RECAP: Reconstructing the Past to Predict the Future in Time-Series Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm for enhancing Time Series Foundation Models (TSFMs) with domain-specific historical knowledge, mitigating their limitations under distribution shifts and non-stationary dynamics. However, existing retrieval-based forecasting approaches typically rely on shallow similarity measures, such as fixed distance metrics or naive correlations, which often retrieve superficially similar patterns that are not truly predictive of future behavior. In this work, we propose a learnable coarse-to-fine retrieval framework centered on a simple principle: *reconstructing the past to predict the future*. Rather than retrieving examples based solely on surface similarity, our method learns to identify historical contexts whose latent temporal structure can best reconstruct the query sequence and whose future trajectories provide the most informative predictive signals. Specifically, we first perform efficient candidate selection using approximate nearest neighbor search, followed by a learned reranking stage operating in a shared feature space. The reranker evaluates retrieved segments according to how well a weighted combination of their representations can reconstruct the original query context, coupling retrieval directly to downstream forecasting quality. In this way, retrieval becomes not merely a similarity search problem, but a reconstruction-driven mechanism for discovering predictive temporal analogies. Importantly, the entire retrieval pipeline operates on top of a frozen TSFM backbone, incurring only some extra lightweight modules. We further introduce a wavelet-based multi-scale regularization loss on the median quantile prediction to improve temporal consistency across different frequency bands. Extensive experiments on public benchmark datasets demonstrate that our method consistently outperforms existing forecasting baselines and retrieval-based approaches.

1 Introduction

Large-scale pretrained time series foundation models (TSFMs) have fundamentally reshaped the forecasting paradigm across domains and temporal frequencies without task-specific retraining (Ansari et al., 2024; Woo et al., 2024). Yet a fundamental limitation persists: the predictive knowledge of these models is *statically encoded* in frozen weights. At inference time, a TSFM applies a fixed prior over the temporal dynamics learned from pretraining corpora, with no mechanism to adapt to the distributional characteristics of the query (input) at hand. This is particularly significant when the query exhibits non-stationarity, domain-specific periodicity, or distribution shift from pretraining. The historical context available at inference time often contains highly localized predictive signals that a frozen backbone alone cannot fully exploit. While the model may encode temporal dynamics *in aggregate* through large-scale pretraining, it lacks the ability to selectively retrieve what it *should* know for a particular query instance. As a result, the model may recognize general temporal regularities, yet fail to identify the most relevant historical analogies needed for accurate forecasting under evolving conditions. This retrieval gap becomes increasingly pronounced as the query distribution drifts farther from the pretraining corpus.

This leads to a critical weakness that pervades existing approaches: *retrieval is decoupled from the forecasting objective*. Existing methods select neighbors via static, heuristic similarity metrics—Euclidean distance in

raw or embedding space, cosine similarity over frozen representations, or DTW (Dynamic Time Warping)-based temporal alignment—that are not optimized for the downstream predictive task. Frozen embeddings, by construction, encode general-purpose representations rather than task-aligned forecasting similarity. As a consequence, retrieved neighbors may share surface-level temporal resemblance with the query but provide no meaningful correction to the base model's prediction. Moreover, retrieval weights are often produced by fixed scores that are disconnected from whether the retrieved futures improve forecasting. Existing methods (Ning et al., 2025) typically address this with a separate gating or fusion module that regulates their influence on the final prediction.

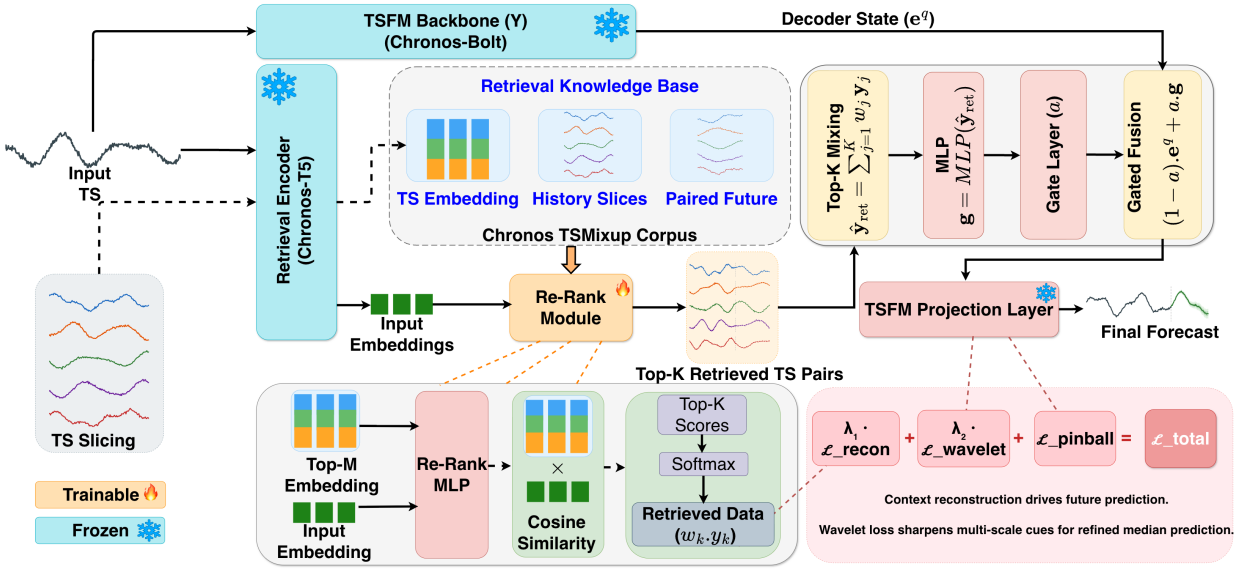


Figure 1: Overview of the our model framework. A query is encoded by the frozen Chronos-T5 encoder to retrieve $Top-K$ neighbors from a 2.1 Million-pair knowledge base via approximate nearest neighbour search and residual MLP reranking. Retrieved contexts are fused with the Chronos-Bolt encoder state through a learned gating network and decoded into nine quantile forecasts. A Haar wavelet auxiliary loss is applied to the median quantile during training. Only the Re-Ranker, projection MLP, and gating network are trainable.

While pragmatically effective, this decoupling leaves the retrieval selection step fixed: candidate ranking is determined by embedding proximity alone, independent of the downstream forecast. We make ranking learnable by training the reranker to score each candidate's predictive coherence with the query, jointly determining retrieval relevance and fusion influence.

In this work, we address these limitations through a differentiable Retrieval-Augmented Forecasting framework that rethinks how retrieved historical context should refine the predictions of a frozen TSFM. Our central insight is simple yet effective: *reconstructing the past to predict the future*. Specifically, if a weighted combination of retrieved historical contexts can faithfully reconstruct the query sequence within a shared semantic space, then the same combination weights applied to their corresponding future trajectories can provide a reliable predictive signal for forecasting. By coupling retrieval directly with reconstruction quality rather than surface-level similarity, the retrieval module learns to identify temporally informative historical analogies instead of merely correlated patterns. This design enables the entire retrieval mechanism to be trained end-to-end together with the fusion network while keeping the TSFM backbone frozen. In addition, we introduce a wavelet-based multi-scale regularization loss on the median quantile prediction to improve temporal consistency across different frequency bands for regularization.

Contributions. Our contributions are summarized as follows:

- We propose a retrieval-augmented forecasting framework, termed **RECAP**, that makes retrieval itself learnable for frozen TSFMs through the principle of *reconstructing the past to predict the future*. Instead of relying on static embedding distances, RECAP learns retrieval relevance via a

reconstruction-driven reranking objective: retrieved contexts are weighted to reconstruct the query sequence in a shared semantic space, encouraging the model to retrieve temporally informative historical analogies rather than superficially similar patterns.

- We introduce a multi-level Haar wavelet loss as a multi-scale regularizer on the median quantile prediction. By decomposing forecasts into approximation and detail coefficients across frequency bands, the loss enforces consistency in both long-term trends and fine-grained temporal structures, complementing the pinball objective with structured frequency-domain supervision.
- Extensive experiments, ablations, and sensitivity analyses on established forecasting benchmarks demonstrate that RECAP consistently outperforms both the frozen Chronos-Bolt backbone and conventional retrieval baselines. Results further show that the wavelet regularizer stabilizes retrieval-augmented training and improves robustness across datasets with heterogeneous temporal dynamics.

2 Related Work

2.1 Time Series Foundation Models

TSFMs have been pretrained on large corpora spanning heterogeneous domains, enabling zero-shot forecasting without task-specific retraining. Representative models include Chronos (Ansari et al., 2024), which tokenizes time series via scaling and quantization and trains a T5-style encoder–decoder (Raffel et al., 2020) with a cross-entropy objective; Chronos-Bolt (Ansari et al., 2024), which replaces discrete tokenization with patch-based embeddings and direct multi-step quantile regression, achieving improved zero-shot accuracy; TimesFM (Das et al., 2024), a decoder-only patch-based model; and Moirai (Woo et al., 2024), which models a mixture of distributions for probabilistic forecasting. Despite their strong zero-shot forecasting performance, these models share a fundamental limitation: their predictive knowledge is fixed after pretraining and remains encoded in frozen model weights. During inference, they rely on static priors over temporal dynamics, lacking mechanisms to incorporate external evidence or adapt to the distributional characteristics of the input sequence. As a result, they often struggle with non-stationary regime shifts and rare or evolving temporal patterns (Kim et al., 2022; Liu et al., 2022). This limitation has motivated the development of retrieval-augmented time series forecasting methods that enhance frozen TSFMs with test-time access to historical evidence.

2.2 Retrieval-Augmented Time Series Forecasting

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Guu et al., 2020) augments parametric models with a non-parametric memory of external documents, allowing predictions to be grounded in retrieved evidence without retraining (Li et al., 2022). Extending this paradigm to time series forecasting presents two fundamental challenges. First, although TSFMs produce sequence embeddings analogous to sentence embeddings in NLP, these representations are optimized for capturing temporal patterns rather than retrieval relevance. Consequently, a high similarity score indicates similar historical dynamics but does not necessarily imply that the retrieved window's future trajectory will be predictive for the query horizon. Second, the retrieved *future* of an analogous historical window constitutes the signal of interest. This fundamentally differs from RAG, where the retrieved document itself serves as supporting evidence (Tire et al., 2026).

A first line of work integrates retrieval into bespoke forecasters that are then trained or fine-tuned end-to-end on a target dataset. ReTime (Jing et al., 2022) performs relational retrieval and content synthesis for retrieval-based forecasting. RATD (Liu et al., 2024) retrieves historical references to guide the denoising process of conditional diffusion models for forecasting. RAFT (Han et al., 2025) integrates retrieval into a multi-resolution forecasting framework, retrieving relevant patches from the entire training series rather than only within the lookback window. RATSF (Wang & Cui, 2024) introduces a cross-attention module (RACA) that fuses retrieved historical segments into a Transformer forecaster and is tailored to non-stationary customer-service workloads. TimeRAG (Yang et al., 2025) constructs a knowledge base by clustering training segments and retrieves analogues with DTW, passing them as natural-language prompts to an LLM forecaster with a trainable reprogramming layer. Although these methods consistently improve

over their non-retrieval baselines, they all require gradient updates on the target distribution, which forfeits the zero-shot flexibility that makes TSFMs attractive in the first place.

A second line targets zero-shot deployment of TSFMs. RAF (Tire et al., 2026) concatenates retrieved context-future pairs directly into the Chronos (Ansari et al., 2024) input, providing inductive bias without training, but inheriting the input-length and scalability constraints of in-context conditioning. TimeRAF (Zhang et al., 2025) introduces an end-to-end learnable retriever and a Channel Prompting mechanism for integrating retrieved sequences into a TSFM, with knowledge bases drawn from multi-domain corpora. TS-RAG (Ning et al., 2025) couples a frozen TSFM with a Mixture-of-Experts adaptive retrieval mixer that fuses retrieved patterns with the model's internal representation while keeping the backbone unchanged. Most of these systems retrieve directly in the raw signal space (or in embedding spaces from off-the-shelf encoders) and combine the retrieved analogues with the query through concatenation, simple averaging, or cross-attention. As a result, they are sensitive to distance-metric choice, perform unconstrained mixing of relevant and noisy retrievals, and provide few interpretability handles.

2.3 Wavelet-Based Multi-Scale Representations for Time Series

Wavelet transforms provide simultaneous time-frequency localization and have a long history in signal processing for non-stationary data (Mallat, 1989; Daubechies, 1992). Compared with the Fourier transform, wavelets retain temporal locality, making them well-suited to time series with regime shifts, abrupt transitions, or heterogeneous spectral content (Percival & Walden, 2000). The Haar wavelet in particular admits an equivalence to dyadic averaging and pairwise differencing (Bosl et al., 2022), providing a computationally efficient orthonormal basis that cleanly separates coarse trend from fine-grained detail at each level.

Several recent forecasting models embed wavelet decomposition into their architectures. W-FENet (Wang et al., 2024a) combines trend decomposition and wavelet transform with a Fourier-enhanced feature extraction module, requiring full end-to-end retraining on each target dataset. WEITS (Guo et al., 2026) introduces a frequency-aware residual framework using the multilevel wavelet decomposition with forward-backward residual links inspired by N-BEATS, again trained from scratch on the target domain. WPMixer (Murad et al., 2025) couples multi-resolution wavelet decomposition with patching and MLP mixing in a supervised forecasting architecture. SWIFT (Xie & Cao, 2025) uses wavelet-based lossless downsampling to map sub-series through a shared lightweight predictor, targeting resource-limited edge deployment.

2.4 Probabilistic and Quantile-Based Forecasting Objectives

Probabilistic forecasting estimates a predictive distribution rather than a point forecast, which is essential for downstream risk-sensitive decision-making. The pinball (quantile) loss originates from quantile regression and remains the canonical proper scoring rule for individual quantiles, with the multi-quantile average forming a discretized approximation to the Continuous Ranked Probability Score (CRPS) (Gneiting & Raftery, 2007). DeepAR (Salinas et al., 2020) popularized parametric likelihood forecasting in deep models, while Lag-Llama (Rasul et al., 2024) and Moirai (Woo et al., 2024) extend this to TSFMs with distribution-mixture heads. In parallel, several models train directly on multi-quantile pinball losses to obtain non-parametric predictive intervals (Wen et al., 2017). Chronos-Bolt (Ansari et al., 2024) adopts the latter approach, predicting a fixed grid of nine quantiles via direct multi-step decoding. Our training objective inherits this pinball-loss formulation and augments it with two task-aligned auxiliary terms—a reconstruction loss on the lookback window and a Haar wavelet loss on the median quantile—weighted by λ_{recon} and λ_{wavelet}. Because the wavelet loss is applied only to the median forecast, it acts as an auxiliary shape regularizer without disrupting the model's probabilistic quantile outputs.

3 Methodology

We propose a differentiable Retrieval-Augmented Forecasting framework that refines the output of a frozen time-series foundation model using learned retrieval from a historical context bank. Conventional retrieval-augmented approaches select relevant historical windows via fixed similarity metrics such as cosine similarity. While computationally efficient, such metrics operate on static representations and are decoupled from

the downstream forecasting objective, limiting their ability to identify context windows whose futures are genuinely informative for the query at hand.

Our framework replaces the static retrieval step with a differentiable pipeline that learns a flexible alignment between retrieved contexts and the query, without retraining the backbone or narrowing its general-purpose behavior. The key insight is that a frozen TSFM already encodes strong temporal representations; what is missing is a mechanism to use those representations to surface *which* retrieved windows are locally predictive for the query, and to adaptively blend the retrieved signal with the backbone's own forecast. By learning this alignment on top of frozen representations rather than fine-tuning the backbone, the framework preserves the zero-shot generality of TSFM while gaining the ability to adapt locally at inference time through retrieval.

The pipeline proceeds in three stages. First, a coarse search over a precomputed embedding index retrieves a broad pool of candidate historical windows whose context resembles the query. Second, a lightweight residual MLP reranker re-scores these candidates using cosine similarity on residual-refined embeddings, trained so that a weighted combination of the top-scoring contexts reconstructs the query, coupling retrieval selection to structural query coherence. Third, a gated fusion module adaptively blends the backbone's base forecast with the aggregated retrieved futures, learning on a per-query basis how much to rely on retrieval. The fused median quantile forecast is further regularized by a multi-level Haar wavelet penalty, which encourages consistency across both coarse trends and fine-grained fluctuations without altering the backbone's probabilistic output. The backbone encoder and decoder remain fully frozen throughout.

3.1 Problem Definition

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ be a historical memory bank, where $\mathbf{x}_i \in \mathbb{R}^{L \times C}$ denotes a context window and $\mathbf{y}_i \in \mathbb{R}^{H \times C}$ its corresponding future horizon. Given a query context $\mathbf{x}_q$, we seek to improve the prediction of a frozen foundation model $f_\theta(\mathbf{x}_q)$ via a retrieval-induced residual. Instead of retrieving neighbors according to a fixed similarity measure, we formulate retrieval as a reconstruction task in the instance-normalized input space. Candidate contexts are assigned reconstruction weights that best represent $\mathbf{x}_q$, and these weights are subsequently propagated to their future horizons to generate a residual forecast correction. This formulation tightly couples retrieval with forecasting utility, encouraging the selection of historical patterns whose future dynamics are most informative for the target prediction.

3.2 Reranking via Learned Similarity

The frozen latent space provides an efficient mechanism for coarse retrieval, but its embeddings are learned in a task-agnostic manner and are not explicitly optimized for forecasting relevance. To improve retrieval quality while preserving the pretrained representation, we introduce a lightweight learnable reranking module.

For each query variate $\mathbf{x}_q^{(c)}$, the coarse retrieval stage produces a candidate set $\mathcal{N}_M^{(c)} = \{j_1, \ldots, j_M\}$ containing the indices of the top-M nearest neighbors in the frozen latent space. Let $\mathbf{e}_q^{(c)}, \mathbf{e}_j^{(c)} \in \mathbb{R}^D$ denote the corresponding frozen embeddings of the query and candidate variates, respectively, obtained from the backbone encoder $\mathcal{F}_\phi$.

To adapt retrieval to the forecasting task while preserving the pretrained representation, we refine these embeddings using a residual MLP g_ψ:

$$\tilde{\mathbf{e}}^{(c)} = \mathbf{e}^{(c)} + g_\psi\left(\mathbf{e}^{(c)}\right), \tag{1}$$

where the residual connection allows g_ψ to learn corrections relative to the original frozen embedding $\mathbf{e}^{(c)}$, enabling task-specific refinement.

The relevance of each candidate is then assessed by the cosine similarity between the refined query and candidate embeddings: $s_j^{(c)} = \cos\left(\tilde{\mathbf{e}}_q^{(c)}, \tilde{\mathbf{e}}_j^{(c)}\right), j \in \mathcal{N}_M^{(c)}$. Based on these scores, we retain the top-K candidates

$\mathcal{K}^{(c)} \subset \mathcal{N}_M^{(c)}$ and convert their similarities into normalized retrieval weights:

$$w_j^{(c)} = \frac{\exp\left(s_j^{(c)}\right)}{\sum_{\ell \in \mathcal{K}^{(c)}} \exp\left(s_\ell^{(c)}\right)}, \qquad j \in \mathcal{K}^{(c)}. \tag{2}$$

The resulting weights $w_j^{(c)}$ form a probability distribution over the top-K candidates, conditioned on the query through the underlying cosine similarities.

3.3 Reconstruction-Regularized Gated Fusion

Given the retrieval weights $\{w_j^{(c)}\}_{j \in \mathcal{K}^{(c)}}$ and the corresponding historical pairs $\{(\mathbf{x}_j, \mathbf{y}_j)\}_{j \in \mathcal{K}^{(c)}}$, we form the retrieval-augmented prediction as the weighted combination $\hat{\mathbf{y}}_{\text{ret}} = \sum_{j \in \mathcal{K}^{(c)}} w_j \, \mathbf{y}_j \in \mathbb{R}^H$. Before fusion, $\hat{\mathbf{y}}_{\text{ret}}$ must be mapped into the same representation space as the frozen backbone's decoder output. We compute a projected representation $\mathbf{g} = f_\rho(\hat{\mathbf{y}}_{\text{ret}}) \in \mathbb{R}^D$ via a learnable MLP f_ρ, where D is the backbone's hidden dimension.

Let $\mathbf{e}^q \in \mathbb{R}^D$ denote the decoder output of the frozen backbone. A gating network Gate : $\mathbb{R}^{2D} \to \mathbb{R}^D$, implemented as a single linear layer, takes the concatenation of the backbone and retrieval representations as input and produces element-wise interpolation coefficients via:

$$\boldsymbol{\alpha} = \sigma(\text{Gate}([\mathbf{e}^q; \mathbf{g}])) \in (0,1)^D, \tag{3}$$

$$\mathbf{h} = (1 - \boldsymbol{\alpha}) \odot \mathbf{e}^q + \boldsymbol{\alpha} \odot \mathbf{g}, \tag{4}$$

where σ is the sigmoid function and $\odot$ denotes the element-wise (Hadamard) product. As training proceeds, the gate learns to modulate retrieval influence per latent dimension based on predictive utility. The fused representation $\mathbf{h}$ is subsequently passed through the frozen output projection head to produce the final quantile forecasts

$$\hat{\mathbf{y}} = f_{\text{proj}}(\mathbf{h}) \in \mathbb{R}^{H \times Q},$$

where Q denotes the number of predicted quantiles.

Retrieval reconstruction regularizer. To ensure the learned retrieval weights $\mathbf{w}$ capture structurally meaningful temporal patterns rather than arbitrary soft assignments, we regularize them with a context reconstruction objective. Let $\tilde{\mathbf{x}}_q$ and $\{\tilde{\mathbf{x}}_j\}_{j \in \mathcal{K}}$ denote the instance-normalized query and candidate context windows. We penalize the weighted reconstruction error:

$$\mathcal{L}_{\text{recon}} = \frac{1}{L} \left\| \tilde{\mathbf{x}}_q - \sum_{j=1}^{K} w_j \, \tilde{\mathbf{x}}_j \right\|_2^2. \tag{5}$$

This term provides a direct supervisory signal to the learnable reranking module (Section 3.2): if the top-K weights faithfully represent the query's temporal structure, then the same weights applied to the corresponding futures may yield a meaningful correction. The regularization strength λ_{recon} controls the trade-off between forecast accuracy and retrieval coherence.

3.4 Haar Wavelet Loss

Time series exhibit structure at multiple temporal scales: a slowly varying trend coexists with faster, fine-grained fluctuations. Within the pinball objective $\mathcal{L}_{\text{quantile}}$, the median ($\tau = 0.5$) component reduces to a point-wise L_1 penalty $|\hat{\tilde{y}}_{0.5,t} - \tilde{y}_t|$ at each timestep t, since the asymmetric quantile check that distinguishes general τ collapses to a constant at the median. This per-timestep penalty treats every position identically regardless of which temporal scale it belongs to, so two forecasts with the same aggregate error can differ substantially in how well they preserve the coarse trajectory versus the local fluctuations of the ground truth, a distinction this term cannot make on its own. We address this with an auxiliary regularizer that explicitly separates forecast error by temporal scale before penalizing it, applied exclusively to the median quantile prediction $\hat{\tilde{\mathbf{y}}}_{0.5} \in \mathbb{R}^H$.

Haar decomposition. A natural way to separate a signal into trend and scale-specific fluctuation is to repeatedly average and difference adjacent pairs of points: averaging two adjacent values produces a smoothed signal at half the resolution, while their difference isolates exactly the local variation removed by that averaging. Applying this operation recursively yields a hierarchy of increasingly coarse trend estimates, each paired with the detail discarded at that step — this is the Haar discrete wavelet transform (Mallat, 1989) (DWT). Formally, given a signal $\mathbf{s}^{(0)} = \mathbf{s} \in \mathbb{R}^T$, the approximation and detail coefficients at level $l = 1, \ldots, L$ are defined recursively from the previous level's approximation $\mathbf{s}^{(l-1)}$:

$$\mathcal{A}_k^{(l)} = \tfrac{1}{\sqrt{2}}\left(s_{2k}^{(l-1)} + s_{2k-1}^{(l-1)}\right), \quad \mathcal{D}_k^{(l)} = \tfrac{1}{\sqrt{2}}\left(s_{2k}^{(l-1)} - s_{2k-1}^{(l-1)}\right), \quad k = 1, \ldots, \lfloor T/2^l \rfloor, \tag{6}$$

with $\mathbf{s}^{(l)} = \mathcal{A}^{(l)}$. Here $\mathcal{A}^{(l)}$ is the smoothed trend at resolution 2^{-l}, obtained by passing the previous level's trend through one further round of averaging, and $\mathcal{D}^{(l)}$ is the high-frequency residual removed at that step. After L levels, the signal is fully represented by a single coarse trend $\mathcal{A}^{(L)}$ together with the detail coefficients $\{\mathcal{D}^{(l)}\}_{l=1}^{L}$ removed at each scale, ordered from finest ($l = 1$) to coarsest ($l = L$).

Multi-scale penalty. The Haar transform is an invertible linear mapping, so exact agreement between two signals' approximation and detail coefficients at every level is equivalent to exact agreement between the signals themselves. We exploit this property to supervise the forecast at every scale separately rather than only point-wise: we decompose both the median prediction and the ground-truth future $\tilde{\mathbf{y}} \in \mathbb{R}^H$ with the same L-level transform and penalize the L_1 distance between corresponding coefficients at each level:

$$\mathcal{L}_{\text{wav}} = \beta_A \left\| \mathcal{A}^{(L)}(\hat{\tilde{\mathbf{y}}}_{0.5}) - \mathcal{A}^{(L)}(\tilde{\mathbf{y}}) \right\|_1 + \sum_{l=1}^{L} \beta_l \left\| \mathcal{D}^{(l)}(\hat{\tilde{\mathbf{y}}}_{0.5}) - \mathcal{D}^{(l)}(\tilde{\mathbf{y}}) \right\|_1, \tag{7}$$

where $\beta_l = \gamma \cdot 2^{-(l-1)}$ applies a geometrically decaying weight across detail levels, assigning the largest weight to the finest scale ($l = 1$) and progressively smaller weights toward the coarsest ($l = L$). Finer-level detail coefficients capture differences between adjacent samples and are typically small in magnitude relative to the coarse trend; without compensation, a point-wise loss on the raw signal would be dominated by trend-level error and could under-supervise this fine-grained structure. The decaying weighting compensates for this imbalance by emphasizing the finest scales most strongly while still providing supervision at every level. The approximation weight $\beta_A = \gamma \cdot 2^{-L}$ is not a separate hyperparameter but the natural continuation of this same sequence: since $\mathcal{A}^{(L)}$ is obtained from one additional round of averaging beyond $\mathcal{D}^{(L)}$, it receives the weight the decay formula would assign at level $L{+}1$. The scalar γ controls the overall strength of the wavelet term and is treated as a hyperparameter.

Median-only application. The wavelet regularizer is applied solely to the median quantile ($\tau = 0.5$). The non-central quantiles are meant to model predictive uncertainty and should spread around the ground truth rather than track its trend and fine-scale structure directly; applying the same multi-scale penalty across the full distribution would conflate frequency-domain matching with quantile spread, two conceptually distinct objectives. The pinball loss $\mathcal{L}_{\text{quantile}}$ already supervises the complete predictive distribution; $\mathcal{L}_{\text{wav}}$ provides complementary multi-scale structure supervision only at the central estimate, where matching the ground truth is unambiguously the correct target.

3.5 Training Objective

Let $\hat{\mathbf{y}} \in \mathbb{R}^{Q \times H}$ denote the final quantile predictions produced by passing the fused representation $\mathbf{h}$ through the frozen output projection head of the frozen-foundational model, where Q is the number of predicted quantiles. Let $\mathbf{y} \in \mathbb{R}^H$ denote the ground-truth future. All predictions and targets are computed in the instance-normalized space of the backbone to ensure scale-invariant optimization. The total training objective combines a quantile forecasting loss, a retrieval reconstruction regularizer, and a wavelet regularizer on the median prediction:

$$\mathcal{L} = \mathcal{L}_{\text{quantile}} + \lambda_{\text{recon}}\, \mathcal{L}_{\text{recon}} + \lambda_{\text{wav}}\, \mathcal{L}_{\text{wav}}. \tag{8}$$

Quantile forecasting loss. We adopt the pinball (quantile) loss to supervise the full predictive distribution. For a set of quantile levels $\{\alpha_q\}_{q=1}^{Q}$, the loss is defined as:

$$\mathcal{L}_{\text{quantile}} = \frac{1}{Q} \sum_{q=1}^{Q} \sum_{t=1}^{H} \rho_{\alpha_q} \left(\tilde{\mathbf{y}}(t) - \hat{\tilde{\mathbf{y}}}_q(t) \right), \tag{9}$$

where $\tilde{\mathbf{y}}$ and $\hat{\tilde{\mathbf{y}}}_q$ denote the instance-normalized target and the q-th quantile prediction respectively, and ρ_α is the pinball loss function:

$$\rho_\alpha(u) = 2\,|u| \cdot |\mathbf{1}[u \leq 0] - \alpha|\,. \tag{10}$$

Optimization details. Only the reranking g_ψ, the projection network f_{proj}, and the gating network are trained; both the backbone encoder and the decoder remain fully frozen. For multivariate series with V variates, each variate is treated independently through the retrieval and fusion pipeline, and the losses are averaged across all variates. Although the backbone's parameters are not updated, gradients propagate through its frozen layers during backpropagation, allowing the upstream trainable components to be updated.

4 Experiments

We compare the effectiveness of our framework on established long-term forecasting benchmarks against both task-specific supervised models and zero-shot foundation model baselines. Beyond aggregate accuracy, we investigate the individual role of each architectural component through targeted ablations, and characterize the conditions under which learned retrieval provides the most substantial gains over a frozen backbone.

4.1 Experimental Setup

Datasets. Following the evaluation protocol of RAFT (Han et al., 2025), we benchmark across diverse time series datasets spanning multiple domains, frequencies, and variate counts. **(1–4)** The ETT datasets (ETTh1, ETTh2, ETTm1, ETTm2) record electricity transformer temperatures at hourly and 15-minute granularities over two years (Zhou et al., 2021). **(5)** The *Exchange* dataset contains the daily exchange rates of eight countries from 1990 to 2016 (Lai et al., 2018). **(6)** The *Weather* dataset contains 21 meteorological indicators measured at a German weather station over one year (Wu et al., 2021). Across these datasets, prediction horizons are set to $H \in \{96, 192, 336, 720\}$, following standard protocol (Wu et al., 2021)

Evaluation metrics. All baseline methods are evaluated using Mean Squared Error (MSE) and Mean Absolute Error (MAE) computed over the prediction horizon, averaged across all variates. Lower values indicate better performance.

Baselines. We compare against three groups of methods. The first is a set of supervised, task-specific forecasters spanning the main architectural families used in long-term forecasting: a patch-based Transformer (Nie et al., 2023), linear and MLP-based mixing models (Zeng et al., 2023; Wang et al., 2024b), and a convolution-based model (Wu et al., 2023). These models are trained directly on each target dataset and serve as a reference for the upper bound of task-specific optimization rather than as direct competitors, since they require full retraining per dataset and do not target zero-shot generalization. The second group is RAFT (Han et al., 2025), a retrieval-augmented forecasting framework trained in-domain and the most direct point of comparison, as it shares our retrieval-augmented paradigm but selects candidates via correlation-based matching rather than a learned, reconstruction- supervised reranker. The third group is Chronos-Bolt (Ansari et al., 2024) evaluated zero-shot, which serves as both our backbone and the reference point for measuring the gain attributable to retrieval augmentation.

Implementation Details. Our framework is built on top of two frozen pretrained backbones: *amazon/chronos-bolt-base*, which serves as the forecasting backbone, and *amazon/chronos-t5-base*, which provides token embeddings for both the knowledge base and query encoding. All backbone parameters are frozen. The retrieval knowledge base consists of approximately 2.1 million context–horizon pairs extracted from the Chronos TSMixup pretraining corpus (Ansari et al., 2024), with context length $L = 720$ and prediction horizon $H = 96$. Embeddings for all knowledge base entries are precomputed offline via the frozen

Table 1: Averaged forecasting results over horizons $H = \{96, 192, 336, 720\}$, evaluated by MSE and MAE (lower is better). **Bold**: best result per row; underline: second best. † Chronos-Bolt is evaluated in zero-shot mode and serves as the direct backbone reference for our method. Our framework achieves the best MSE on the majority of datasets and consistently improves over the frozen Chronos-Bolt backbone in MSE across all datasets, demonstrating that learned retrieval augmentation provides predictive signal beyond what the pretrained backbone encodes in its weights.

	Ours		RAFT		Chronos-Bolt†		TimeMixer		PatchTST		TimesNet		DLinear	
Dataset	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.396**	0.409	0.420	0.436	0.417	**0.408**	0.447	0.440	0.516	0.484	0.495	0.450	0.461	0.481
ETTh2	**0.334**	**0.369**	0.359	0.409	0.360	0.379	0.364	0.395	0.391	0.412	0.414	0.427	0.563	0.532
ETTm1	0.356	**0.367**	**0.348**	0.378	0.409	0.380	0.381	0.396	0.406	0.408	0.400	0.406	0.404	0.422
ETTm2	0.264	**0.306**	**0.254**	0.320	0.282	0.314	0.275	0.323	0.290	0.334	0.291	0.333	0.354	0.402
Weather	**0.228**	**0.254**	0.241	0.286	0.251	0.270	0.484	0.272	0.529	0.286	0.620	0.294	0.625	0.321
Exchange	**0.367**	0.407	0.441	0.441	0.400	0.445	0.386	0.416	0.564	0.443	0.416	**0.404**	0.643	0.478

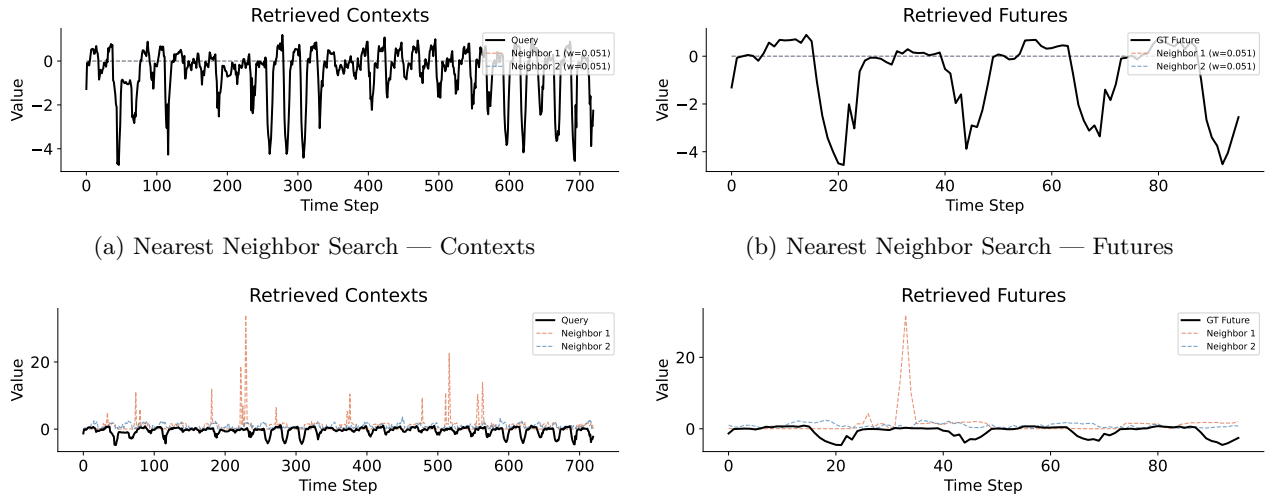
Chronos-T5 encoder and indexed with nearest-neighbor search. The reconstruction regularization weight λ_{recon} and wavelet loss weight λ_{wavelet} in the training objective (Eq. 8) are set based on grid search. For multivariate series, we adopt a channel-independent strategy, processing each variate independently through the retrieval and fusion pipeline. All experiments are conducted on 4× NVIDIA A6000 GPUs, with a single training run completing in approximately 4 hours. All hyperparameters are selected on the held-out validation split to prevent test-set leakage.

All experiments use an input context length of $L = 720$, matching the grid-searched lookback window reported by RAFT (Han et al., 2025) for every dataset used in this comparison. As the frozen Chronos-Bolt backbone natively predicts at most 64 steps per forward pass, forecasts at horizons $H > 64$ are produced through iterative rolling: at each step the model generates up to 64 future values, the median ($\tau = 0.5$) quantile prediction is appended to the context, the context window is truncated to the most recent L steps, and the process repeats until the full horizon H is reached. This rolling strategy follows the approach adopted by TS-RAG (Ning et al., 2025), which similarly extends retrieval-augmented forecasts beyond the backbone's native prediction length through iterative rollout. At each roll, retrieval and reranking are repeated using the updated context, allowing the retrieved candidates to adapt as the forecast extends rather than remaining fixed to the initial query. Training is performed exclusively at the backbone's native horizon , the iterative rolling procedure described above is applied only at inference time for horizons $H > 64$.

4.2 Main Results

Our method achieves the best MSE on four of six datasets and the second-best on the remaining two, where RAFT records a marginally lower MSE on ETTm1 and ETTm2. On MAE, our method achieves the best or tied-best result on four of six datasets. The consistent MSE improvement over the frozen Chronos-Bolt backbone across all datasets confirms that the reranker and fusion components extract predictive signal that the pretrained backbone cannot recover from its weights alone. The most pronounced gains appear on Weather and Exchange MSE, where our method outperforms all supervised baselines by a notable margin, suggesting that retrieval augmentation is particularly effective when the target distribution contains temporal patterns well-represented in the cross-domain knowledge base.

Figure 2 illustrates qualitatively why naive embedding-distance retrieval can be insufficient. Under nearest-neighbor search, both retrieved neighbors do not explicitly align with the query, indicating the embedding space treats structurally dissimilar contexts as equidistant from the query; their futures are correspondingly flat and uninformative relative to the ground truth's sharp decline. The re-ranking stage selects neighbors whose context dynamics more closely resemble the query, yielding retrieved futures with more varied local structure.



(a) Nearest Neighbor Search — Contexts

(b) Nearest Neighbor Search — Futures

(c) Re-Ranking — Contexts

(d) Re-Ranking — Futures

Figure 2: Top-2 retrieved contexts (left) and their corresponding futures (right) for a representative ETTh1 query under Nearest Neighbor Search (top) and Re-Ranking (bottom). Naive Nearest Neighbor search retrieves smooth, periodic neighbors structurally unlike the query; re-ranking selects neighbors with more similar local dynamics, producing futures that better reflect the query's behavior.

4.3 Ablation Studies

We conduct a comprehensive set of ablation experiments to isolate the contribution of each architectural and training design choice. Unless stated otherwise, all ablations use ETTh1 and ETTm1 with prediction horizon $H = 96$.

4.3.1 Effect of Retrieval Strategy and Auxiliary Objectives

To isolate the contribution of each architectural component, we evaluate three progressive configurations that incrementally extend the simplest retrieval baseline toward the full proposed framework. This design allows us to assess whether learned retrieval improves over static embedding-distance selection, and whether the wavelet auxiliary loss provides additional benefit beyond the reranker alone. Hence we evaluate three configurations to disentangle the contributions of each module:

- **Retrieval:** Top-K neighbors retrieved by L2 distance in frozen Chronos-T5 decoder embedding space; retrieval weights are assigned as $\mathbf{w} = \text{softmax}(-\mathbf{d})$, where $\mathbf{d}$ is the distance.
- **Retrieval + Re-ranking:** A residual MLP reranker refines frozen decoder embeddings to produce learned similarity scores, with the top-K neighbors selected from a larger candidate pool of size $M = 1000$. A reconstruction regularization loss $\mathcal{L}_{\text{recon}}$ is added to align retrieved contexts with the query in normalized space.
- **Retrieval + Re-ranking + Wavelet (Ours):** The full proposed model, adding the Haar wavelet auxiliary loss $\mathcal{L}_{\text{wav}}$ applied exclusively to the median quantile ($\tau = 0.5$) forecast, atop the reranker.

Figure 3 show three observations. Naive nearest neighbor search already improves over the frozen Chronos-Bolt baseline, indicating that the historical context bank contains useful predictive signal even without learned selection. Adding the reconstruction-supervised reranker makes these gains more stable and task-aligned. The Haar wavelet loss then yields the strongest overall configuration, with a larger reduction on MSE than on MAE, consistent with multi-scale regularization helping to suppress larger forecast deviations rather than only the average absolute error.

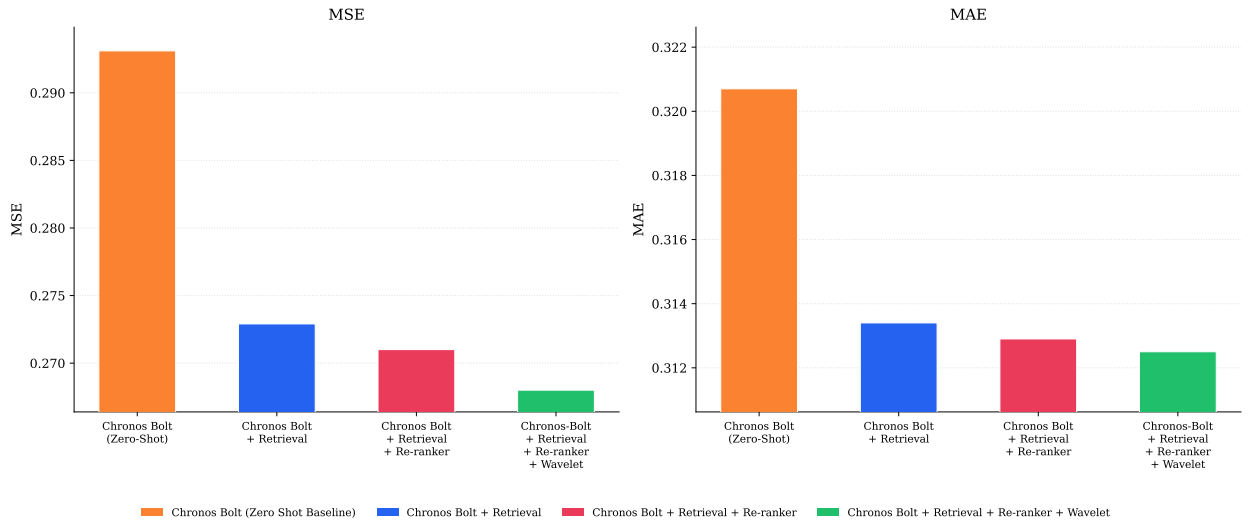


Figure 3: Retrieval Strategy pipeline aggregated across all four ETT datasets at $H = 96$. Each bar reports the mean MSE (left) and MAE (right) averaged over the four per-dataset means, each itself computed as the mean over three random seeds.

Table 2: Fusion strategy ablation across all four ETT datasets ($H = 96$). Soft gating outperforms additive residual fusion on all eight reported columns, indicating that adaptive modulation of retrieval influence is more effective than unconditional addition. Chronos-Bolt zero-shot serves as a retrieval-free reference. **Bold**: best result per column; underline: second best.

Fusion Strategy	**ETTh1**		**ETTh2**		**ETTm1**		**ETTm2**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Chronos-Bolt (Zero-Shot)†	0.3769	0.3774	0.2903	0.3271	0.3266	0.3330	0.1803	0.2461
Additive Residual	0.3687	0.3805	0.2834	0.3265	0.3156	0.3321	0.1799	0.2460
Soft Gating (Ours)	**0.3462**	**0.3737**	**0.2677**	**0.3170**	**0.2873**	**0.3205**	**0.1642**	**0.2379**

4.3.2 Effect of Fusion Strategy

The degree to which retrieved futures should influence the final forecast may vary depending on how well the retrieved candidates align with the query. This ablation examines whether a fusion mechanism that can modulate retrieval influence outperforms one that applies it unconditionally. We compare two configurations while holding all other components fixed:

- **Additive Residual:** The retrieval projection $\mathbf{g}$ is added directly to the decoder state $\mathbf{e}^q$ without any gating, $\mathbf{h} = \mathbf{e}^q + \mathbf{g}$, applying the retrieval signal unconditionally regardless of its relevance.
- **Soft Gating (Ours):** Dimension-wise interpolation via learned gates as described in Eq. 4, allowing the model to learn on a per-query basis how strongly to defer to the retrieved signal versus the backbone's own representation.

The soft gating formulation is strictly more expressive than additive fusion: when the gate collapses to a constant, it recovers the additive case, while learned per-dimension weights allow the model to selectively amplify or suppress individual feature channels of the retrieval signal depending on how well the retrieved candidates match the structural properties of the current query.

Table 2 shows that soft gating records lower MSE and MAE than additive residual fusion on all eight reported columns across all four ETT datasets. The consistent gap suggests that unconditional addition

Table 3: Effect of top-K across all four ETT datasets ($H = 96$). Chronos-Bolt zero-shot serves as a retrieval-free reference. $K = 20$ achieves the best or near-best MSE across all four datasets, indicating that a moderate candidate pool balances retrieval diversity against the inclusion of lower-relevance neighbors; performance degrades modestly at $K = 30$, confirming diminishing returns beyond the optimal pool size. **Bold**: best result per column; underline: second best.

K	**ETTh1**		**ETTh2**		**ETTm1**		**ETTm2**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Chronos-Bolt(Zero-Shot)	0.3769	0.3774	0.2903	0.3271	0.3266	0.3330	0.1803	0.2461
5	0.3525	**0.3721**	0.2769	0.3248	0.2962	0.3220	0.1770	0.2470
10	0.3537	0.3730	0.2731	0.3206	0.2930	0.3216	0.1681	0.2411
20	**0.3462**	0.3737	**0.2670**	**0.3170**	**0.2873**	**0.3205**	**0.1660**	**0.2380**
30	0.3470	0.3739	0.2716	0.3189	0.2939	0.3231	0.1667	0.2385

of the retrieval signal is harmful when that signal is poorly aligned with the query, and that a learned per-dimension gate provides the flexibility needed to suppress unhelpful retrieval corrections selectively.

4.3.3 Effect of Number of Retrieved Neighbors (K)

We study how the number of top-K neighbors selected after reranking affects forecast accuracy. We vary $K \in \{5, 10, 20, 30\}$ while keeping the coarse candidate pool fixed at $M = 1000$. Table 3 shows that across all four datasets, MSE follows the same overall trajectory: retrieval benefit grows as K increases, but this gain is conservative. Majority of the datasets reach retrieval saturation at $K = 20$, with only a small variation at higher numbers where additional retrieval no longer improves the forecast enough to offset the inclusion of lower-relevance candidates. The contribution of each additional neighbor is therefore not constant: it is substantial when the pool is small and the model is selecting from a limited set of candidates, but it gradually decreases and becomes marginal once the pool already contains the most relevant matches.

4.3.4 In-Domain vs. Cross-Domain Knowledge Base

To evaluate the importance of knowledge base diversity, we compare two retrieval database configurations:

- **In-domain KB:** The knowledge base is constructed exclusively from the target dataset's own training split, using the same context length $L = 720$ and forecast horizon $H = 96$ as the cross-domain configuration. This represents the conventional retrieval assumption that the most useful neighbors come from the same domain as the query.
- **Cross-domain KB (Ours):** The knowledge base is built from the Chronos TSMixup pretraining corpus, spanning roughly 2.1M context–horizon pairs aggregated from heterogeneous domains and sampling frequencies, with no overlap with any ETT dataset. The reranker and fusion pipeline are identical to the in-domain configuration, with no architectural or training changes specific to this setting. This configuration tests whether the learned reranker can identify predictively useful candidates without relying on domain proximity between the query and the retrieved pattern.

Table 4 shows that the cross-domain KB outperforms the in-domain KB on both metrics across all three datasets, despite containing no ETT data. The more informative observation is not that a larger corpus helps, but that the learned reranker successfully identifies structurally relevant patterns from a heterogeneous external memory without any domain-specific adaptation. The in-domain KB, constructed from the target dataset's own training split, is consistently outperformed by the cross-domain KB and in two of three datasets even by the frozen zero-shot baseline, indicating that a small in-domain pool with limited pattern diversity offers weaker retrieval candidates than a diverse corpus with a well-trained selector. Together, these results

Table 4: Knowledge base generalization at $H = 96$. *In-domain KB* is constructed from the target dataset's own training split; *Cross-domain KB* is the Chronos TSMixup pretraining corpus (∼2.1M pairs, no ETT overlap). Chronos-Bolt† zero-shot serves as a retrieval-free reference. **Bold**: best result per column; underline: second best. The cross-domain KB outperforms the in-domain KB across all datasets without any domain-specific adaptation; notably, the in-domain KB is outperformed by the frozen zero-shot baseline on six of eight columns, confirming that a small domain-matched pool with limited pattern diversity offers weaker retrieval signal than a large heterogeneous corpus with a well-trained reranker.

Test Dataset	**In-domain KB**		**Cross-domain KB (Ours)**		**Chronos-Bolt†**	
	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.371	0.379	**0.346**	**0.373**	0.376	0.377
ETTh2	0.273	0.321	**0.267**	**0.317**	0.290	0.327
ETTm1	0.370	0.389	**0.291**	**0.322**	0.326	0.333
ETTm2	0.201	0.287	**0.166**	**0.238**	0.180	0.246

Table 5: Effect of Haar decomposition level L on test MSE and MAE ($H = 96$). **Bold**: best per column; underline: second best. $L = 6$ is the theoretical maximum for $H = 96$.

L	**ETTh1**		**ETTh2**		**ETTm1**		**ETTm2**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
3	0.3643	0.3746	**0.2672**	**0.3169**	0.2938	**0.3213**	0.1637	0.2369
4	0.3636	0.3745	0.2694	0.3175	0.2944	**0.3213**	**0.1635**	**0.2368**
5	0.3566	0.3730	0.2704	0.3180	0.2955	0.3217	0.1648	0.2379
6	**0.3532**	**0.3721**	0.2707	0.3183	**0.2935**	0.3215	0.1657	0.2381

suggest that the retrieval mechanism generalizes beyond domain-specific memorization and can transfer predictive signal from unrelated domains without retraining.

4.3.5 Effect of Wavelet Decomposition Level

We evaluate the sensitivity of forecast accuracy to the number of Haar decomposition levels $L \in \{3, 4, 5, 6\}$, with all other hyperparameters held fixed at their default values. The maximum feasible level for $H = 96$ is $L = 6$ ($\lfloor \log_2 96 \rfloor = 6$). Results are shown in Table 5.

Table 5 shows that optimal decomposition level varies by dataset and no single value of L dominates across all four benchmarks. On ETTh1, finer decomposition is consistently preferred, with accuracy improving monotonically from $L = 3$ to $L = 6$. On ETTh2, the opposite trend holds, with fewer levels yielding lower error. On ETTm1 and ETTm2, the results are substantially more stable.

4.3.6 Effect of Loss Weights

The two auxiliary loss weights λ_{recon} and λ_{wav} (Eq. 8) control the relative contributions of the reconstruction and wavelet objectives on top of the primary pinball forecasting loss. We assess the sensitivity of both weights via one-axis sweeps around our chosen configuration on ETTh1 and ETTm1 ($H = 96$), with the wavelet decomposition level fixed at $L = 6$ and the number of retrieved neighbors fixed at $K = 20$ throughout; full results are reported in Fig. 4 and Fig. 5. We find that performance is broadly stable across the swept range for both weights, with no single setting required for competitive accuracy.

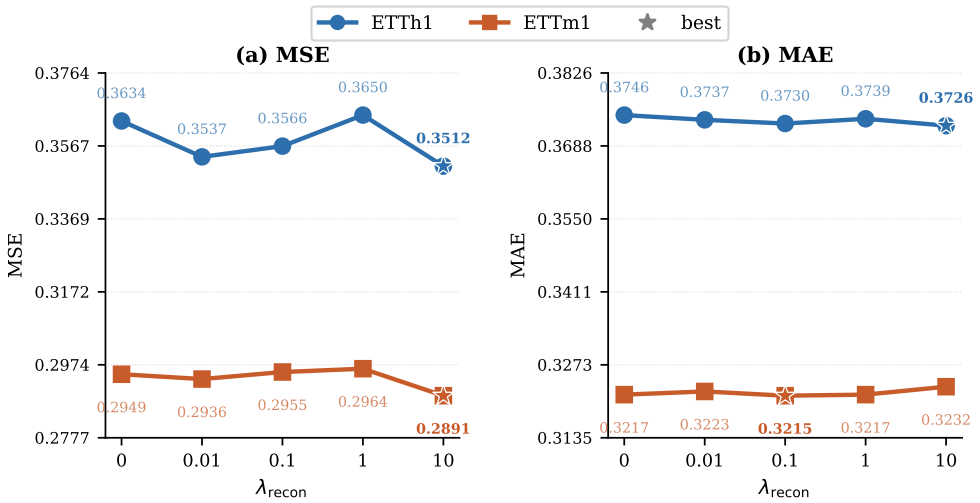


Figure 4: Sensitivity of MSE (left) and MAE (right) to λ_{recon} on ETTh1 and ETTm1 ($H = 96$). Stars mark the per-dataset optimum. Increasing λ_{recon} consistently improves MSE on both datasets, with the lowest error achieved at the largest tested value ($\lambda_{\text{recon}} = 10$), suggesting the reconstruction signal primarily benefits squared-error minimization.

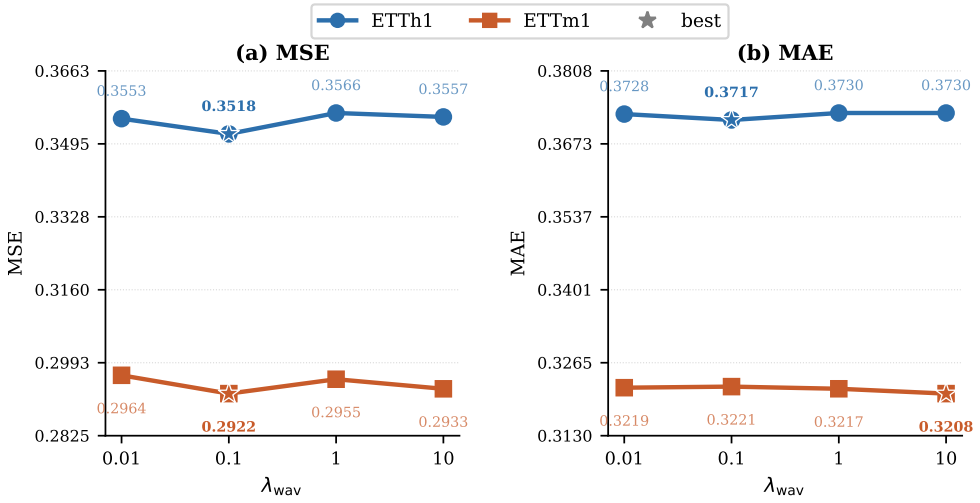


Figure 5: Sensitivity of MSE (left) and MAE (right) to λ_{wav} on ETTh1 and ETTm1 ($H = 96$). Stars mark the per-dataset optimum. MSE is best at $\lambda_{\text{wav}} = 0.1$ on both datasets and rises when the wavelet weight is increased further, indicating that excessive wavelet supervision diverts optimization pressure away from squared-error minimization.

Table 6: Transfer learning from **ETTh1** to target datasets.

Target	**Transfer (ETTh1→)**		**In-Domain**		**Zero-Shot**	
	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	<u>0.2825</u>	0.3346	**0.2678**	**0.3174**	0.2903	<u>0.3274</u>
ETTm1	0.3620	0.3828	**0.2910**	**0.3221**	<u>0.3266</u>	<u>0.3330</u>
ETTm2	<u>0.1783</u>	0.2628	**0.1664**	**0.2384**	0.1803	<u>0.2461</u>

Table 7: Transfer learning from **ETTh2** to target datasets.

Target	Transfer (ETTh2→)		In-Domain		Zero-Shot	
	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.3688	0.3767	**0.3516**	**0.3736**	0.3769	0.3774
ETTm1	0.3291	0.3516	**0.2910**	**0.3221**	0.3266	0.3330
ETTm2	0.1759	0.2498	**0.1664**	**0.2384**	0.1803	0.2461

Table 8: Transfer learning from **ETTm1** to target datasets.

Target	Transfer (ETTm1→)		In-Domain		Zero-Shot	
	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.3836	0.4127	**0.3516**	**0.3736**	0.3769	0.3774
ETTh2	0.3050	0.3447	**0.2678**	**0.3174**	0.2903	0.3271
ETTm2	0.1753	0.2484	**0.1664**	**0.2384**	0.1803	0.2461

4.3.7 Transfer Learning Across Datasets

To assess our model's generalization beyond its training domain, we conduct cross-dataset transfer experiments in which a single model is trained on a source ETT dataset and evaluated on held-out target datasets *without any fine-tuning*. Results are reported at horizon H=96 (MSE / MAE) and compared against two reference points: **In-domain** — model trained and tested on each target dataset independently and **Chronos (Zero-shot)**, the frozen Chronos-Bolt backbone evaluated with no task-specific supervision. *In Tables 6–9, **bold** denotes the best result per row and underline denotes the second best. All experiments are at H=96.*

Several consistent patterns emerge across the four transfer configurations tables (6, 7, 8, 9). Transfer performance is strongest when the source and target share the same temporal granularity, demonstrating that the retrieval and fusion pipeline can carry useful predictive structure across datasets when their sampling frequencies are aligned. Cross-granularity transfer is systematically weaker, and in most cases fails to improve over the zero-shot backbone, suggesting that the reranker's similarity judgments are sensitive to differences in temporal scale that are not normalized away by instance normalization alone. ETTm2 as a source dataset produces the largest degradation in all three target evaluations, indicating that its learned retrieval weights encode patterns specific to its statistical regime that do not generalize. Across all configurations, transfer performance remains below the in-domain reference, as expected, but the fact that same-granularity transfer can match or surpass a frozen pretrained model without any fine-tuning on the target domain demonstrates that the learned retrieval mechanism captures structurally transferable patterns rather than purely memorizing the training distribution.

5 Conclusion

We presented a retrieval-augmented forecasting framework that augments a frozen Chronos-Bolt backbone with learned retrieval and multi-scale wavelet supervision. The central contribution is a learnable reranker trained through a reconstruction regularizer that couples candidate ranking directly to predictive coherence, ensuring that retrieved contexts are selected for their temporal representativeness rather than their proximity in a generic embedding space. A multi-level Haar wavelet loss applied to the median quantile provides complementary multi-scale supervision, encouraging the model to match both the coarse trend and fine-grained structure of the target forecast.

Experiments on six benchmarks show that learned retrieval augmentation consistently improves over the frozen Chronos-Bolt zero-shot baseline, and ablations confirm that both the reconstruction regularizer and the wavelet loss contribute to this improvement. Within our framework, the cross-domain knowledge base

Table 9: Transfer learning from **ETTm2** to target datasets.

Target	Transfer (ETTm2→)		In-Domain		Zero-Shot	
	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.4661	0.4673	**0.3516**	**0.3736**	0.3769	0.3774
ETTh2	0.3285	0.3782	**0.2678**	**0.3174**	0.2903	0.3271
ETTm1	0.5767	0.4941	**0.2910**	**0.3221**	0.3266	0.3330

constructed from the Chronos TSMixup pretraining corpus outperforms an in-domain knowledge base across all evaluated datasets. This advantage likely reflects two compounding factors: scale, as the cross-domain corpus provides ∼2.1M diverse context–horizon pairs against a comparatively small in-domain training split; and learned selection, as the reconstruction-supervised reranker identifies predictively relevant sequences from a heterogeneous corpus that static distance metrics alone would fail to separate.

Finally, the channel-independent design discards cross-variate dependencies that are predictive in highly correlated multivariate settings such as Electricity and Traffic; extending the reranker to operate on joint multivariate embeddings while preserving the scalability of per-variate retrieval is a natural but non-trivial extension.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

William J. Bosl, Tobias Loddenkemper, and Solveig Vieluf. Coarse-Graining and the Haar Wavelet Transform for Multiscale Analysis. *Bioelectronic Medicine*, 8(1):3, 2022. doi: 10.1186/s42234-022-00085-z. URL https://springer.com.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Ingrid Daubechies. *Ten Lectures on Wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992. doi: 10.1137/1.9781611970104. URL https://siam.org.

Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org.

Ziyou Guo, Yan Sun, and Tieru Wu. WEITS: A Wavelet-Enhanced Residual Framework for Interpretable Time Series Forecasting. *Neurocomputing*, 674:132788, 2026. doi: 10.1016/j.neucom.2026.132788. URL https://www.sciencedirect.com/science/article/pii/S0925231226001852.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL http://mlr.press.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2505.04163*, 2025.

Baoyu Jing, Si Zhang, Yada Zhu, Bin Peng, Kaiyu Guan, Andrew Margenot, and Hanghang Tong. Retrieval based time series forecasting. *arXiv preprint arXiv:2209.13525*, 2022.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*, 2022.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with Deep Neural Networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–104, 2018. doi: 10.1145/3209978.3210006. URL https://dl.acm.org/doi/10.1145/3209978.3210006.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-Augmented Diffusion Models for Time Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://openreview.net/forum?id=dRJJt0Ji48.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pp. 9881–9893, 2022.

Stéphane G. Mallat. A theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. doi: 10.1109/34.192463. URL https://ieee.org.

Md Mahmuddun Nabi Murad, Mehmet Aktukmak, and Yasin Yilmaz. WPMixer: Efficient Multi-Resolution Mixing for Long-Term Time Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19581–19588, 2025. doi: 10.1609/aaai.v39i18.34156. URL https://ojs.aaai.org/index.php/AAAI/article/view/34156.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, 2023. arXiv:2211.14730.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Y. Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. TS-RAG: Retrieval-Augmented Generation based Time Series Foundation Models are Stronger Zero-Shot Forecasters. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=TJuUelhGQr.

Donald B. Percival and Andrew T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. doi: 10.1017/CBO9780511841040. URL https://doi.org.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2024.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001. URL https://doi.org.

Kutay Tire, Ege Onur Taga, M. Emrullah Ildiz, and Samet Oymak. Retrieval Augmented Time Series Forecasting. In *International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, 2026. URL https://openreview.net/forum?id=UD76JhLswg.

Hai-Kun Wang, Xuewei Zhang, Haicheng Long, Shunyu Yao, and Pengjin Zhu. W-FENet: Wavelet-based Fourier-Enhanced Network Model Decomposition for Multivariate Long-Term Time-Series Forecasting. *Neural Processing Letters*, 56(1):43, 2024a. doi: 10.1007/s11063-024-11478-3. URL https://link.springer.com/article/10.1007/s11063-024-11478-3.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=7oLshfEIC2.

Tianfeng Wang and Gaojie Cui. RATSF: Empowering Customer Service Volume Management through Retrieval-Augmented Time-Series Forecasting. *arXiv preprint arXiv:2403.04180*, 2024. URL https://arxiv.org/abs/2403.04180.

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A Multi-Horizon Quantile Recurrent Forecaster. *arXiv preprint arXiv:1711.11053*, 2017. URL https://arxiv.org.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. arXiv:2402.02592.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with Auto-Correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pp. 22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-Variation modeling for general time series analysis. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, 2023. arXiv:2210.02186.

Wenxuan Xie and Fanpu Cao. SWIFT: Mapping Sub-series with Wavelet Decomposition Improves Time Series Forecasting. *arXiv preprint arXiv:2501.16178*, 2025. URL https://arxiv.org/abs/2501.16178.

Silin Yang, Dong Wang, Haoqi Zheng, and Ruochun Jin. TimeRAG: Boosting LLM Time Series Forecasting via Retrieval-Augmented Generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. doi: 10.1109/ICASSP49660.2025.10889933. URL https://arxiv.org/abs/2412.16643.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. URL https://ojs.aaai.org/index.php/AAAI/article/view/26317.

Huanyu Zhang, Chang Xu, Yi-Fan Zhang, Zhang Zhang, Liang Wang, Jiang Bian, and Tieniu Tan. TimeRAF: Retrieval-Augmented Foundation Model for Zero-Shot Time Series Forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 37(8):5654–5667, 2025. doi: 10.1109/TKDE.2025.11031238. URL https://ieee.org.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

A Appendix

A.1 Dataset Statistics

Table 10 summarizes the statistics of all datasets used in our evaluation. For ETT datasets, we follow the standard 6:2:2 chronological train/validation/test split; for Exchange and Weather, we follow the 7:1:2 split convention established by prior work (Wu et al., 2021; Nie et al., 2023).

Table 10: Dataset statistics. Dataset Size denotes the number of time points in (Train, Validation, Test) splits. Frequency denotes the sampling interval.

Dataset	Variates	Dataset Size	Frequency	Domain	Split
ETTh1	7	(8545, 2881, 2881)	1 hour	Electricity	6:2:2
ETTh2	7	(8545, 2881, 2881)	1 hour	Electricity	6:2:2
ETTm1	7	(34465, 11521, 11521)	15 min	Electricity	6:2:2
ETTm2	7	(34465, 11521, 11521)	15 min	Electricity	6:2:2
Exchange	8	(5120, 665, 1422)	Daily	Economics	7:1:2
Weather	21	(36792, 5271, 10540)	10 min	Meteorology	7:1:2

A.2 Baseline Methods

PatchTST (Nie et al., 2023) PatchTST segments time series into fixed-length patches and applies a Transformer encoder to capture long-range temporal dependencies with reduced sequence length. A channel-independent design is adopted, processing each variate separately through the same set of parameters.

DLinear (Zeng et al., 2023) DLinear decomposes the input into trend and remainder components using a moving average kernel and applies a separate linear layer to each, serving as a strong and lightweight baseline that challenges more complex architectures.

TimeMixer (Wang et al., 2024b) TimeMixer is a fully MLP-based architecture that performs multiscale mixing of decomposed seasonal and trend components in both bottom-up and top-down directions across downsampled resolutions of the input.

TimesNet (Wu et al., 2023) TimesNet transforms 1D time series into 2D representations by reshaping the signal according to its dominant periods, then applies a parameter-efficient Inception block to capture intraperiod and interperiod temporal variations jointly.

RAFT (Han et al., 2025) RAFT augments in-domain forecasting models by retrieving historical segments whose contexts are most similar to the query via multi-period matching and Pearson correlation, then incorporates their future values as additional input alongside the query context.

Chronos-Bolt (Ansari et al., 2024) Chronos-Bolt is a patch-based encoder-decoder foundation model for probabilistic time series forecasting, pretrained on a large corpus of real and synthetic data and evaluated here in zero-shot mode without any task-specific fine-tuning.

A.3 Knowledge Base Construction

The retrieval knowledge base is constructed offline prior to training and remains fixed throughout all experiments. We construct and evaluate two configurations of the knowledge base. The **cross-domain KB** is built from the Chronos TSMixup pretraining corpus (Ansari et al., 2024), spanning heterogeneous domains and sampling frequencies with no overlap with any evaluation dataset used in this paper. The **in-domain**

Table 11: Selected hyperparameters per dataset. K: number of retrieved neighbors; λ_{wav}: wavelet loss weight; J: number of Haar decomposition levels; λ_{recon}: reconstruction loss weight. All values chosen by validation MSE.

Dataset	K	λ_{wav}	J (wavelet levels)	λ_{recon}
ETTh1	20	0.1	6	10
ETTh2	20	0.01	3	0.01
ETTm1	20	0.1	6	10
ETTm2	20	10	4	0.01
Weather	20	0.001	6	0.1
Exchange Rate	20	1	5	10

KB is built exclusively from the training split of the target dataset, providing a smaller but domain-matched pool of retrieval candidates.

For both configurations, context–horizon pairs are extracted using a stride-1 sliding window with context length $L_{\text{kb}} = 720$ and horizon length $H_{\text{kb}} = 64$ (unless specified), discarding any windows shorter than the required size or containing non-finite values. The cross-domain construction yields approximately 2.1 million pairs; the in-domain pool is limited to the number of valid windows extractable from each dataset's training split. Each context window is encoded offline using the frozen `amazon/chronos-t5-base` encoder, extracting the final token embedding to produce a 768-dimensional representation per window. The resulting embeddings are indexed for L2 nearest-neighbor search via FAISS, enabling efficient retrieval at inference time .

A.4 Hyperparameter Selection per Dataset

All hyperparameters are selected via grid search on the validation split, with no access to test labels. The search ranges are: $K \in \{3, 5, 10, 20\}$, $\lambda_{\text{wav}} \in \{0.01, 0.1, 1, 10\}$, $J \in \{3, 4, 5, 6\}$, $\lambda_{\text{recon}} \in \{0, 0.01, 0.1, 1, 10\}$.

A.5 Component Analysis

To systematically evaluate the architectural choices that define RECAP, we conduct a component ablation study to demonstrate how each design layer progressively enhances forecasting performance over the zero-shot baseline.

A.6 Qualitative Results

We present five representative forecasts at $H = 96$, illustrating how the retrieved context guides the final prediction. In each example, the weighted reconstruction $\hat{x} = \sum_k w_k \, x_{q(k)}$ summarizes the query context through the reranker-assigned weights, directly informing the downstream trajectory correction. While the Chronos-Bolt zero-shot baseline tends to revert toward the local context mean, the retrieval-augmented forecast tracks the directional behavior of the ground truth more faithfully.

Qualitative forecasting examples on ETT and Exchange Rate test windows at $H = 96$ shows that context reconstruction weights indeed help in correcting the future trajectory of the zero-shot model prediction, refining the final forecast (green) based on the historical patterns captured in the weighted reconstruction (red dashed). While the Chronos-Bolt zero-shot baseline (orange) tends to revert toward the local context mean, our adaptively augmented prediction successfully tracks the complex directional behavior of the ground truth (black).

Table 12: Component ablation on ETTh1, ETTh2, ETTm1, and ETTm2 ($H = 96$). Each column adds one component cumulatively on top of the previous: **+ Retrieval** augments the frozen backbone with retrieval and weighted aggregation; **+ Re-ranking** replaces the naive weights with a learned residual MLP reranker trained via reconstruction supervision; **+ Wavelet Loss** adds the multi-scale Haar auxiliary regularizer on top of reranker, on the median prediction. Trainable configurations are reported as mean±std over three random seeds. †Zero-shot Chronos-Bolt is deterministic (no training)

Dataset	**Metric**	**Chronos-Bolt (Zero-shot)**†	**+ Retrieval**	**+ Re-ranking**	**+ Wavelet Loss**
ETTh1	MSE	0.3760	$0.3628_{\pm 0.0029}$	$0.3545_{\pm 0.0003}$	$0.3472_{\pm 0.0008}$
	MAE	0.3770	$0.3736_{\pm 0.0007}$	$0.3729_{\pm 0.0006}$	$0.3731_{\pm 0.0005}$
ETTh2	MSE	0.2900	$0.2689_{\pm 0.0007}$	$0.2687_{\pm 0.0024}$	$0.2694_{\pm 0.0012}$
	MAE	0.3270	$0.3178_{\pm 0.0004}$	$0.3181_{\pm 0.0013}$	$0.3181_{\pm 0.0008}$
ETTm1	MSE	0.3266	$0.2942_{\pm 0.0022}$	$0.2958_{\pm 0.0048}$	$0.2909_{\pm 0.0029}$
	MAE	0.3330	$0.3231_{\pm 0.0003}$	$0.3223_{\pm 0.0018}$	$0.3210_{\pm 0.0004}$
ETTm2	MSE	0.1800	$0.1658_{\pm 0.0001}$	$0.1648_{\pm 0.0011}$	$0.1646_{\pm 0.0004}$
	MAE	0.2460	$0.2390_{\pm 0.0001}$	$0.2384_{\pm 0.0004}$	$0.2380_{\pm 0.0001}$

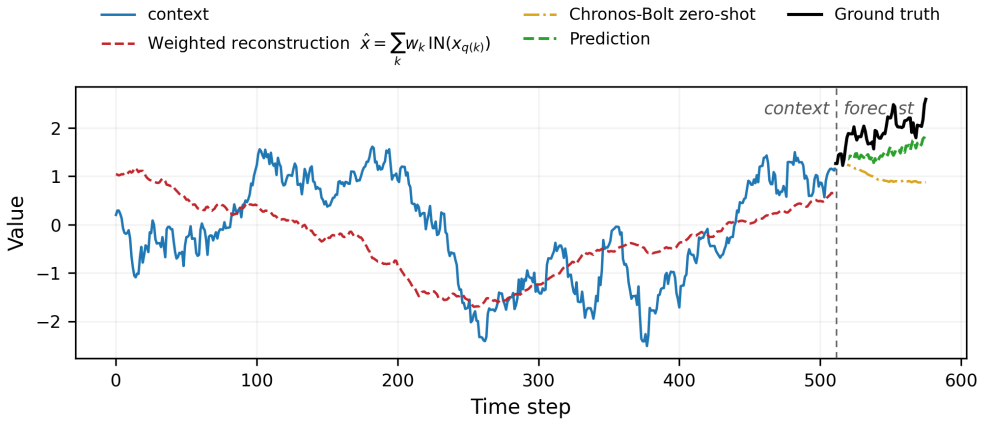


Figure 6: Qualitative forecasting on ETTh1 at $H = 96$.

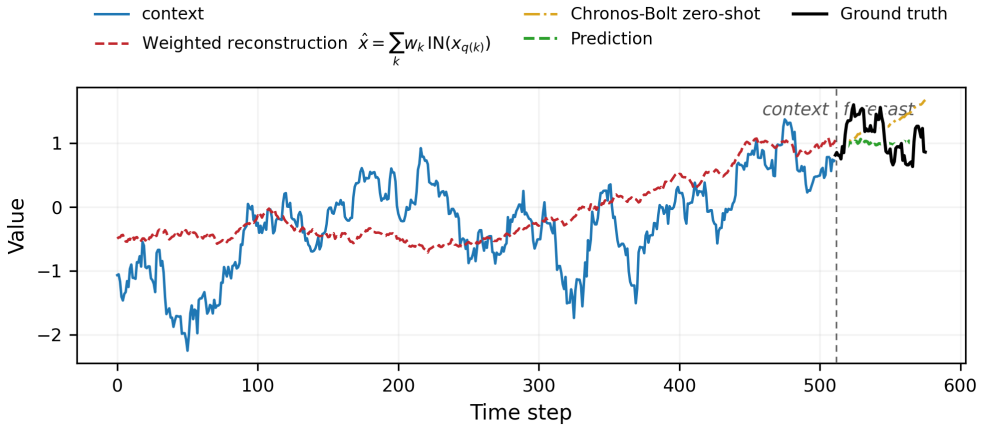


Figure 7: Qualitative forecasting on ETTh1 at $H = 96$.

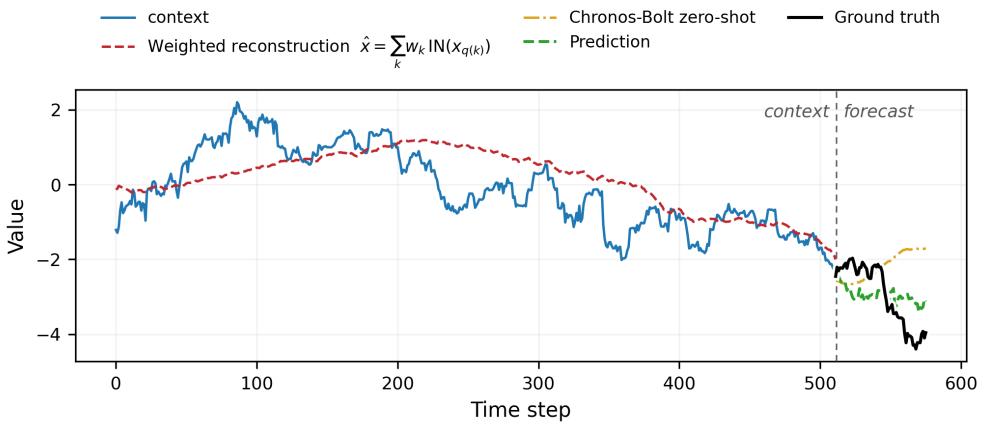


Figure 8: Qualitative forecasting on Exchange Rate at $H = 96$.

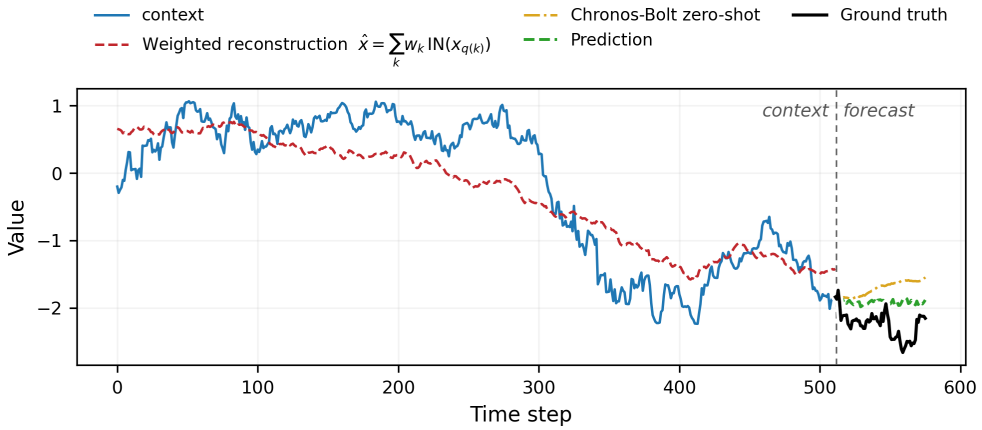


Figure 9: Qualitative forecasting on Exchange Rate at $H = 96$.

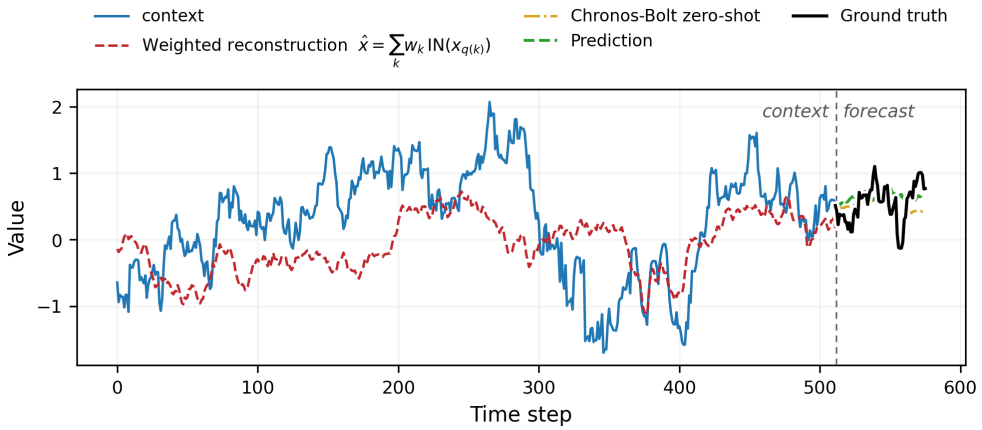


Figure 10: Qualitative forecasting on Exchange Rate at $H = 96$.

Table 13: Forecasting MSE / MAE across prediction horizons $\{96, 192, 336, 720\}$. **Bold**: best.

Dataset	Horizon	Chronos-Bolt (Zeroshot)		RAFT		Ours	
	H	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.376	0.377	0.367	0.397	**0.351**	**0.373**
	192	0.430	0.409	0.411	0.427	**0.394**	**0.407**
	336	0.392	0.399	0.436	0.442	**0.424**	**0.426**
	720	0.473	0.449	0.467	0.478	**0.415**	**0.432**
ETTh2	96	0.290	0.327	0.276	0.344	**0.267**	**0.317**
	192	0.354	0.368	0.347	0.393	**0.333**	**0.362**
	336	0.392	0.399	0.376	0.425	**0.364**	**0.388**
	720	0.404	0.422	0.436	0.473	**0.375**	**0.409**
ETTm1	96	0.326	0.333	0.302	0.349	**0.291**	**0.322**
	192	0.379	0.364	**0.329**	0.367	0.337	**0.355**
	336	0.422	0.390	**0.355**	0.383	0.365	**0.375**
	720	0.509	0.434	**0.406**	**0.413**	0.434	0.417
ETTm2	96	0.180	0.246	**0.164**	0.256	0.166	**0.238**
	192	0.248	0.292	**0.219**	0.296	0.228	**0.284**
	336	0.306	0.331	**0.275**	0.336	0.282	**0.320**
	720	0.395	0.388	**0.359**	0.392	0.380	**0.384**
Weather	96	0.171	0.204	0.164	0.256	**0.149**	**0.186**
	192	0.215	0.246	0.219	0.296	**0.195**	**0.230**
	336	0.270	0.288	0.275	0.336	**0.246**	**0.272**
	720	0.351	0.342	0.359	0.392	**0.325**	**0.328**
Exchange Rate	96	0.101	0.223	0.091	0.209	**0.086**	**0.205**
	192	0.200	0.320	0.205	0.324	**0.191**	**0.309**
	336	0.350	0.429	0.353	0.431	**0.323**	**0.413**
	720	0.951	0.808	1.115	0.801	**0.866**	**0.704**

A.7 Performance Across Longer Forecast Horizons

We additionally evaluate our framework across longer prediction horizons to assess whether retrieval gains persist beyond short-range forecasting. Results are reported in Table 13.

Table 13 presents per-horizon results for the three methods. Our framework consistently outperforms the frozen Chronos-Bolt backbone across all datasets and all evaluated horizons on both metrics, confirming that learned retrieval augmentation provides a reliable improvement over the pretrained backbone regardless of forecast length. Against RAFT, the picture is dataset-dependent: on ETTh1, ETTh2, Weather, and Exchange Rate, our method achieves lower MSE and MAE at every horizon; on ETTm1 and ETTm2, RAFT holds an advantage on MSE, particularly at longer horizons, while our method remains competitive or superior on MAE. A key structural advantage of our approach at longer horizons is that retrieval is not fixed to the initial query: at each roll, the reranker re-scores candidates against the updated context, allowing the retrieved patterns to adapt as the forecast extends rather than anchoring to a single snapshot of the input.