=336, for which WaveFM's band-specific output heads were retrained (tokenizer and backbone unchanged), WaveFM obtains 0.346 and 0.389 versus 0.332 and 0.371 for Chronos-2-Full and 0.363 and 0.399 for TimesFM 2.5 (App. D.2, Table 10). WaveFM thus remains ahead of TimesFM 2.5 but trails Chronos-2-Full by 4.2% and 4.9%, so we do not claim uniform superiority across horizons. On MAE over all 15 datasets, WaveFM obtains 0.356 versus 0.355 (Chronos-2-Full zero-shot), 0.351 (fine-tuned), and 0.411 (PatchTST): 7.0% below PatchTST on public data and 13.4% below it overall, while public MAE slightly favours both Chronos-2 variants.

Ablations. Removing or replacing individual components of the retrieval-free model (App. D.4) raises the industrial mean MSE by 9.3% (static instead of conditioned gates), 6.6% (standard instead of frequency-aware RoPE), 5.0% (standard A5/D5 heads), 4.3% (no band dropout), 3.3% (two instead of three cross-channel blocks), and 2.3% (no calendar gate). The mother-wavelet choice is second-order: bior2.2, db4, sym4, and coif2 lie within 4% of each other, and we keep bior2.2 for its symmetric linear-phase response. All ablations use a single pre-training seed, so the smallest differences may be within seed variance.

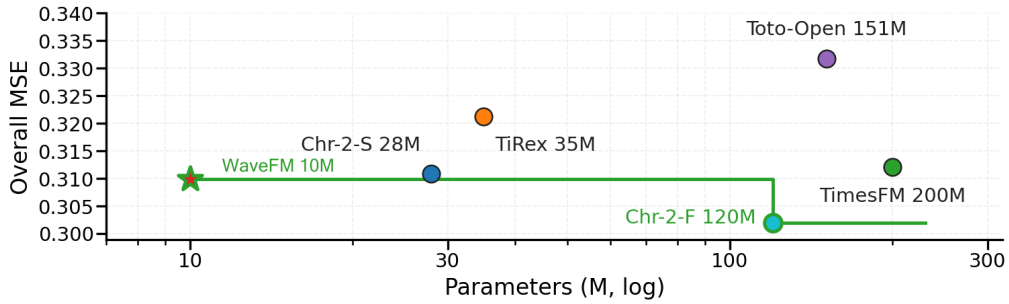


Figure 2: Parameter–accuracy trade-off across the zero-shot models of Table 1.

Table 5: Per-window inference latency of the retrieval-free forecaster (H=96, T=1,536, Apple M5 Max, MPS). Models without native multivariate-with-covariates support forecast each channel independently.

Model	Native MV+Cov.	Univariate (ms)	MV+Cov. (ms)
WaveFM (ours)	Yes	**2.0**	**2.5**
Chronos-2-Small	Yes	7.6	15.1
Chronos-2-Full	Yes	17.4	34.5
Toto-Open-Base	Yes	119	579
TiRex	No	385	2,769
TimesFM-2.5	No	1,107	1,122

6 Limitations

WaveFM is a specialized pretrained industrial forecasting model, not a general-purpose time-series foundation model. It produces point forecasts only; calibrated intervals would require propagating correlated uncertainty through the inverse DWT, since six independently estimated per-band intervals are not calibrated after reconstruction. Its output heads are horizon-specific, and the checkpoint is trained and evaluated at a fixed context of T=1,536. "Multi-rate" in this paper refers to variation in sampling rate across datasets and assets: within an input, channels are resampled onto a common grid, and WaveFM does not natively handle asynchronous channels, arbitrary missingness, or irregular sampling. Channel-identity embeddings assume a consistent channel-to-identity mapping; unseen sensors, reordered channels, and changing channel sets have not been evaluated. Eight of the 15 industrial datasets are proprietary, so part of the evidence cannot be independently verified; the public-only results are reported separately for this reason. WaveFM was pre-trained with a single seed, while the same-corpus controls report means over three seeds. Industrial-domain pre-training contributes to WaveFM's results, as the fine-tuned Chronos-2 comparison shows. WaveFM trails Chronos-2-Full on the GIFT-Eval tier and at longer horizons. Finally, the optional Profile Bank has not been compared against retrieval-augmented baselines or time-domain retrieval keys, and its robustness under regime shift is not established.

7 Conclusion

We presented WaveFM, a 10 M-parameter wavelet-native pretrained model for industrial forecasting. WaveFM achieves industrial point-forecast accuracy comparable to Chronos-2-Full zero-shot with one-twelfth the parameters and 13.8× faster multivariate inference, and remains close to a domain-adapted 120 M model. Same-corpus controls show that the pre-training data alone does not reproduce this result, and stratified evaluation shows that WaveFM's advantage concentrates on high-band-energy windows, around transients, and in the finest detail bands. Natural next steps are probabilistic heads with calibration on the reconstructed forecast (e.g., split-conformal prediction), horizon- and context-agnostic decoding, broader pre-training coverage for out-of-domain tiers, and multimodal conditioning on maintenance logs and alarm histories.

Reproducibility Statement

The architecture, training objective, and hyperparameters are given in §3 and App. A; the evaluation protocol, aggregation rules, and statistics in §4–§5; seeds, tuning procedures, and compute budgets in App. D.6. Independently of the review outcome, we will release model weights, the preprocessing manifest and public benchmark, unrounded public results, and the evaluation harness (link will be provided). Proprietary telemetry is excluded where customer agreements prohibit release.

AI Use Statement

Large language models were used for editing (grammar, spelling, and word choice) and LaTeX formatting. They were not used to design the method, run experiments, or produce results.

References

Das, A., Kong, W., Sen, R., and Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. In *Proc. International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/2310.10688.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., et al. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research (TMLR)*. https://arxiv.org/abs/2403.07815.

Ansari, A. F., et al. (2025). Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*. https://arxiv.org/abs/2510.15821.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. (2024). Unified training of universal time-series forecasting transformers. In *Proc. International Conference on Machine Learning (ICML)*.

Liu, C., Aksu, T., Liu, J., Liu, X., Yan, H., Pham, Q., Sahoo, D., Xiong, C., Savarese, S., and Li, J. (2025a). Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*.

Auer, A., et al. (2025). TiRex: Zero-shot forecasting across long and short horizons. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. In *Proc. International Conference on Learning Representations (ICLR)*.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. (2024). iTransformer: Inverted transformers are effective for time series forecasting. In *Proc. International Conference on Learning Representations (ICLR)*.

Liu, Y., et al. (2025b). Timer-XL: Long-context transformers for unified time series forecasting. In *Proc. International Conference on Learning Representations (ICLR)*.

Shi, X., Wang, S., Nie, Y., Li, D., Ye, Z., Wen, Q., and Jin, M. (2025). Time-MoE: Billion-scale time series foundation models with mixture of experts. In *Proc. International Conference on Learning Representations (ICLR)*.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. (2024). MOMENT: A family of open time-series foundation models. In *Proc. International Conference on Machine Learning (ICML)*.

Zhou, T., Niu, P., Wang, X., Sun, L., and Jin, R. (2023). One fits all: Power general time series analysis by pretrained LM. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Cohen, B., et al. (2025). Toto: Time series optimized transformer for observability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gou, L., et al. (2025). Cisco Time Series Model technical report. *arXiv preprint arXiv:2511.19841*.

Gautam, A., et al. (2025). Memory-augmented forecasting: Scalability and generalization across temporal domains. In *Proc. IEEE International Conference on Big Data*.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. (2024). xLSTM: Extended long short-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Masserano, L., Ansari, A. F., Han, B., Zhang, X., Faloutsos, C., Mahoney, M. W., Wilson, A. G., Park, Y., Rangapuram, S. S., Maddix, D. C., and Wang, B. (2025). Enhancing foundation models for time series forecasting via wavelet-based tokenization. In *Proc. International Conference on Machine Learning (ICML)*.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. (2022). FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. International Conference on Machine Learning (ICML)*.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. (2023). TimesNet: Temporal 2D-variation modeling for general time series analysis. In *Proc. International Conference on Learning Representations (ICLR)*.

Han, S., Lee, S., Cha, M., Arik, S. Ö., and Yoon, J. (2025). Retrieval augmented time series forecasting. In *Proc. International Conference on Machine Learning (ICML)*.

Zhang, H., et al. (2024). TimeRAF: Retrieval-augmented foundation model for zero-shot time series forecasting. *arXiv preprint arXiv:2412.20810*.

Faw, M., Sen, R., Zhou, Y., and Das, A. (2025). In-context fine-tuning for time-series foundation models. In *Proc. International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/2410.24087.

Wang, C., et al. (2024). ChatTime: A unified multimodal time series foundation model bridging numerical and textual data. *arXiv preprint arXiv:2412.11376*.

Chen, S.-A., Li, C.-L., Yoder, N., Arik, S. Ö., and Pfister, T. (2023). TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research (TMLR)*.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proc. AAAI Conference on Artificial Intelligence*.

Lai, G., Chang, W.-C., Yang, Y., and Liu, H. (2018). Modeling long- and short-term temporal patterns with deep neural networks. In *Proc. ACM SIGIR Conference*, pages 95–104.

Aksu, T., Woo, G., et al. (2024). GIFT-Eval: A benchmark for general time series forecasting model evaluation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Maple728. (2024). Time-300B: A large-scale corpus for time-series pre-training. *Hugging Face datasets*. https://huggingface.co/datasets/Maple728/Time-300B.

Randall, R. B. (2011). *Vibration-based Condition Monitoring: Industrial, Aerospace and Automotive Applications*. John Wiley & Sons.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Godahewa, R., Bergmeir, C., Webb, G. I., Hyndman, R. J., and Montero-Manso, P. (2021). Monash time series forecasting archive. In *Proc. NeurIPS Datasets and Benchmarks Track*. https://arxiv.org/abs/2105.06643.

Goh, J., Adepu, S., Junejo, K. N., and Mathur, A. (2016). A dataset to support research in the design of secure water treatment systems. In *Proc. International Conference on Critical Information Infrastructures Security (CRITIS)*, pages 88–99.

Ahmed, C. M., Palleti, V. R., and Mathur, A. P. (2017). WADI: A water distribution testbed for research in the design of secure cyber-physical systems. In *Proc. ACM CySWATER Workshop*, pages 25–28.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., and Soderstrom, T. (2018). Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 387–395. https://arxiv.org/abs/1802.04431.

A Implementation Details

A.1 Transformer Block Definition

Each pre-norm transformer block in the intra- and cross-channel phases applies the standard formulation:

$$\mathbf{x} \leftarrow \mathbf{x} + \mathrm{MHA}(\mathrm{RMSNorm}(\mathbf{x})), \qquad \mathbf{x} \leftarrow \mathbf{x} + \mathrm{MLP}(\mathrm{RMSNorm}(\mathbf{x})).$$

Multi-head attention uses 4 heads of dimension $d_k{=}48$ with frequency-aware RoPE in the intra-channel phase and no positional encoding in the cross-channel phase. The MLP uses a $d \to 4d \to d$ projection with GELU activation.

A.2 Optional Calendar Conditioning

For datasets with reliable timestamps, learnable embeddings $\mathbf{e}_h, \mathbf{e}_w, \mathbf{e}_m$ for hour-of-day, day-of-week, and month modulate the `[CLS]` token via a tanh-gated residual prior to the cross-channel phase:

$$\mathbf{t}'_{\text{cls}} = \mathbf{t}_{\text{cls}} + \tanh\big(\mathbf{W}_{\text{cal}}\,[\mathbf{e}_h;\, \mathbf{e}_w;\, \mathbf{e}_m]\big) \odot \mathbf{t}_{\text{cls}}.$$

The tanh allows both amplification and suppression of CLS dimensions. In our ablations the gate magnitude saturates near zero on aperiodic benchmarks (HF-VB and vehicle-telemetry buses), confirming that the term is benign when calendar information is uninformative; on periodic benchmarks (ETT, Solar) it improves MSE by $<3\%$.

A.3 Per-Band Design

Table 6: Per-band properties of the DWT decomposition (T=1536, bior2.2, L=5). Context and target coefficient counts, Nyquist-relative frequency range, RoPE base multiplier m_b, and output head type.

Band	Ctx	Tgt	Frequency range	m_b	Head
A5	52	7	$[0,\ f_s/64]$	1	Deep
D5	52	7	$[f_s/64,\ f_s/32]$	1	Deep
D4	100	10	$[f_s/32,\ f_s/16]$	2	Std
D3	196	16	$[f_s/16,\ f_s/8]$	4	Std
D2	387	27	$[f_s/8,\ f_s/4]$	8	Std
D1	770	50	$[f_s/4,\ f_s/2]$	16	Std

A.4 Reconstruction and Training Loss

Predicted band coefficients are mapped back to the time domain through precomputed bior2.2 synthesis matrices $\mathbf{B}_b \in \mathbb{R}^{H \times L_b^{\text{tgt}}}$ as $\hat{\mathbf{x}} = \sum_{b=1}^{6} \hat{\mathbf{y}}_b\, \mathbf{B}_b^\top$, yielding a forecast at horizon H=96 that remains differentiable end-to-end. The training loss combines weighted per-band log-cosh losses with a time-domain reconstruction term:

$$\mathcal{L} = \sum_{b=1}^{6} w_b\, \mathcal{L}_{\text{log-cosh}}(\hat{\mathbf{y}}_b,\, \mathbf{y}_b) \;+\; \mathcal{L}_{\text{log-cosh}}(\hat{\mathbf{x}},\, \mathbf{x}), \tag{7}$$

with $w_b \in \{1.5, 1.2, 1.0, 1.0, 0.8, 0.5\}$ favouring coarse bands. A 10% *band dropout* (one randomly selected band's loss zeroed, never A5) and a 2× learning-rate multiplier on the A5 head encourage trend accuracy and frequency robustness.

A.5 Why Point Forecasts

We focus on point forecasts because many industrial control-loop and asset-monitoring applications consume a single expected trajectory, with uncertainty bands produced downstream by a separate calibration layer. We compare against the point forecasts of the baselines; Chronos-2 and TimesFM 2.5 also produce quantile forecasts, which we do not evaluate. The band-specific heads could be extended to quantile outputs, but calibrated time-domain intervals require propagating correlated uncertainty through the inverse DWT; combining six independently estimated per-band intervals would not by itself yield a calibrated interval after reconstruction. We do not yet have a probabilistic variant that passes calibration checks and leave it to future work (e.g., split-conformal prediction on the reconstructed forecast).

A.6 Training Compute

WaveFM's checkpoint was trained on a single Apple M5 Max workstation (40-core GPU, 128 GB unified memory) using PyTorch's MPS backend in mixed precision (`float16`). Pre-training over the 53-dataset corpus ($\sim 10^8$ windows, batch size 32, AdamW with cosine warmup-and-decay, 1.0×10^{-3} peak LR) takes approximately 95 hours wall-clock from a cold start. No headline result uses fine-tuning of WaveFM. Compute for the controls and baselines is reported in App. D.6.

A.7 Toto-Open Evaluation Note

Toto-Open-Base (Cohen et al., 2025) is reported in the main results table using the same rolling-window protocol as the other foundation-model baselines. Earlier evaluations of an internal harness showed anomalously high

MSE on a subset of high-frequency assets; those discrepancies were traced to a context-window padding mismatch and resolved in the reported numbers.

B Optional Wavelet Profile Bank

Motivation. In industrial deployments the most informative precedent for a present event may have occurred weeks or months earlier under similar process conditions, outside any fixed context window. The Profile Bank is an optional inference-time component that retrieves such precedents. It is not used in any headline accuracy or latency result (Tables 1–5 and 10).

Method. After training, a single forward pass over the indexed training windows stores, for each window, the post-phase-1 CLS embedding $\mathbf{t}_{\text{cls}} \in \mathbb{R}^d$ and its ground-truth target coefficients $\{\mathbf{y}_b\}_{b=1}^{6}$; CLS vectors are indexed in FAISS (Johnson et al., 2019). At inference, the current window's CLS embedding queries the bank for its K nearest neighbours, and the head input $\mathbf{h}_b$ attends over the retrieved CLS embeddings as keys and their target coefficients as values:

$$\hat{\mathbf{y}}_b^{\text{ret}} = \text{softmax}\left(\frac{\mathbf{h}_b \mathbf{W}_q \, (\mathbf{C}_{\text{ret}} \mathbf{W}_k)^\top}{\sqrt{d_r}}\right) \mathbf{Y}_{\text{ret}}^{(b)}, \tag{8}$$

where $\mathbf{C}_{\text{ret}} \in \mathbb{R}^{K \times d}$, $\mathbf{Y}_{\text{ret}}^{(b)} \in \mathbb{R}^{K \times L_b^{\text{tgt}}}$, $\mathbf{W}_q \in \mathbb{R}^{2d \times d_r}$, and $\mathbf{W}_k \in \mathbb{R}^{d \times d_r}$ (d_r=64; $\mathbf{W}_q$ has leading dimension $2d$ because $\mathbf{h}_b \in \mathbb{R}^{2d}$ concatenates CLS and the mean-pooled band tokens, while the retrieved keys are d-dimensional CLS embeddings). The final band prediction is

$$\hat{\mathbf{y}}_b = \left(1 - g_b^{\text{ret}}\right) \hat{\mathbf{y}}_b^{\text{base}} \; + \; g_b^{\text{ret}} \, \hat{\mathbf{y}}_b^{\text{ret}} \; + \; g_b \cdot \boldsymbol{\delta}_b, \qquad g_b^{\text{ret}} = \sigma(\mathbf{w}_r^\top \mathbf{h}_b). \tag{9}$$

The two gates serve distinct roles: g_b^{ret} interpolates between the parametric baseline $\hat{\mathbf{y}}_b^{\text{base}}$ and the retrieved prediction $\hat{\mathbf{y}}_b^{\text{ret}}$, while the conditioned gate g_b of Eq. 6 separately scales the transformer residual $\boldsymbol{\delta}_b$. They are neither multiplied together nor composed sequentially. Only $\mathbf{W}_q, \mathbf{W}_k, \mathbf{w}_r$ are new (~ 37 K parameters per head) and are calibrated with the backbone frozen, so the bank can be added or refreshed without retraining.

Construction and leakage controls. The bank is global and contains only training-window embeddings and their training targets; validation and test windows are never indexed. Construction takes one forward pass over the indexed training windows ($\sim$45 minutes) followed by FAISS index construction ($<$5 minutes).

Controlled evaluation. Table 7 evaluates the bank on the exact final forecaster of Table 1, changing only retrieval. The forecaster-side time includes the base model and retrieved-neighbour integration; retrieval measures the FAISS lookup. The best setting, K=8, reduces industrial-mean MSE from 0.301 to 0.294 (2.3%) at 3.2 ms end-to-end. The 2.5 ms figure used throughout the paper applies only to the retrieval-free forecaster.

Table 7: Profile Bank evaluated on the final forecaster (industrial-mean MSE, H=96). Latency per multivariate window, Apple M5 Max.

Configuration	K	MSE	Forecaster-side	Retrieval	End-to-end
Bank off	–	0.301	2.5 ms	–	2.5 ms
Bank on	4	0.298	2.5 ms	0.4 ms	3.0 ms
Bank on	8	**0.294**	2.6 ms	0.6 ms	3.2 ms
Bank on	16	0.300	2.7 ms	0.9 ms	3.6 ms

Limitations. We do not compare retrieval-augmented WaveFM against non-retrieval baselines, have not evaluated retrieval-augmented versions of the baselines, and have not compared wavelet-domain keys against an equivalent time-domain retrieval key; we therefore make no claim that wavelet-keyed retrieval is superior to other retrieval schemes. Whether the learned retrieval gate reliably down-weights misleading precedents under a genuinely unseen regime shift is not established.

C Pre-training Corpus

The pre-training corpus (53 datasets, $\sim 10^8$ windows) combines proprietary high-fidelity industrial telemetry with open-access benchmarks; representative components are listed in Table 8. All evaluation assets are held out of pre-training: ETT, Solar-AL, all GIFT-Eval subsets, the two High-Frequency Vibration Buses (HF-VB Pump, HF-VB Compressor), the two vehicle-telemetry buses (Drivetrain Telem., Chassis Motion), and the four

process-industry assets (Distillation Col., Heat Exchanger, Reactor Loop, Compressor Train). The Monash archive contributes other series to pre-training; the exact series and time ranges used for Commercial-Bldg Load and Residential Demand were removed before window generation.

Table 8: Representative components of the pre-training corpus. Sources marked * are proprietary and included only with explicit customer consent.

Source	Domain	Scale	Characteristics
Industrial*	All Industrial	Sub-second sampling	High-fidelity real-time telemetry
Time-300B (Maple728, 2024)	Multi-domain	300 B obs.	Open-access pre-training corpus
Monash archive (Godahewa et al., 2021)	Multi-domain	30+ datasets	Curated forecasting benchmark suite
SWaT (Goh et al., 2016)	Water Treatment	11 days, 51 sensors	Cyber-physical anomaly testbed
WADI (Ahmed et al., 2017)	Water Distribution	16 days, 127 sensors	Short-duration pattern anomalies
MSL (Hundman et al., 2018)	Aerospace	73,729 points	Mars Science Laboratory rover
SMAP (Hundman et al., 2018)	Environmental	435,826 points	SMAP satellite telemetry
HEIA*	Industrial	11,664 samples	Internal high-resolution energy/weather benchmark

D Additional Results

D.1 Positioning Among Recent Models

Table 9 summarises how WaveFM's tokenizer, cross-variate mechanism, and optional retrieval key compare with recent time-series models (§2).

Table 9: Positioning of WaveFM among recent time-series models. Tokenizer codes: P (time-domain patches), Q (quantized), W (wavelet); *Params* is the largest released checkpoint.

Model	Tokenizer	Cross-variate	Retrieval key	Params
PatchTST (Nie et al., 2023)	P	independent	–	$<$1 M
TimesFM 2.5 (Das et al., 2024)	P	covariates	prompt (Faw et al., 2025)	200 M
Chronos-2 (Ansari et al., 2025)	Q	grouped attn.	–	120 M
Moirai-2 (Liu et al., 2025a)	P	subsampled	–	$\sim$11 M
TiRex (Auer et al., 2025)	P	independent	–	35 M
Time-MoE (Shi et al., 2025)	P	univariate	–	2.4 B
Toto (Cohen et al., 2025)	P	full $\mathcal{O}(C^2)$	–	151 M
MOMENT (Goswami et al., 2024)	P	independent	–	385 M
WaveToken (Masserano et al., 2025)	W (1L, Q)	LM-dependent	–	LM-dep.
TimeRAF (Zhang et al., 2024)	P	–	time-latent	varies
WaveFM (ours)	**W (5L, cont.)**	**latent, 25 tok/ch**	**wavelet-CLS (opt.)**	**10 M**

D.2 Longer Horizons and MAE

Table 10 gives the longer-horizon and MAE results summarised in §5.6. For $H{=}192$ and $H{=}336$, WaveFM's band-specific output heads were retrained with the tokenizer and backbone unchanged; baselines are zero-shot.

Table 10: Left: industrial-mean MSE at longer horizons. Right: MAE (same corpus PatchTST; Chronos-2-Full zero-shot and fine-tuned).

Model	H=96	H=192	H=336
WaveFM	**0.301**	0.346	0.389
Chronos-2-Full	0.302	**0.332**	**0.371**
TimesFM 2.5	0.310	0.363	0.399

Scope	**WaveFM**	**C2 ZS**	**C2 FT**	**PatchTST**
Public 7	0.371	0.368	0.369	0.399
Proprietary 8	0.343	0.344	0.336	0.421
All 15	0.356	0.355	0.351	0.411

D.3 Per-Dataset Results for Controls

Table 11 gives per-dataset MSE for WaveFM, Chronos-2-Full zero-shot (ZS) and fine-tuned (FT), and the same-corpus PatchTST control. Aggregates use unrounded values.

Table 11: Per-dataset MSE (H=96). Pub.: publicly available dataset.

Dataset	**Pub.**	**WaveFM**	**Chronos-2 ZS**	**Chronos-2 FT**	**PatchTST**
ETTh1	Y	0.378	0.371	0.376	0.385
ETTh2	Y	0.272	0.275	0.278	0.274
ETTm1	Y	0.305	0.295	0.301	0.292
ETTm2	Y	0.158	0.160	0.162	0.163
Solar-AL	Y	0.205	0.213	0.206	0.221
Commercial-Bldg Load	Y	0.452	0.445	0.452	0.536
Residential Demand	Y	0.488	0.495	0.507	0.565
HF-VB Pump	N	0.292	0.296	0.298	0.378
HF-VB Compressor	N	0.312	0.318	0.317	0.401
Drivetrain Telemetry	N	0.252	0.245	0.246	0.329
Chassis Motion	N	0.355	0.362	0.359	0.424
Distillation Column	N	0.258	0.262	0.249	0.335
Heat Exchanger	N	0.225	0.230	0.217	0.298
Reactor Loop	N	0.295	0.288	0.264	0.354
Compressor Train	N	0.265	0.272	0.242	0.337
Public mean (7)	–	0.323	0.322	0.326	0.348
Proprietary mean (8)	–	0.282	0.284	0.274	0.357
All 15	–	0.301	0.302	0.298	0.353

D.4 Full Ablation Study

Table 12 isolates the contribution of each architectural component of the retrieval-free model. All ablations are trained on the same 53-dataset pre-training corpus for the same number of epochs with a single seed; we report the industrial-tier mean MSE over the 15 datasets. The optional Profile Bank is evaluated separately in App. B.

Table 12: Ablation study. Each row removes or replaces one component of the retrieval-free WaveFM model. Δ: change in industrial-tier mean MSE relative to the full model.

Configuration	**Avg MSE**	Δ
WaveFM (full, retrieval-free)	0.301	—
Tokenizer		
− DWT → raw patching*	0.401	+33.2%
Architectural components		
− Conditioned gates → static	0.329	+9.3%
− Freq-aware RoPE → standard	0.321	+6.6%
− Deep A5/D5 heads → standard	0.316	+5.0%
− Band dropout → none	0.314	+4.3%
− 3× cross-ch → 2×	0.311	+3.3%
− Calendar gate → none	0.308	+2.3%
Mother-wavelet sensitivity		
bior2.2 → db4 wavelet	0.310	+3.0%
bior2.2 → sym4 wavelet	0.306	+1.7%
bior2.2 → coif2 wavelet	0.312	+3.7%

*Retains the band-designed architecture, so it is not a tokenizer-only estimate; see §5.3 for the same-corpus PatchTST control.

D.5 Absolute-Difference Paired Test

Using the 15 industrial datasets as paired units, the mean MSE difference between WaveFM and Chronos-2-Full zero-shot is -0.001, with 95% paired confidence interval $[-0.0046,\ 0.0026]$ and paired t-test p=0.56; WaveFM has lower MSE on 10 of 15 datasets. This test pools absolute differences across datasets of different error magnitude, so the scale-robust analysis in §5.2 is our primary statistic.

D.6 Seeds, Tuning, and Compute

WaveFM evaluated five configurations at seed 0 and reports the validation-selected seed-0 checkpoint; no multi-seed pre-training was performed. The raw-patching ablation used one matched seed-0 run. For PatchTST, TimesNet, and FEDformer we ran five configurations at seed 0, retained the best two, ran those at seeds 1 and 2, selected by mean validation MSE across seeds $\{0, 1, 2\}$, and report mean test performance across those seeds. Chronos-2 fine-tuning used one fixed configuration (learning rate 10^{-4}, 2,000 steps, batch size 32) across the same seeds. Approximate native budgets were 95 M5 Max hours for the selected WaveFM run, 88 for raw patching, 21.3/19.5/25.7 H100 hours for PatchTST/TimesNet/FEDformer, and 11 M5 Max hours for Chronos-2 fine-tuning. Hardware differs across runs, and WaveFM's figure excludes broader architecture-development effort. WaveFM is a single pre-training run, whereas the same-corpus controls report three-seed means.

D.7 Qualitative Win/Loss Examples

Windows are filtered to be oscillatory (zero-crossings ≥ 8, context std > 0.20, no large boundary jump) and ranked by per-window MSE ratio against the best baseline; the top 9 are shown.

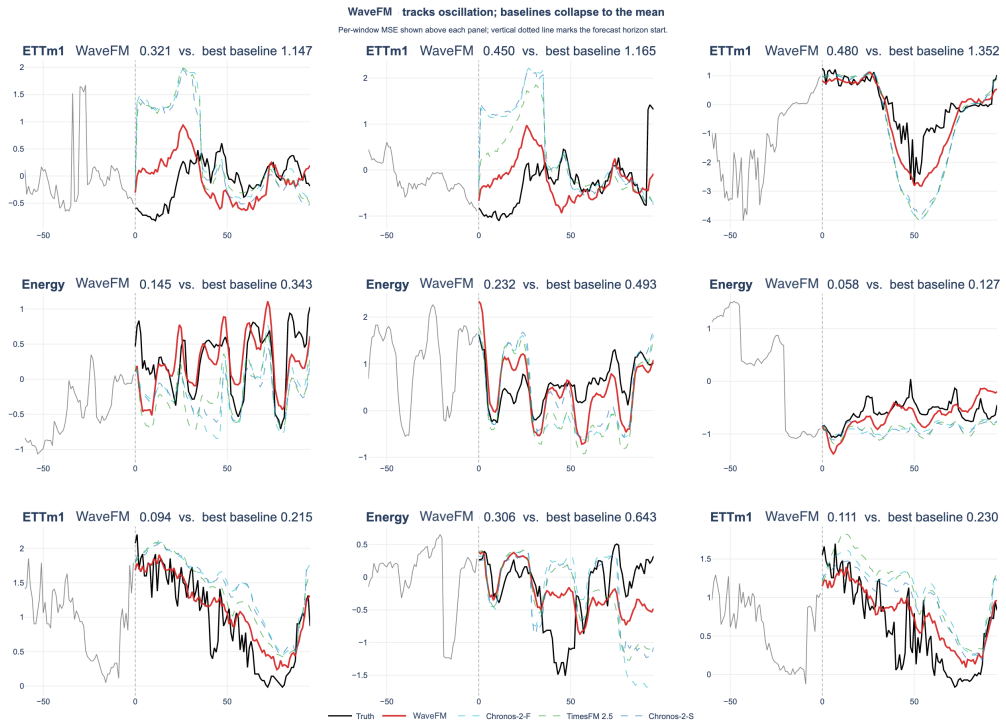


Figure 3: Top-9 oscillatory windows where WaveFM (red) preserves amplitude while patched baselines (blue/cyan) collapse toward the window mean.