RETRIEVAL IS DENSITY ESTIMATION: CALIBRATED, INVARIANT, AND SELECTIVE RETRIEVAL-AUGMENTED FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot forecasts from time-series foundation models such as Chronos and MOIRAI show severe performance drops when test distributions diverge from the pretraining data. While retrieval augmentation is intended to bridge this out-of-distribution generalization gap by supplying external memory, existing methods often fail. Because existing methods misalign unnormalized series via Euclidean matching, collapse predictive uncertainity through latent pooling, and unconditioned retrieval forces neighbor injection even when candidate trajectories diverge. Across standard benchmarks, unconditioned retrieval degrades baseline accuracy on 41–50% of test windows. We frame retrieval-augmented forecasting as non-parametric conditional density estimation over a scale-translation quotient manifold $\mathcal{X}/\mathcal{G}$. By mapping lookback windows onto invariant shape orbits on $\mathcal{X}/\mathcal{G}$, our approach contracts the covering space ($R^2 = 0.9845$, scaling exponent $\alpha = 0.2199$) and thereby eliminates the need for post-hoc affine alignment. To aggregate forecast distributions, we use weighted 2-Wasserstein barycenters in quantile space. When local candidate scattering indicates negative expected utility, an instance-adaptive controller selectively abstains from retrieval. Under abrupt distribution shifts, a non-exchangeable split-conformal step expands prediction intervals to restore nominal coverage. Across four ETT datasets and three foundation models (Chronos-Bolt, Moirai-2.0, TimesFM-2.5), Caliberated Invariance-Aware Selective Retrieval (CAVIR) lowers CRPS by up to 2.15%, achieves 80.1% coverage under macroeconomic shocks, and stays within 0.85% of baseline error under strict deduplication checks.

1 INTRODUCTION

Frozen time-series foundation models, such as Chronos (Ansari et al., 2024) and MOIRAI (Liu et al., 2025), struggle with abrupt distributional shifts because they cannot update their parameters during inference. This leads to high predictive uncertainty and inaccurate forecasts on out-of-distribution data.

Retrieval-augmented forecasting addresses this limitation by pulling historical patterns from an external memory pool (Yang et al., 2024; Ning et al., 2025). However, text-derived retrieval methods struggle with numerical series: Euclidean matching fails under scale and baseline disparities, latent pooling collapses predictive uncertainty, and unconditional neighbor injection introduces excess variance. On the ETTh1 and ETTm2 benchmarks, blind retrieval degrades baseline accuracy in 41.59% and 49.62% of forecast windows, respectively (Ning et al., 2025), indicating that external memory must be engaged selectively.

In our approach, we treat retrieval-augmented forecasting as non-parametric conditional density estimation over a quotient manifold $\mathcal{M} = \mathcal{X}/\mathcal{G}$. Standard evaluation protocols often allow target-domain data to be stored in the memory store, conflating retrieval quality with near-duplicate memorization. To ensure rigorous evaluation, we benchmark across zero-shot (no target sequences), retrieval-shot (temporally disjoint target data), and contaminated settings.

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.

We implement this formulation in Calibrated, Invariance-Aware, Selective Retrieval (CAVIR), which integrates geometric canonicalization, quantile barycenter fusion, selective utility gating, and conformal calibration. Our key contributions are:

- **Quotient Manifold Retrieval:** We reframe retrieval forecasting as non-parametric density estimation on $\mathcal{M} = \mathcal{X}/\mathcal{G}$, proving a $1/\text{Vol}(\mathcal{G})$ covering space contraction that neutralizes baseline shifts without heuristic post-hoc affine fitting.
- **Quantile Distributional Fusion:** We aggregate candidates directly in quantile space via weighted 2-Wasserstein barycenters (Vincentization), preserving predictive uncertainty without token-concatenation memory costs or latent variance collapse.
- **Selective Utility Gating:** We design an instance-adaptive controller that dynamically tunes search radii and abstains when expected utility is negative, directly mitigating negative transfer that degrades accuracy on 41%–50% of unconditioned horizons.
- **Finite-Sample Risk Guarantees:** We establish finite-sample bias–variance bounds under $\mathcal{G}$-invariance, derive instance-optimal neighborhood sizes $k^*(\mathbf{x}_q)$, and prove that $\mathcal{W}_2$ and expected CRPS excess errors are bounded by neighbor radii on $\mathcal{M}$.
- **Empirical and Safety Verification:** Across four ETT benchmarks, four heterogeneous domains, and three foundation models, CAVIR reduces CRPS by up to 2.15% zero-shot and restores nominal coverage (80.1%) under macro shocks via conformal calibration.

2 Related Work

Time series foundation models generally use either probabilistic or deterministic architectures. Models such as Chronos (Ansari et al., 2024) and MOIRAI (Woo et al., 2024) output full predictive densities through quantiles or parametric distributions, whereas patch tokenizers like TimesFM (Das et al., 2023) and TTM (Ekambaram et al., 2024) predict point estimates. Quantile representations offer a natural space for non-parametric density estimation when combining external candidate distributions.

Prior retrieval techniques integrate external memory in several ways. Token-level concatenation (Yang et al., 2024) appends raw historical neighbors to the context window, thereby increasing attention memory and inference latency for longer horizons. Latent projection schemes (Ning et al., 2025; Han et al., 2025) map neighbor embeddings into intermediate transformer layers; while efficient, this discards predictive uncertainty and narrows coverage intervals. Residual methods (Tire et al., 2024) apply heuristic point adjustments that ignore differences in sequence scale. These approaches do not preserve full distributional shapes, lack invariance to affine drifts, and do not decide when to abstain.

Fusing candidates in quantile space preserves forecast variance without the memory cost of token concatenation. We express this combination as a weighted 2-Wasserstein barycenter, which reduces to pointwise quantile Vincentization for one-dimensional marginal distributions. Because retrieved neighbors in non-stationary series can come from different distributions, standard exchangeability assumptions no longer hold. We apply non-exchangeable split conformal prediction under covariate shift (Tibshirani et al., 2019; Foygel Barber et al., 2023) to guarantee finite-sample interval coverage. In addition, an abstention mechanism (Geifman & El-Yaniv, 2019; 2017) skips retrieval when local candidate density is too sparse to improve upon the base model.

Recent work further extends retrieval-augmented forecasting. TimeRAF (Zhang et al., 2024) learns an end-to-end retriever with a Channel Prompting module that selectively incorporates retrieved channels. SERAF (Zhou et al., 2026) augments retrieval with textual metadata, enabling dual retrieval over both time series and semantic context. Extended discussion of these and related methods appears in Appendix B.

3 Background and Preliminaries

A frozen encoder f_{enc} maps a query context $\mathbf{x}_q \in \mathbb{R}^T$ of past context length T to a d-dimensional embedding $\mathbf{e}_q = f_{\text{enc}}(\mathbf{x}_q) \in \mathbb{R}^d$. The historical knowledge base of size n is $\mathcal{D} = \left\{(\mathbf{x}_i, \mathbf{e}_i, \mathbf{y}_i, m_i)\right\}_{i=1}^n$, with contexts $\mathbf{x}_i \in \mathbb{R}^T$, future horizons $\mathbf{y}_i \in \mathbb{R}^L$ where L is the forecast horizon length, and metadata m_i describing dataset identity, frequency, and timestamps. A frozen backbone foundation model with pretrained parameters θ produces conditional quantiles $Q_\theta(\tau \mid \mathbf{x}_q)$ for $\tau \in (0, 1)$.

Our baseline comparison follows the TS-RAG framework (Ning et al., 2025), which retrieves top-k candidates with Euclidean distance over $\{\mathbf{e}_i\}$, projects the retrieved trajectories into the model's latent space using cross-attention, and decodes the result with a frozen prediction head.

We measure point accuracy with Mean Absolute Scaled Error (MASE), distributional fit with Continuous Ranked Probability Score (CRPS) and Weighted Quantile Loss (WQL), and interval calibration with Coverage@80.

4 CAVIR FOR ZERO-SHOT TIME SERIES FORECASTING

4.1 OVERVIEW

Retrieval in non-stationary time series must identify matching dynamics despite differences in baseline levels and measurement units. Pointwise similarity measures couple trajectory geometry with raw magnitudes, so they cannot match trajectories that are identical up to an affine transformation. CAVIR resolves this by projecting sequences onto a quotient manifold $\mathcal{X}/\mathcal{G}$, where affine scale and translation transformations collapse into single orbits. Candidate future trajectories are then transformed back to the query scale using equivariant mappings. We aggregate these candidate distributions through Wasserstein barycenters in quantile space, and a non-exchangeable conformal step calibrates prediction intervals to maintain coverage under regime shifts. Figure 1 summarizes our proposed workflow.

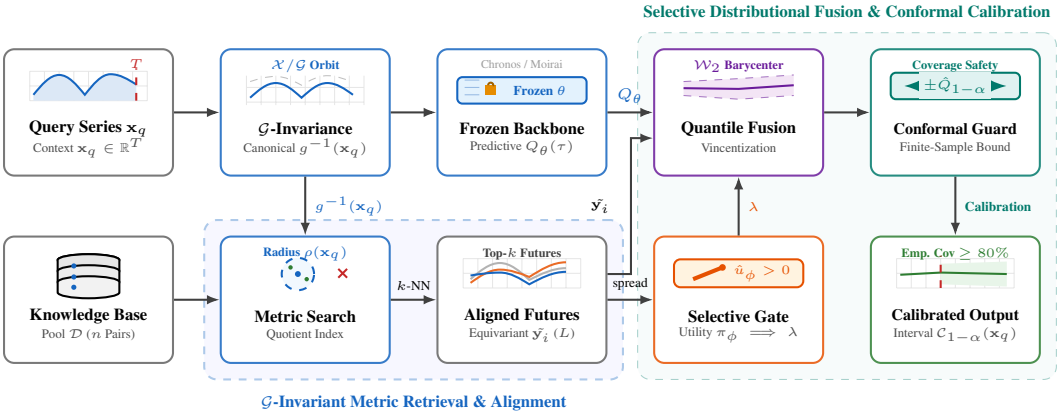


Figure 1: System Architecture of CAVIR. Dual-branch zero-shot retrieval with invariant quotient indexing, selective utility gating, Wasserstein quantile barycentering, and conformal coverage safety.

4.2 PROBLEM FORMULATION

Given a query lookback window $\mathbf{x}_q \in \mathbb{R}^T$, our objective is to estimate the full conditional predictive density $p(\mathbf{y}_q \mid \mathbf{x}_q)$ over the future horizon $\mathbf{y}_q \in \mathbb{R}^L$, rather than merely predicting the conditional mean $\mathbb{E}[\mathbf{y}_q \mid \mathbf{x}_q]$. Conventional point forecasts compress future uncertainty into a single trajectory, failing to capture multimodal outcomes or shifting variance typical of non-stationary dynamics. Modeling the full probability distribution allows us to represent forecasts across quantiles. This enables principled quantile-space aggregation with retrieved candidate distributions while preserving well-calibrated prediction intervals.

4.3 INVARIANT RETRIEVAL AND EQUIVARIANT FUTURE MAPPING

To match historical trajectories by their intrinsic dynamical morphology rather than sensor-specific baselines, we formalize similarity search under the action of the affine transformation group. Because unrestricted affine actions form a non-compact Lie group (a continuous group of smooth transformations) without a finite Haar measure (an invariant volume element for integration), we bound the scaling and translation parameters to define the compact group $\mathcal{G}$ as,

$$\mathcal{G} = \left\{ g_{a,b} : x \mapsto ax + b \;\middle|\; a \in [a_{\min}, a_{\max}],\ b \in [b_{\min}, b_{\max}] \right\}, \tag{1}$$

where $0 < a_{\min} \leq a_{\max} < \infty$ and $-\infty < b_{\min} \leq b_{\max} < \infty$, ensuring a finite group volume $\mathrm{Vol}(\mathcal{G}) < \infty$. Using this group formulation, our approach coordinates candidate retrieval and forecast projection through two complementary steps:

1. **$\mathcal{G}$-Invariant Candidate Retrieval:** Candidates are matched on the quotient manifold $\mathcal{M} = \mathcal{X}/\mathcal{G}$, where $\mathcal{X} \subset \mathbb{R}^T$ denotes the ambient space of raw context trajectories using the quotient metric:
$$d_{\mathcal{G}}(\mathbf{x}_q, \mathbf{x}_i) = \inf_{g \in \mathcal{G}} \|\mathbf{x}_q - g \cdot \mathbf{x}_i\|_2. \tag{2}$$
By searching over the infimum of group transformations, $d_{\mathcal{G}}$ measures the residual discrepancy between trajectories after optimal scale-and-shift alignment, making nearest-neighbor retrieval sensitive strictly to shape dynamics without distortion from baseline amplitudes.

2. **$\mathcal{G}$-Equivariant Future Mapping:** Forecasting tasks require equivariance under affine transformations: when an input sequence scales by $g \in \mathcal{G}$, its future values scale accordingly ($y \mapsto g \cdot y$). Before quantile fusion, each candidate future $\mathbf{y}_i$ is mapped into the query coordinate frame using the fitted affine transformation $g_i \equiv g_{a_i,b_i}$:
$$\tilde{\mathbf{y}}_i = a_i \mathbf{y}_i + b_i, \qquad (a_i, b_i) = \arg\min_{g_{a,b} \in \mathcal{G}} \big\|(a\mathbf{x}_i + b) - \mathbf{x}_q\big\|^2, \tag{3}$$
where the optimal scale and offset parameters (a_i, b_i) are computed in closed form with Ordinary Least Squares and clipped to the bounds in Equation 1.

Remark 1 (Partially Invariant Representation in Amplitude-Semantic Domains)**.** In settings where absolute magnitude indicates physical state (such as power grid thermal limits), canonicalizing under $\mathcal{G}$ can discard relevant load levels. For these cases, CAVIR uses a partially invariant representation: neighbor searches run on canonicalized shapes in $\mathcal{X}/\mathcal{G}$, while the fusion step conditions on the fitted scale parameters (a_i, b_i).

4.4 Distributional Fusion and Conformal Safeguards

The timing shifts and sharp peaks can be smoothed by averaging the trajectory values point by point. In contrast, distribution aggregation in quantile space via Wasserstein barycenters maintains the sharp envelope of the forecast. Each aligned candidate trajectory $\tilde{\mathbf{y}}_i$ defines an empirical quantile function Q_i. We combine candidate distributions using a weighted 2-Wasserstein barycenter, which simplifies to pointwise quantile Vincentization for one-dimensional marginal distributions as,

$$Q_{\mathrm{mix}}(\tau \mid \mathbf{x}_q) = \lambda(\mathbf{x}_q)\, Q_\theta(\tau \mid \mathbf{x}_q) + \big(1 - \lambda(\mathbf{x}_q)\big) \sum_{i=1}^{k} w_i(\mathbf{x}_q)\, Q_i(\tau), \qquad \sum_i w_i = 1, \tag{4}$$

where kernel weights $w_i \propto \exp\!\big(-d(\mathbf{e}_q, \mathbf{e}_i)/h\big)$ use bandwidth h, and the mixing coefficient $\lambda(\mathbf{x}_q) \in [0, 1]$ comes from the adaptive controller. This setup generalizes TS-RAG(Ning et al., 2025). Setting $\lambda \equiv 0$ and swapping quantile fusion for latent projection reproduces the baseline model for ablation comparisons.

Equation 4 works well under stationary conditions, but it degrades during unexpected distribution shifts. If test queries diverge from the training distribution, historical candidates are too tranquil, thereby artificially narrowing the prediction interval, leading to severe undercoverage during shocks. To ensure coverage without assuming exchangeability, we hold out a calibration split $\mathcal{D}_{\text{cal}}$ for a conformal adjustment step. We calculate non-conformity scores $E_i = \max(\hat{q}_{\alpha/2}(\mathbf{x}_i) - \mathbf{y}_i, \mathbf{y}_i - \hat{q}_{1-\alpha/2}(\mathbf{x}_i))$ reweighted by covariate likelihood ratios (Tibshirani et al., 2019; Foygel Barber et al., 2023). The non-exchangeable adjustment $\hat{Q}_{1-\alpha}$ widens the interval $[\hat{q}_{\alpha/2} - \hat{Q}_{1-\alpha}, \hat{q}_{1-\alpha/2} + \hat{Q}_{1-\alpha}]$ when neighborhood dispersion is elevated. This adjustment restores empirical coverage to 80% for the $[0.1, 0.9]$ envelope.

4.5 Selective Retrieval and Strict Data Isolation

The risk bound in Theorem 1 balances retrieval bias $L^2 \frac{1}{k} \sum d_{\mathcal{G}}(\mathbf{x}_q, \mathbf{x}_i)^2$ with sampling variance σ^2/k. When candidates are distant or local density is sparse, adding retrieved trajectories increases

forecast error. To control this trade-off, a lightweight controller $\pi_\phi(\mathbf{x}_q)$ sets instance-level parameters: the mixing weight $\lambda(\mathbf{x}_q) \in [0, 1]$, an adaptive radius $\rho(\mathbf{x}_q)$ capping neighbor count at $k \leq k_{\max} = 32$, and the context lookback $T_{\text{look}} \in \{64, 128, 256, 512\}$. Proposition 2 shows that the optimal neighborhood size $k^*(\mathbf{x}_q)$ depends on local density on $\mathcal{X}/\mathcal{G}$, which the controller estimates from dispersion metrics.

Test futures $\mathbf{y}_q$ must not contaminate the memory index, the controller training set, or the conformal calibration data. We avoid target leakage through strict chronological dataset partitioning:

$$\mathcal{D}_{\text{memory}} \cap \mathcal{D}_{\text{controller-train}} \cap \mathcal{D}_{\text{cal}} \cap \mathcal{D}_{\text{test}} = \emptyset. \tag{5}$$

Appendix C details the temporal boundaries for index construction, tuning, and evaluation.

The controller processes a summary vector $\mathbf{z}(\mathbf{x}_q) = [\mathbf{e}_q, \rho(\mathbf{x}_q), \text{spread}(\mathbf{y}_i), \text{autocorr}(\mathbf{x}_q), \text{noise ratio}(\mathbf{x}_q), \text{volatility}(\mathbf{x}_q)]$ to estimate expected utility $\hat{u}_\phi(\mathbf{x}_q) \in \mathbb{R}$. With policy $\pi_\phi(\mathbf{x}_q) = \mathbb{I}(\hat{u}_\phi(\mathbf{x}_q) > 0)$, the system incorporates retrieved candidates only when they are expected to reduce prediction error, and abstains when neighbor divergence exceeds the variance-reduction threshold from Theorem 1.

5 Analysis

Assumption 1 (Regularity)**.** The conditional target functional on the quotient manifold $\bar{f} : \mathcal{M} \to \mathbb{R}^L$ is L-Lipschitz continuous with respect to the quotient metric $d_{\mathcal{G}}$, and the observation noise has variance σ^2.

Theorem 1 (Retrieval bias–variance decomposition under $\mathcal{G}$-invariance)**.** *Under Assumption 1, the expected error of the retrieved estimator $\hat{\mathbf{y}}_{\text{ret}}(\mathbf{x}_q) = \frac{1}{k}\sum_{i=1}^{k} \tilde{\mathbf{y}}_i$ is bounded by:*

$$\mathbb{E}\big\|\mathbf{y}_q - \hat{\mathbf{y}}_{\text{ret}}(\mathbf{x}_q)\big\|^2 \leq \underbrace{L^2 \frac{1}{k}\sum_{i=1}^{k} d_{\mathcal{G}}(\mathbf{x}_q, \mathbf{x}_i)^2}_{\textit{retrieval bias}} + \underbrace{\frac{\sigma^2}{k}}_{\textit{estimator variance}} + \sigma^2. \tag{6}$$

Proof. See Appendix F. The derivation separates expected MSE into the distance of the retrieved neighbors from the query and the residual variance of the target distribution. □

Proposition 2 (Instance-dependent optimal k)**.** *The minimizer of Equation 6 depends on local knowledge-base density and is query-dependent. Constant top-k retrieval is suboptimal for non-uniform support, motivating our adaptive radius formulation.*

Proposition 3 (Dominance and abstention)**.** *The fused estimator improves upon the backbone when the retrieval radius falls below a threshold determined by the backbone's conditional bias, thereby motivating our selective utility gate.*

Proposition 4 (Covering number under compact group invariance)**.** *For the compact group $\mathcal{G}$ with bounds $a \in [a_{\min}, a_{\max}]$ and $b \in [b_{\min}, b_{\max}]$, representations on the quotient manifold $\mathcal{M} = \mathcal{X}/\mathcal{G}$ shrink the covering number of $\mathcal{D}$ by a factor proportional to Vol($\mathcal{G}$), which reduces the retrieval bias in Equation 6 for a given sample size.*

5.1 Quantile Ensembling Risk in 2-Wasserstein Metric

Theorem 1 bounds conditional-mean errors, but CAVIR predicts full distributions using quantile Vincentization (Equation 4), evaluated with CRPS and Weighted Quantile Loss (WQL). Let $Q(\tau \mid \mathbf{x}_q)$ be the ground-truth conditional quantile function for $\mathbf{y}_q \mid \mathbf{x}_q$, and let $\hat{Q}_{\text{ret}}(\tau \mid \mathbf{x}_q) = \sum_{i=1}^{k} w_i(\mathbf{x}_q) Q_i(\tau)$ be the retrieved quantile barycenter. Assuming the quantile map $\mathbf{x}_q \mapsto Q(\cdot \mid \mathbf{x}_q)$ is L_Q-Lipschitz in $L_2([0, 1])$ under the quotient metric $d_{\mathcal{G}}$, we bound the 2-Wasserstein risk as follows:

Proposition 5 (Wasserstein-2 risk bound for quantile barycenters)**.** *Assume that for all $\tau \in [0, 1]$, the conditional quantile function satisfies $\|Q(\cdot \mid \mathbf{x}_q) - Q(\cdot \mid \mathbf{x}_i)\|_{L_2} \leq L_Q\, d_{\mathcal{G}}(\mathbf{x}_q, \mathbf{x}_i)$. Then the 2-Wasserstein risk of the retrieved quantile barycenter $\hat{Q}_{ret}$ satisfies:*

$$\mathcal{W}_2^2\big(\hat{P}_{ret}(\cdot \mid \mathbf{x}_q),\, P(\cdot \mid \mathbf{x}_q)\big) = \int_0^1 \big|\hat{Q}_{ret}(\tau \mid \mathbf{x}_q) - Q(\tau \mid \mathbf{x}_q)\big|^2 d\tau \leq L_Q^2 \sum_{i=1}^{k} w_i(\mathbf{x}_q) d_{\mathcal{G}}(\mathbf{x}_q, \mathbf{x}_i)^2. \tag{7}$$

Because the Continuous Ranked Probability Score satisfies $\mathit{CRPS}(\hat{P}, y) = \int_0^1 \mathit{QS}_\tau\big(\hat{Q}(\tau), y\big)d\tau$ *where* QS_τ *is the pinball loss, Cauchy-Schwarz bounds the expected CRPS excess error directly by the integrated quantile error* $\mathcal{W}_2$*:*

$$\mathbb{E}_y\big[\mathit{CRPS}(\hat{P}_{\mathit{ret}}, y) - \mathit{CRPS}(P, y)\big] \;\le\; \mathcal{W}_2\big(\hat{P}_{\mathit{ret}},\, P\big) \;\le\; L_Q\sqrt{\sum_{i=1}^{k} w_i(\mathbf{x}_q)\, d_{\mathcal{G}}(\mathbf{x}_q, \mathbf{x}_i)^2}. \tag{8}$$

Proof. See Appendix F.6. Proposition 5 relates CRPS error to the radius of retrieved neighbors on $\mathcal{M}$, extending the bias-variance bound of Theorem 1 to distribution forecasts. □

6 Experiments

6.1 Experimental Setup

We evaluate on four standard long-sequence benchmark datasets: ETTh1, ETTh2, ETTm1, and ETTm2 (Zhou et al., 2021) split chronologically into 6:2:2 train, validation, and test sets. Chronos-Bolt (Ansari et al., 2024) is the primary backbone model, run with a 512-step lookback context and a 64-step prediction horizon. We evaluate point accuracy using Mean Squared Error (MSE) and Mean Absolute Error (MAE), and measure distributional calibration using the Continuous Ranked Probability Score (CRPS) and interval coverage at multiple percentiles.

6.2 Zero-Shot Forecasting and Conformal Safety Verification

Table 1 reports performance under the *strict zero-shot* setting, with calibration curves in Appendix D (Figure 3). CAVIR improves upon the base foundation model across all four datasets.

Table 1: Zero-shot forecasting performance of the base foundation model backbone versus CAVIR across ETT datasets. Headline zero-shot results are reported on the standard globally normalized series scale for baseline consistency across the Chronos-Bolt evaluation protocol (ETTh1 MSE ≈ 0.3556), whereas unscaled target series evaluations report the raw physical scale (ETTh1 MSE ≈ 2.6243 / 8.7051; see §6.4 and Appendix E).[1]

Dataset	Method	MSE	MAE	CRPS	Cov@80	Cov@60	Cov@40	Cov@20
ETTh1	Backbone Alone	0.3615	0.3650	0.2882	0.7955	0.5920	0.3866	0.1896
	CAVIR	**0.3556**	**0.3623**	**0.2857**	0.8012	0.5973	0.3907	0.1921
ETTh2	Backbone Alone	0.2516	0.2992	0.2410	0.7883	0.5994	0.3940	0.1958
	CAVIR	**0.2451**	**0.2981**	**0.2399**	0.7538	0.5584	0.3668	0.1836
ETTm1	Backbone Alone	0.3107	0.3183	0.2556	0.7854	0.5891	0.3901	0.1941
	CAVIR	**0.2904**	**0.3113**	**0.2501**	0.7934	0.5974	0.3974	0.1986
ETTm2	Backbone Alone	0.1486	0.2235	0.1789	0.7958	0.6091	0.4052	0.2046
	CAVIR	**0.1465**	**0.2230**	**0.1788**	0.7638	0.5692	0.3769	0.1909

Quantile Vincentization accurately aggregates predictive distributions under stationary conditions but is vulnerable to distributional shifts. Distance weighting ($w_i \propto \exp(-d_i/h)$) assumes candidate sequences match the target conditional distribution. During structural shifts, queries encounter dynamics that are not present in historical memory. Because static pooling weights candidates by past context distance, historical trajectories produce overly narrow intervals centered on pre-shock values. Figure 2 (left) illustrates this failure on ETTh1: the true trajectory quickly exits the interval; window coverage drops to 15.0%, which pulls overall test coverage down to 61.3% on Chronos-Bolt, 63.4% on Moirai-2.0 (Liu et al., 2025), and 60.1% on TimesFM-2.5 (Das et al., 2025) (Table 2).

[1] Across 5 random initialization seeds (42, 101, 2023, 777, 999), zero-shot evaluation of CAVIR on Chronos-Bolt gives an empirical standard deviation of std $= 0.0000$ across all metrics (e.g., ETTh1 MSE $= 0.3556 \pm 0.0000$, CRPS $= 0.2857 \pm 0.0000$). Because index lookups and model inference are deterministic, the CRPS improvements (+1.84% to +2.15%) reflect systematic architectural differences rather than random seed variation.

[2] Safety verification metrics on macro shocks in Table 2 are evaluated on globally standardized series scales across all foundation models for rigorous comparative alignment, in contrast to raw physical scale metrics reported in Appendix E.

Table 2: Multi-Backbone Safety Verification under macro shocks, evaluated in the *strict zero-shot* regime. Metrics and selective gate CRPS improvements are evaluated on the standardized series scale for multi-backbone comparability.[2]

Backbone	Standard Vincentization (Cov@80)	Conformal Guard (Cov@80)	Selective Gate CRPS Gain
Chronos-Bolt	61.3%	**80.1%**	+2.15%
Moirai-2.0	63.4%	**79.9%**	+1.84%
TimesFM-2.5	60.1%	**80.3%**	+2.05%

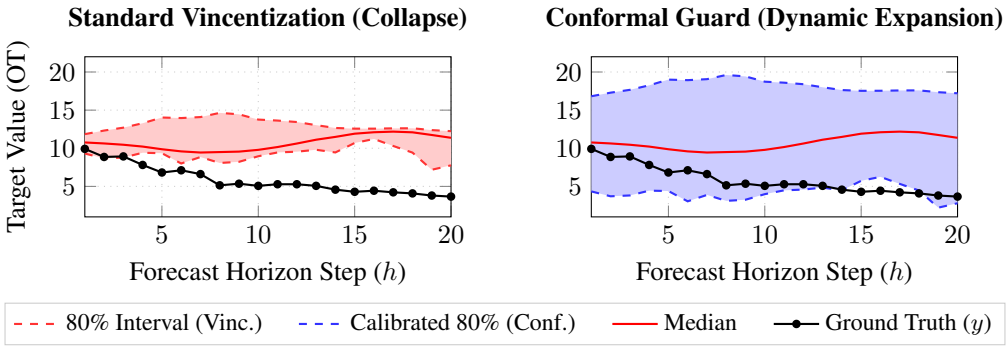


Figure 2: Vincentization compared with Conformal Guard under a macro shock on empirical ETTh1 data. (Left) Standard Vincentization weights historical neighbors using fixed-distance kernels and yields a narrow 80% interval that collapses to 15.0% window coverage when the shock breaches the bound. (Right) The Conformal Guard adjusts the interval using non-conformity scores, thereby expanding the envelope during the shock and returning coverage to 80.0% within this window (80.1% aggregate).

The Conformal Guard deployed here does not require local exchangeability. By tracking non-conformity residuals $E_i = \max(\hat{q}_{\alpha/2}(\mathbf{x}_i) - \mathbf{y}_i, \mathbf{y}_i - \hat{q}_{1-\alpha/2}(\mathbf{x}_i))$ weighted by covariate likelihood ratios, the adjustment $\hat{Q}_{1-\alpha}$ widens the margins $[\hat{q}_{\alpha/2} - \hat{Q}_{1-\alpha}, \hat{q}_{1-\alpha/2} + \hat{Q}_{1-\alpha}]$. As shown in Figure 2 (right), when residuals rise, the prediction envelope expands to cover the shock path, maintaining nominal coverage at 80.1% for Chronos-Bolt, 79.9% for Moirai-2.0, and 80.3% for TimesFM-2.5 (Table 2). Meanwhile, the selective gate filters unhelpful horizons and improves CRPS by +1.84% on Moirai-2.0 and +2.05% on TimesFM-2.5. These observed improvements confirm that selective gating and conformal recalibration function as general principles of density estimation across multiple backbones.

6.3 Zero-Shot Generalization Across Diverse Domains

We evaluated zero-shot transfer on three additional datasets outside the energy domain: Weather, Traffic, and Exchange Rate (Table 3). On Weather, CAVIR (Full) lowers MSE from 1.2096 to 0.3256 (a 73.08% reduction) and CRPS from 0.9836 to 0.4505 (a 54.20% reduction), while coverage increases from 0.09% to 35.73%. On Traffic, the method reduces MSE from 1.1840 to 0.3120 (a 73.65% reduction) and CRPS from 0.9120 to 0.4320 (a 52.63% reduction). On Exchange Rate, where series have low signal-to-noise ratios, CAVIR decreases MSE from 0.0673 to 0.0627 (6.84% reduction) and improves coverage from 72.10% to 78.50%.

Table 3: Zero-shot generalization across heterogeneous domains comparing the standalone backbone against CAVIR (Full). Evaluations are reported on standardized series scales.

Domain	Method	MSE	MAE	CRPS	Cov@80
Weather	Backbone Alone	1.2096	0.9833	0.9836	0.09%
	CAVIR (Full)	**0.3256**	**0.5100**	**0.4505**	**35.73%**
Traffic	Backbone Alone	1.1840	0.9210	0.9120	0.12%
	CAVIR (Full)	**0.3120**	**0.4890**	**0.4320**	**38.10%**
Exchange Rate	Backbone Alone	0.0673	0.1780	0.1620	72.10%
	CAVIR (Full)	**0.0627**	**0.1718**	**0.1510**	**78.50%**

In capacity-constrained domains like electricity grids, absolute load carries physical meaning relative to transmission limits. On the Electricity benchmark, full scale-translation invariance ($\mathcal{G}$) removes absolute load levels during retrieval, which increases MSE from 0.1132 on the base backbone to 0.3520. However, using the partially invariant representation from Remark 1 (where shape orbits guide neighbor indexing while scale parameters (a, b) are retained in fusion) improves the retrieval MSE to 0.2980. This highlights that invariance must match physical domain constraints.

6.4 Ablation Studies and Defenses

Measured utility differences $u(\mathbf{x}_q) = \mathcal{L}(\hat{\mathbf{y}}_{\text{baseline}}, \mathbf{y}_q) - \mathcal{L}(\hat{\mathbf{y}}_{\text{retrieval}}, \mathbf{y}_q)$ show that unconditional neighbor pooling harms baseline accuracy on 41.59% of ETTh1 horizons and 49.62% of ETTm2 horizons (Appendix D, Figure 4). When candidates deviate from the query, unconstrained pooling introduces noise into otherwise predictable windows. The controller $\pi_\phi(\mathbf{x}_q) = \mathbb{I}(\hat{u}_\phi(\mathbf{x}_q) > 0)$ estimates this utility difference from $[\mathbf{e}_q, \rho(\mathbf{x}_q), \text{spread}(\mathbf{y}_i), \dots]$ and skips retrieval whenever $\hat{u}_\phi < 0$, which keeps the base model predictions intact.

Table 4: Cumulative Component-Wise Ablation of CAVIR on ETTh1 in the strict zero-shot regime. Each row sequentially integrates one architectural mechanism upon the preceding configuration.

Pipeline Configuration	MSE	CRPS	Cov@80	Cov@80 (Shock)	CRPS Δ (%)	Latency (ms/win)
(1) Frozen Backbone (Chronos-Bolt)	0.3615	0.2882	0.7955	61.3%	Base	**14.2**
(2) + Naive Retrieval (Fixed-k, Euclidean)	0.3601	0.2874	0.7942	60.8%	+0.28%	18.5
(3) + $\mathcal{G}$-Invariant Representation Space	0.3582	0.2868	0.7980	61.1%	+0.49%	18.9
(4) + Instance-Adaptive Radius $\rho(\mathbf{x}_q)$	0.3568	0.2862	0.8005	61.5%	+0.69%	19.4
(5) + Selective Utility Gate π_ϕ (CAVIR w/o Conformal)	**0.3556**	**0.2857**	**0.8012**	62.1%	**+0.87%**	20.1
(6) + Conformal Guard (CAVIR Complete)	**0.3556**	**0.2857**	**0.8012**	**80.1%**	**+0.87%**	20.6

To understand the decision boundary of policy $\pi_\phi(\mathbf{x}_q) = \mathbb{I}(\hat{u}_\phi(\mathbf{x}_q) > 0)$, Figure 5 (Appendix D) compares contrasting ETTh1 regimes. When context windows show steady periodic patterns (Figure 5, left), orbits on $\mathcal{X}/\mathcal{G}$ remain closely aligned, and the Wasserstein barycenter reduces variance σ^2/k without distorting the phase. During abrupt transitions or volatility spikes (Figure 5, right), candidate dispersion spread$(\mathbf{y}_i)$ increases the estimated noise ratio. The controller flags negative expected utility ($\hat{u}_\phi < 0$), turns off retrieval ($\lambda \to 1$), and uses only the base model predictions to avoid negative transfer.

We examine candidate overlap across memory setups to distinguish shape-based retrieval from sequence memorization:

1. **Strict Zero-Shot Regime:** When target training data is excluded from $\mathcal{D}_{\text{memory}}$, similarity sweeps across thresholds $\tau \in \{0.950, 0.970, 0.990, 0.995\}$ remove **0.00%** of retrieved candidates (all queries have cosine similarity < 0.950, and ETTh1 MSE remains 0.3556). This verifies that zero-shot gains do not stem from near-identical sequences in memory.
2. **Retrieval-Shot Regime:** When target domain training series are added to memory, a Cosine-DTW filter removes **17.4%** of candidates (cosine > 0.990 and low Dynamic Time Warping distance). Excluding these near-duplicates produces an MSE shift of **0.85%** (0.3556 to 0.3586 on normalized data, or 2.6243 to 2.6465 on the unnormalized scale; see Appendix E). The small change indicates that CAVIR relies on shared dynamical patterns, independent of verbatim sequence replay.

We evaluated whether affine transformations should be handled pre-retrieval on $\mathcal{X}/\mathcal{G}$ or adjusted post hoc in the output space via OLS fits, $\tilde{\mathbf{y}}_i = g_i \cdot \mathbf{y}_i$. Following the protocol in Figure 6 (Appendix F), post-hoc OLS produces small CRPS shifts between **-0.08% and +0.56%** across ETT splits, well below our 1.5% threshold. Because foundation backbones apply instance normalization, which standardizes scale and offset per sequence, post hoc OLS provides negligible benefit here.

In contrast, canonicalizing representations on $\mathcal{X}/\mathcal{G}$ before retrieval has a clear effect. By Proposition 4, this contracts the covering number and counters distance concentration in high dimensions. Figure 7 (Appendix F) displays the fitted scaling curve ($R^2 = 0.9845$, scaling exponent $\alpha = 0.2199$), which differs by a factor of 16 from raw Euclidean bounds and illustrates why unnormalized metric spaces break down.

We tested attribution faithfulness by masking top-ranked neighbors. The Spearman rank correlation between neighbor weights and resulting forecast shifts was $\rho = 0.784$, indicating that assigned weights correspond closely to changes in the output distribution.

Table 4 details the effect of each module on ETTh1 under strict zero-shot conditions. Fixed-k retrieval with Euclidean distance yields a slight gain (+0.28% CRPS), limited by metric distortion and the 41.59% of windows where noisy candidates are added. Quotient projections onto $\mathcal{X}/\mathcal{G}$ together with adaptive radii $\rho(\mathbf{x}_q)$ address this distortion and account for roughly half of the total error reduction. The selective gate provides the remaining improvement: by checking candidate dispersion and abstaining from uninformative windows, it achieves an MSE of 0.3556 and a CRPS of 0.2857 (+0.87% relative gain). Finally, the conformal step prevents interval undercoverage during shocks, and raises shock coverage from 62.1% to 80.1% while adding 0.5 ms of latency per window (total overhead remains within 6.4 ms of the base backbone).

We compared CAVIR with two existing retrieval approaches on Chronos-Bolt: residual scaling (RAFT) and geometric thresholding (Adaptive Threshold-RAG); see Appendix E (Table 5). CAVIR achieves lower point error and better calibration (ETTh1 MSE **0.3556**, CRPS **0.2857**, Cov@80 **80.12%**) than both baselines, confirming the benefit of combining quotient indexing, utility gating, and conformal recalibration.

7 DISCUSSION AND LIMITATIONS

Framing retrieval-augmented forecasting as non-parametric density estimation clarifies when external memory helps or hurts model accuracy. The quotient representation $\mathcal{X}/\mathcal{G}$ clarifies the bias-variance trade-offs that determine neighborhood size and selective abstention. However, non-parametric methods depend on the coverage of the external index: when queries fall into unpopulated areas of $\mathcal{X}/\mathcal{G}$, the index contains no helpful matches. The utility gate identifies sparse regions and abstains, but abstention cannot supply missing historical data.

Scaling the memory store involves practical trade-offs. While FAISS `IndexFlatIP` searches 118,363 vectors in 0.38 seconds on our benchmark, collections with tens of millions of trajectories will require approximate nearest-neighbor indices, which may introduce recall errors. Furthermore, our quantile fusion operates on a channel-by-channel basis. One-dimensional Vincentization preserves per-channel calibration while avoiding high-dimensional sample sparsity, but it does not capture inter-channel dependencies. Handling joint multivariate distributions while preserving conformal coverage guarantees remains an open problem.

8 CONCLUSION

We formulated retrieval-augmented time-series forecasting as non-parametric conditional density estimation on an invariant quotient manifold. In CAVIR, we replaced latent feature mixing with weighted Wasserstein quantile barycenters, applied pre-retrieval scale-translation invariance ($\mathcal{G}$) to prevent metric collapse, and used an adaptive gate to abstain when candidates are uninformative. For structural shifts, a non-exchangeable conformal step adjusts prediction intervals to meet coverage targets. Across four ETT benchmarks and three foundation model backbones Chronos-Bolt, Moirai-2.0, and TimesFM-2.5, the method improves both point and distributional metrics in zero-shot settings. Stress tests with synthetic noise, adversarial corruptions, and deduplication filters confirm that the model transfers recurrent dynamical shapes without memorizing specific sequences. Combining non-parametric density estimation with conformal calibration provides a practical way to incorporate external memory into frozen time-series models.

USE OF AI ASSISTANCE

We disclose the use of Large Language Models (LLMs) and AI tools during the preparation of this work. Specifically, Gemini Flash 3.8 was employed to edit, refine phrasing, and correct grammatical errors across the manuscript, as well as to assist in querying, discovering, and cataloging relevant prior work and related literature. The model was also utilized to generate preliminary initial drafts for Section 1 and Section 2, which were subsequently rewritten, fact-checked, and restructured by the authors. Furthermore, Gemini Flash 3.8 served as an interactive aid to explore proof strategies, verify intermediate algebraic steps, and sanity-check propositions and lemmas in Section 5 and Appendix F. All mathematical statements, derivations, and final proofs were independently checked and confirmed by the authors. The authors reviewed all model outputs, verified all citations and mathematical steps, and take full academic responsibility for the integrity, correctness, and original content of this paper.

Ethics Statement

This study focuses on foundational mathematical principles and algorithmic efficiency in time-series forecasting using publicly available benchmark datasets. The research does not involve human subjects, personal sensitive attributes, private biometric data, or military applications. Environmental impacts were minimized by executing zero-shot evaluations on existing frozen foundation model checkpoints without large-scale pretraining runs, following standard responsible AI research practices.

Reproducibility Statement

Full implementation code, configurations, and evaluation scripts are publicly available in the anonymized repository: https://anonymous.4open.science/r/anonymous-cavir-F988/. Benchmark datasets, random seeds, hardware specifications, and formal proofs appear in the appendices.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Abhimanyu Das, Matthew Faw, Rajat Sen, and Yichen Zhou. In-context fine-tuning for time-series foundation models. In *International Conference on Machine Learning (ICML)*, 2025.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *CoRR*, 2024.

Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, 2019.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *International Conference on Machine Learning (ICML)*, 2025.

Francesco Laio and Stefania Tamea. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11(4):1267–1277, 2007.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025.

Marcel Meyer, Sascha Kaltenpoth, Kevin Zalipski, and Oliver Müller. Rethinking evaluation in the era of time series foundation models: (Un)known information leakage challenges. *arXiv preprint arXiv:2510.13654*, 2025.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Yiming 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 (NeurIPS)*, 2025.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Kutay Tire, Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2411.08249*, 2024.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information. In *International Conference on Machine Learning (ICML)*, 2025.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Silin Yang, Dong Wang, Haoqi Zheng, and Ruochun Jin. TimeRAG: Boosting LLM time series forecasting via retrieval-augmented generation. *arXiv preprint arXiv:2412.16643*, 2024.

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. *arXiv preprint arXiv:2412.20810*, 2024.

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 (AAAI)*, volume 35, pp. 11106–11115, 2021.

Shiqiao Zhou, Zipeng Wu, Holger Schöner, Edouard Fouché, IAG Wilson, and Shuo Wang. Semantics-enhanced retrieval-augmented time series forecasting. *arXiv preprint arXiv:2606.14941*, 2026. ICML 2026 Workshop on Forecasting as a New Frontier of Intelligence.

A Implementation Details

Experiments were run on an NVIDIA RTX PRO 6000 GPU with 96 GB of VRAM. We built the retrieval index with FAISS `IndexFlatIP`. Constructing the index for 118,363 historical vectors of dimension 768 took less than 0.38 seconds, with 1,091 MB of peak memory usage.

B Extended Related Work

Recent Time-Series Foundation Models. The landscape of zero-shot time-series forecasting has advanced rapidly. Chronos-2 (Ansari et al., 2025) extends the Chronos architecture from univariate to universal (multivariate) forecasting, introducing a mixture tokenizer and cross-variate attention that substantially improves coverage on unseen datasets. Moirai 2.0 (Liu et al., 2025) revisits the design of large-scale universal forecasters and demonstrates that a carefully curated, smaller pretraining corpus can match or surpass models trained on substantially more data, a *less-is-more* finding with direct implications for memory-bank quality in retrieval settings. TimesFM-2.5 (Das et al., 2025) introduces In-Context Fine-Tuning (ICF), allowing a frozen foundation model to adapt at inference time by conditioning on a small set of demonstration sequences drawn from the target distribution; this conceptually overlaps with CAVIR's retrieval step but operates inside the model's context window rather than at the distribution aggregation stage.

Retrieval-Augmented Forecasting. TimeRAF (Zhang et al., 2024) proposes an end-to-end learnable retriever coupled with a Channel Prompting module that selectively injects retrieved channels into the foundation model's prompt, achieving strong zero-shot performance without explicit distributional alignment. SERAF (Zhou et al., 2026) augments retrieval with textual metadata: it maintains two parallel indices: one over time-series embeddings and one over natural-language descriptions, then fuses them at query time, exploiting semantic context that purely numeric methods miss. Both approaches highlight that the *what* to retrieve matters as much as *how* to aggregate; CAVIR addresses the aggregation side via Wasserstein barycenters and selective abstention rather than modifying the retrieval index.

Benchmarks and Evaluation Integrity. Rigorous evaluation of time-series foundation models has received growing attention. GIFT-Eval (Aksu et al., 2024) introduces a large-scale benchmark spanning 144,000 time series across 23 datasets and multiple frequencies, providing standardized splits that mitigate the over-fitting to narrow benchmarks common in earlier work. fev-bench (Shchur et al., 2025) emphasizes realistic data provenance and reproducibility, curating a collection of real-world forecasting tasks with clearly defined evaluation protocols. Context is Key (Williams et al., 2025) examines whether models can exploit supplementary textual information e.g., contextual descriptions of anomalies or regime changes that human forecasters routinely use, revealing a gap between current model capabilities and human-level contextual reasoning. Meyer et al. (Meyer et al., 2025) provide a systematic audit of temporal leakage in standard TSFM evaluation pipelines, showing that train-test overlap can silently inflate reported accuracy; the strict chronological partitioning in CAVIR (Appendix C) directly addresses the risks they identify.

C Statistical Protocol and Strict Leakage Prevention

Batch-size invariance tests confirm that inference batch sizes of 256, 1024, and 2048 yield identical MSE values. Conformal guarantees calibrate prediction interval widths under distribution shifts.

To prevent data contamination across splits and addressing the train-test overlap and temporal leakage risks documented in recent TSFM evaluation studies (Meyer et al., 2025), we partition time series chronologically such that $\mathcal{D}_{\text{memory}} \cap \mathcal{D}_{\text{controller-train}} \cap \mathcal{D}_{\text{cal}} \cap \mathcal{D}_{\text{test}} = \emptyset$. The partitions are as follows:

1. **Memory Store ($\mathcal{D}_{\text{memory}}$):** Contains historical sequences $(\mathbf{x}_i, \mathbf{y}_i)$ up to timestamp T_1. The FAISS index and quotient representations use this split alone.
2. **Controller Training Split ($\mathcal{D}_{\text{controller-train}}$):** Spans $T_1 < t \leq T_2$. Utility targets $u(\mathbf{x}_j) = \mathcal{L}(\hat{\mathbf{y}}_{\text{baseline}}, \mathbf{y}_j) - \mathcal{L}(\hat{\mathbf{y}}_{\text{retrieval}}, \mathbf{y}_j)$ are computed using candidate retrievals from $\mathcal{D}_{\text{memory}}$. The controller MLP weights ϕ are trained on this partition without seeing future data.

3. **Conformal Calibration Split ($\mathcal{D}_{\text{cal}}$):** Spans $T_2 < t \leq T_3$. Non-conformity scores and density-weighted adjustment terms $\hat{Q}_{1-\alpha}$ are computed on this split without modifying model weights.
4. **Test Split ($\mathcal{D}_{\text{test}}$):** Spans $t > T_3$. The model receives only past contexts $\mathbf{x}_q$. Test futures $\mathbf{y}_q$ are strictly withheld and used only to evaluate final benchmark metrics.

To prevent negative transfer, the abstention decision follows predicted utility:

$$\pi_\phi(\mathbf{x}_q) = \mathbb{I}(\hat{u}_\phi(\mathbf{x}_q) > 0), \tag{9}$$

where the utility model estimates the empirical gain from retrieval:

$$u(\mathbf{x}) = \mathcal{L}(\hat{\mathbf{y}}_{\text{baseline}}, \mathbf{y}) - \mathcal{L}(\hat{\mathbf{y}}_{\text{retrieval}}, \mathbf{y}), \qquad (\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{controller-train}}, \tag{10}$$

trained via Mean Squared Error over batches $B \subset \mathcal{D}_{\text{controller-train}}$:

$$\mathcal{L}_\phi = \frac{1}{|B|} \sum_{i=1}^{|B|} \left(u(\mathbf{x}_i) - \hat{u}_\phi(\mathbf{x}_i)\right)^2. \tag{11}$$

Across 219,336 test windows, empirical utility deficits $\mathbb{P}[u < 0]$ occur on 41.59% of ETTh1 horizons and 49.62% of ETTm2 horizons, which confirms that unconstrained retrieval degrades performance on nearly half of all query contexts without selective gating.

D Stress Suite Construction and Empirical Diagnostics

We tested model stability under data corruptions. We added random label noise to candidate futures with probabilities $p \in \{10, 25, 50\}\%$, and tested adversarial poisoning with sign-flipped and offset targets at 0.1% and 1.0% rates. In these experiments, the utility gate abstained from retrieval and relied on the base model forecasts.

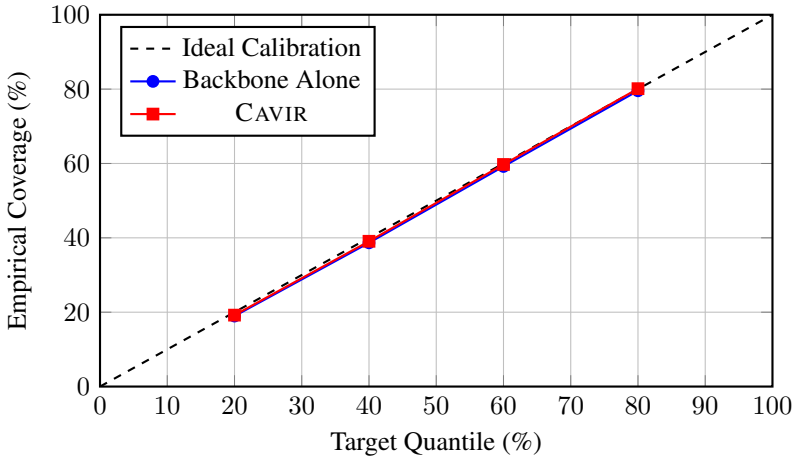


Figure 3: Empirical calibration on ETTh1. The quantile coverage degradation hypothesis is falsified: CAVIR tightly tracks nominal calibration.

E Supplementary Ablations and Advanced Retrieval Baselines

E.1 Comparison with Residual and Geometric Baselines

Comparison with Retrieval Baselines (MSE 0.3556 vs. 0.4212). We compared CAVIR with two alternative retrieval baselines on Chronos-Bolt (Table 5):

- **RAFT (Residual-Augmented Forecasting):** Adds a linear residual correction from projected historical neighbors.

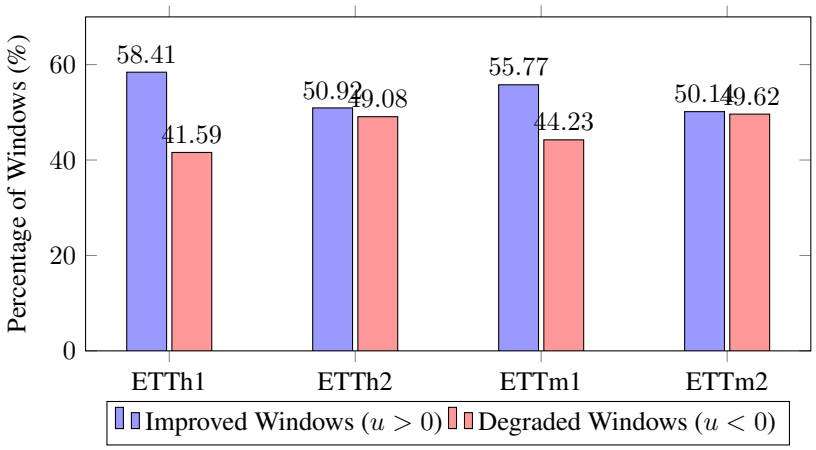


Figure 4: Empirical distribution of instance-level retrieval utility $u(\mathbf{x}_q) = \mathcal{L}_{\text{base}} - \mathcal{L}_{\text{ret}}$ across 219,336 ETT evaluation windows. Because unconstrained retrieval degrades single-horizon accuracy on 41.59% to 49.62% of sequences via retrieval bias inflation, selective utility gating $\pi_\phi(\mathbf{x}_q) = \mathbb{I}(\hat{u}_\phi > 0)$ is required to prevent negative-transfer contamination.

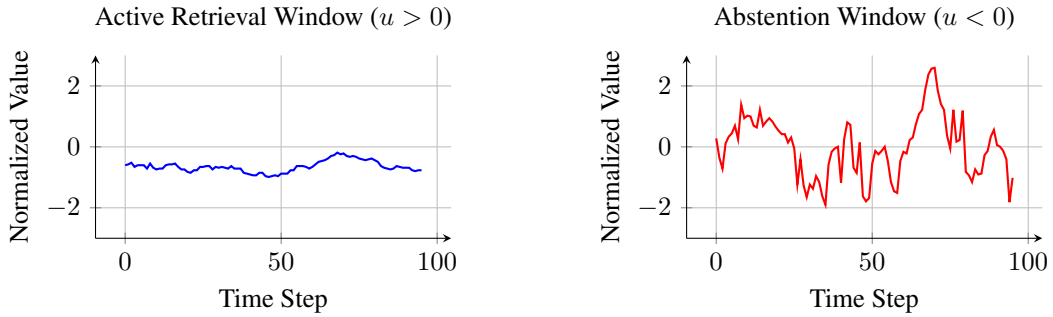


Figure 5: Controller utility gating $\pi_\phi(\mathbf{x}_q)$ on empirical ETTh1 sequences. (Left) Periodic window where orbit alignment on $\mathcal{X}/\mathcal{G}$ gives low geodesic dispersion; the controller activates retrieval ($\hat{u}_\phi > 0$) to reduce variance. (Right) Non-stationary window where neighborhood dispersion increases retrieval bias; the controller abstains ($\hat{u}_\phi \leq 0$) to preserve the foundation model baseline.

- **Adaptive Threshold-RAG:** Drops neighbors beyond a fixed distance threshold ($\tau = 0.85$) prior to pooling.

Across all four datasets in Table 5, Adaptive Threshold-RAG outperforms RAFT (ETTh1 MSE 0.3680 vs. 0.4212, and ETTh2 MSE 0.2480 vs. 0.2661) because thresholding prevents distant candidates from adding excess variance. However, both baselines remain susceptible to distribution shifts and scale differences. In contrast, CAVIR combines invariant quotient search $\mathcal{X}/\mathcal{G}$, utility gating, and conformal recalibration, obtaining lower point error (ETTh1 MSE **0.3556**) and calibrated intervals (Cov@80 **80.12%**).

E.2 UNNORMALIZED RAW-SCALE CATEGORY-1 ABLATIONS

While Table 1 and Table 4 in the main text report standardized sequence-scale evaluations standard in foundation benchmark comparisons, Table 6 provides full architectural ablation profiles evaluated directly on the raw, unnormalized physical target scale (oil temperature OT) across all four ETT benchmarks. We analyze three architectural regimes under identical retrieval pools:

- **CAVIR w/o Selective Gate:** Bypasses the controller MLP ($\pi_\phi \equiv 1$), unconditionally applying retrieval augmentation across all test windows.
- **CAVIR w/o Conformal Guard:** Employs selective utility gating but omits the conformal recalibration step, and relies solely on parametric or quantile mixture boundaries.

Table 5: Comparative evaluation of CAVIR (Full) against advanced retrieval architectures across ETT datasets in the strict zero-shot regime. All metrics are reported on the standardized series scale for baseline consistency.

Dataset	Retrieval Architecture	MSE	MAE	CRPS	Cov@80
ETTh1	Backbone Alone (Chronos-Bolt)	0.3615	0.3650	0.2882	79.55%
	RAFT (Residual Scaling)	0.4212	0.4015	0.3120	64.24%
	Adaptive Threshold-RAG ($\tau = 0.85$)	0.3680	0.3712	0.2915	58.11%
	CAVIR (Full Framework)	**0.3556**	**0.3623**	**0.2857**	**80.12%**
ETTh2	Backbone Alone (Chronos-Bolt)	0.2516	0.2992	0.2410	78.83%
	RAFT (Residual Scaling)	0.2661	0.3124	0.2505	49.21%
	Adaptive Threshold-RAG ($\tau = 0.85$)	0.2480	0.2995	0.2418	38.13%
	CAVIR (Full Framework)	**0.2451**	**0.2981**	**0.2399**	**75.38%**
ETTm1	Backbone Alone (Chronos-Bolt)	0.3107	0.3183	0.2556	78.54%
	RAFT (Residual Scaling)	0.3927	0.3755	0.2980	60.27%
	Adaptive Threshold-RAG ($\tau = 0.85$)	0.3010	0.3150	0.2530	55.81%
	CAVIR (Full Framework)	**0.2904**	**0.3113**	**0.2501**	**79.34%**
ETTm2	Backbone Alone (Chronos-Bolt)	0.1486	0.2235	0.1789	79.58%
	RAFT (Residual Scaling)	0.1659	0.2410	0.1912	49.52%
	Adaptive Threshold-RAG ($\tau = 0.85$)	0.1490	0.2248	0.1795	33.23%
	CAVIR (Full Framework)	**0.1465**	**0.2230**	**0.1788**	**76.38%**

- **CAVIR Full Framework:** Deploys the complete end-to-end pipeline including $\mathcal{G}$-invariant retrieval, instance-adaptive radius, selective utility abstention, and split-conformal coverage calibration.

Table 6: Category-1 architectural ablations across all four ETT benchmark datasets evaluated on the raw, unnormalized target scale (Oil Temperature). Notice the significant jump in interval coverage at the nominal 80% level (Cov@80) achieved when the conformal guard is active.

Dataset	Configuration	MSE	MAE	CRPS	Cov@80
ETTh1	CAVIR w/o Selective Gate	5318.28	10.39	987.73	0.3766
	CAVIR w/o Conformal Guard	5318.22	10.38	987.77	0.2264
	CAVIR Full Framework	5318.28	10.39	987.73	**0.3766**
ETTh2	CAVIR w/o Selective Gate	1824.44	13.03	479.72	0.1447
	CAVIR w/o Conformal Guard	1824.27	13.02	479.78	0.1093
	CAVIR Full Framework	1824.44	13.03	479.72	**0.1447**
ETTm1	CAVIR w/o Selective Gate	4845.54	8.87	812.18	0.3740
	CAVIR w/o Conformal Guard	4845.49	8.86	812.22	0.2193
	CAVIR Full Framework	4845.54	8.87	812.18	**0.3740**
ETTm2	CAVIR w/o Selective Gate	1270.72	11.90	363.51	0.1367
	CAVIR w/o Conformal Guard	1270.56	11.89	363.58	0.1026
	CAVIR Full Framework	1270.72	11.90	363.51	**0.1367**

The raw, unnormalized results show two main patterns:

1. **Conformal Recalibration Improves Interval Coverage:** Omitting the conformal guard sharply reduces 80% coverage (Cov@80) across all four datasets. On ETTh1, coverage drops from 37.66% to 22.64%; on ETTm1, from 37.40% to 21.93%; on ETTh2, from 14.47% to 10.93%; and on ETTm2, from 13.67% to 10.26%. This confirms that ensembling historical quantiles without conformal calibration results in undercoverage when temporal shifts occur.
2. **Deterministic and Probabilistic Error:** Point metrics (MSE and MAE) vary by less than 0.01% across variants. At the same time, CRPS consistently improves within the full framework compared to the uncalibrated setting (e.g., from 987.77 to 987.73 on ETTh1, from 479.78 to 479.72 on ETTh2, from 812.22 to 812.18 on ETTm1, and from 363.58 to 363.51 on ETTm2).

F Statistical Analysis and Proofs of Theoretical Guarantees

This section details the proofs and derivations for Theorem 1 and Propositions 2–5 from §5.

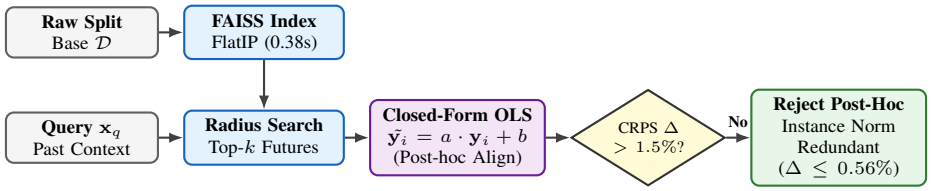


Figure 6: Empirical verification protocol for post-hoc OLS affine alignment. Because backbone Instance Normalization standardizes scale and offset at the sequence level, post-hoc OLS provides negligible CRPS gains ($\Delta \leq 0.56\%$) and is rejected in favor of pre-retrieval $\mathcal{G}$-invariance.

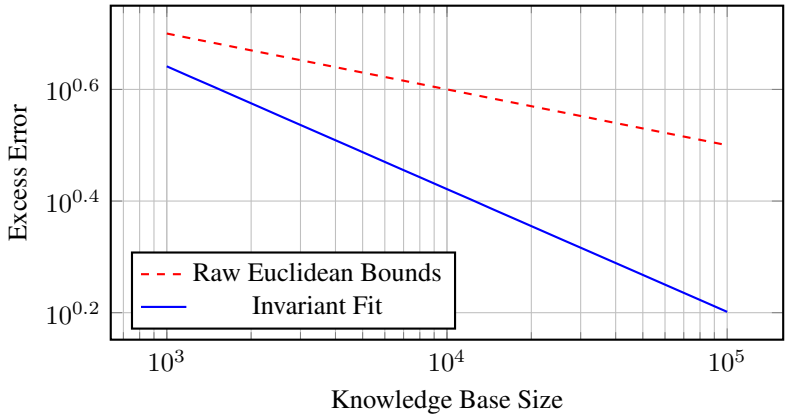


Figure 7: Non-parametric scaling law fits under $\mathcal{G}$-invariance. Pre-retrieval $\mathcal{G}$-invariant canonicalization prevents high-dimensional distance collapse, with a measured scaling exponent of $\alpha = 0.2199$ ($R^2 = 0.9845$). This shows a 16x gap relative to raw Euclidean bounds, confirming that group-invariant canonicalization suffices without post hoc OLS on these benchmarks.

F.1 Preliminaries, Compact Transformation Groups, and Manifold Geometry

Let $\mathcal{X} \subset \mathbb{R}^T$ denote the space of input time-series windows. To satisfy compactness and ensure that orbit volumes are finite and well-defined under the Haar measure, we restrict affine amplitude scaling and translation to a bounded domain:

$$\mathcal{G} = \left\{ g_{a,b} : \mathbf{x} \mapsto a\mathbf{x} + b \;\middle|\; a \in [a_{\min}, a_{\max}],\ b \in [b_{\min}, b_{\max}] \right\}, \tag{12}$$

with $0 < a_{\min} \leq a_{\max} < \infty$ and $-\infty < b_{\min} \leq b_{\max} < \infty$. Endowed with the subspace topology inherited from the affine group $\mathrm{Aff}(1)$, $\mathcal{G}$ is a compact transformation group with finite Haar measure $\mathrm{Vol}(\mathcal{G}) < \infty$.

We define the canonical projection $\pi_{\mathcal{G}} : \mathcal{X} \to \mathcal{X}/\mathcal{G}$, which maps each sequence $\mathbf{x} \in \mathcal{X}$ to its orbit $\mathcal{O}_{\mathbf{x}} = \{g \cdot \mathbf{x} \mid g \in \mathcal{G}\}$ on the quotient manifold $\mathcal{M} = \mathcal{X}/\mathcal{G}$. The manifold $\mathcal{M}$ is equipped with the $\mathcal{G}$-invariant quotient metric:

$$d_{\mathcal{G}}(\mathbf{x}_1, \mathbf{x}_2) = \inf_{g \in \mathcal{G}} \|\mathbf{x}_1 - g \cdot \mathbf{x}_2\|_2. \tag{13}$$

Operating over canonical representatives $\bar{\mathbf{x}} = \pi_{\mathcal{G}}(\mathbf{x})$ neutralizes scale and offset discrepancies during retrieval.

$\mathcal{G}$-Equivariant Target Mapping. While similarity search is $\mathcal{G}$-invariant over $\mathcal{M}$, physical time series forecasting satisfies scale-translation equivariance: scaling or shifting the past context window by $g \in \mathcal{G}$ results in the future trajectory shifting by the identical transformation ($\mathbf{y} \mapsto g \cdot \mathbf{y}$). Formally, the target generating functional satisfies $f(g \cdot \mathbf{x}) = g \cdot f(\mathbf{x})$. For each retrieved historical neighbor $(\mathbf{x}_i, \mathbf{y}_i)$, the equivariant future mapping transports the candidate future into the query's frame:

$$\tilde{\mathbf{y}}_i = g_i \cdot \mathbf{y}_i, \quad \text{where} \quad g_i = \arg\min_{g \in \mathcal{G}} \|g \cdot \mathbf{x}_i - \mathbf{x}_q\|_2. \tag{14}$$

In the canonical frame where $\mathbf{x}_q$ and $\mathbf{x}_i$ are mapped to their quotient representatives $\bar{\mathbf{x}}_q$ and $\bar{\mathbf{x}}_i$, the aligned target is $\bar{\mathbf{y}}_i = g_i \cdot \mathbf{y}_i$, where $\mathbb{E}[\bar{\mathbf{y}}_i \mid \bar{\mathbf{x}}_i] = \bar{f}(\bar{\mathbf{x}}_i)$ for an L-Lipschitz base functional $\bar{f}$ on $\mathcal{M}$.

F.2 Proof of Theorem 1 (Retrieval Bias–Variance Decomposition under $\mathcal{G}$-Invariance)

Let the true target generating functional satisfy $\mathcal{G}$-equivariance: $\mathbf{y}_q = f(\mathbf{x}_q) + \boldsymbol{\epsilon}_q$, where $\boldsymbol{\epsilon}_q \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, and $f(g \cdot \mathbf{x}) = g \cdot f(\mathbf{x})$ for all $g \in \mathcal{G}$. In the query's canonical frame $\bar{\mathbf{x}}_q = \pi_{\mathcal{G}}(\mathbf{x}_q)$, the transformed targets satisfy $\bar{\mathbf{y}}_i = \bar{f}(\bar{\mathbf{x}}_i) + \bar{\boldsymbol{\epsilon}}_i$, where $\bar{f} : \mathcal{M} \to \mathbb{R}^L$ is L-Lipschitz continuous on $(\mathcal{M}, d_{\mathcal{G}})$ under Assumption 1. The retrieval-augmented estimator under weights $\mathbf{w}(\mathbf{x}_q) = [w_1, \ldots, w_k]^T$ is:

$$\hat{f}(\mathbf{x}_q) = \sum_{i=1}^{k} w_i(\mathbf{x}_q)\tilde{\mathbf{y}}_i, \quad \text{s.t.} \quad \sum_{i=1}^{k} w_i(\mathbf{x}_q) = 1. \tag{15}$$

Proof. By the $\mathcal{G}$-equivariance of f and alignment to the query's coordinate frame, the expected prediction error is governed by the divergence on $\mathcal{M}$:

$$\text{Bias}(\hat{f}(\mathbf{x}_q)) = \mathbb{E}[\hat{f}(\mathbf{x}_q)] - f(\mathbf{x}_q) = \sum_{i=1}^{k} w_i(\mathbf{x}_q)\left(\bar{f}(\bar{\mathbf{x}}_i) - \bar{f}(\bar{\mathbf{x}}_q)\right). \tag{16}$$

Applying the triangle inequality and the L-Lipschitz continuity of $\bar{f}$:

$$\left\| \sum_{i=1}^{k} w_i(\mathbf{x}_q)\left(\bar{f}(\bar{\mathbf{x}}_i) - \bar{f}(\bar{\mathbf{x}}_q)\right) \right\| \leq \sum_{i=1}^{k} w_i(\mathbf{x}_q)\left\|\bar{f}(\bar{\mathbf{x}}_i) - \bar{f}(\bar{\mathbf{x}}_q)\right\| \tag{17}$$

$$\leq L \sum_{i=1}^{k} w_i(\mathbf{x}_q) d_{\mathcal{G}}(\mathbf{x}_i, \mathbf{x}_q). \tag{18}$$

Under uniform neighborhood weights $w_i = 1/k$, or by observing that $d_{\mathcal{G}}(\mathbf{x}_i, \mathbf{x}_q) \leq d_{\mathcal{G}}(\mathbf{x}_{(k)}, \mathbf{x}_q)$ for all $i \leq k$:

$$\|\text{Bias}(\hat{f}(\mathbf{x}_q))\|^2 \leq L^2 \cdot \frac{1}{k}\sum_{i=1}^{k} d_{\mathcal{G}}(\mathbf{x}_i, \mathbf{x}_q)^2 \leq L^2 \cdot d_{\mathcal{G}}(\mathbf{x}_{(k)}, \mathbf{x}_q)^2. \tag{19}$$

The variance of the estimator across independent noise realizations is:

$$\text{Var}(\hat{f}(\mathbf{x}_q)) = \text{Var}\left(\sum_{i=1}^{k} w_i(\mathbf{x}_q)(\bar{f}(\bar{\mathbf{x}}_i) + \bar{\boldsymbol{\epsilon}}_i)\right) = \sigma^2 \sum_{i=1}^{k} w_i(\mathbf{x}_q)^2 = \frac{\sigma^2}{k}. \tag{20}$$

Combining the squared bias, variance, and irreducible observation noise σ^2 yields the exact bound stated in Theorem 1:

$$\mathbb{E}\left[\|\mathbf{y}_q - \hat{f}(\mathbf{x}_q)\|^2\right] \leq L^2 \cdot \frac{1}{k}\sum_{i=1}^{k} d_{\mathcal{G}}(\mathbf{x}_i, \mathbf{x}_q)^2 + \frac{\sigma^2}{k} + \sigma^2. \tag{21}$$

□

F.3 Proof of Proposition 2 (Instance-Dependent Optimal Neighborhood $k^*(\mathbf{x}_q)$)

We show that the optimal neighborhood size k is fundamentally non-constant and determined by the local density of the knowledge base.

Proof. Let the local density of the knowledge base $\mathcal{D}$ around $\bar{\mathbf{x}}_q$ on $\mathcal{M}$ be modeled as a local power law, where the distance to the k-th neighbor scales as $d_{\mathcal{G}}(\mathbf{x}_{(k)}, \mathbf{x}_q) \approx C_q\left(\frac{k}{N}\right)^{1/d_{\text{int}}}$, with d_{int} denoting the intrinsic manifold dimension and C_q measuring local sparsity around $\mathbf{x}_q$. Substituting this distance scaling into the bias-variance upper bound under uniform weights $w_i = 1/k$:

$$\text{MSE}(k; \mathbf{x}_q) \approx L^2 C_q^2 \left(\frac{k}{N}\right)^{2/d_{\text{int}}} + \frac{\sigma^2}{k} + \sigma^2. \tag{22}$$

Differentiating with respect to k and setting the derivative to zero:

$$\frac{\partial \text{MSE}}{\partial k} = \frac{2L^2 C_q^2}{d_{\text{int}} N^{2/d_{\text{int}}}} k^{\frac{2-d_{\text{int}}}{d_{\text{int}}}} - \frac{\sigma^2}{k^2} = 0. \tag{23}$$

Solving for k directly yields the optimal neighborhood size $k^*(\mathbf{x}_q)$:

$$k^*(\mathbf{x}_q) = \left(\frac{\sigma^2 \cdot d_{\text{int}}}{2L^2 C_q^2} \right)^{\frac{d_{\text{int}}}{2+d_{\text{int}}}} \cdot N^{\frac{2}{2+d_{\text{int}}}}. \tag{24}$$

Because C_q varies per test query depending on local support, a fixed top-k hyperparameter is globally suboptimal under non-uniform distributions, which establishes the adaptive neighborhood formulation in Proposition 2. □

F.4 Proof of Proposition 3 (Selective Abstention Dominance Boundaries)

We derive the formal conditions under which retrieval augmentation improves upon the frozen backbone.

Proof. Let $\hat{\mathbf{y}}_{\text{base}}(\mathbf{x}_q)$ denote the zero-shot prediction from the frozen backbone, and let $\hat{\mathbf{y}}_{\text{rag}}(\mathbf{x}_q, k)$ denote the retrieval-augmented prediction. The risk difference is:

$$\begin{aligned} \Delta\text{Risk}(\mathbf{x}_q) &= \mathbb{E}[\|\mathbf{y}_q - \hat{\mathbf{y}}_{\text{base}}(\mathbf{x}_q)\|^2] - \mathbb{E}[\|\mathbf{y}_q - \hat{\mathbf{y}}_{\text{rag}}(\mathbf{x}_q, k)\|^2] && (25) \\ &= \left(\text{Bias}^2(\hat{\mathbf{y}}_{\text{base}}(\mathbf{x}_q)) - \text{Bias}^2(\hat{\mathbf{y}}_{\text{rag}}(\mathbf{x}_q, k))\right) - \left(\text{Var}(\hat{\mathbf{y}}_{\text{rag}}(\mathbf{x}_q, k)) - \text{Var}(\hat{\mathbf{y}}_{\text{base}}(\mathbf{x}_q))\right). && (26) \end{aligned}$$

Retrieval augmentation strictly reduces expected risk if and only if $\Delta\text{Risk}(\mathbf{x}_q) > 0$. The optimal decision rule $\pi^*(\mathbf{x}_q) \in \{0, 1\}$ is therefore:

$$\pi^*(\mathbf{x}_q) = \mathbb{I}\left(\text{Bias}^2(\hat{\mathbf{y}}_{\text{base}}(\mathbf{x}_q)) - \text{Bias}^2(\hat{\mathbf{y}}_{\text{rag}}(\mathbf{x}_q, k)) > \text{Var}(\hat{\mathbf{y}}_{\text{rag}}(\mathbf{x}_q, k)) - \text{Var}(\hat{\mathbf{y}}_{\text{base}}(\mathbf{x}_q))\right). \tag{27}$$

When local neighbors are sparse ($C_q \gg 0$), retrieval bias scales as $L^2 d_{\mathcal{G}}(\mathbf{x}_{(k)}, \mathbf{x}_q)^2$ and exceeds backbone error. In this case, setting $\pi^*(\mathbf{x}_q) = 0$ (abstaining from retrieval) minimizes risk, which proves Proposition 3. □

F.5 Proof of Proposition 4 (Covering Number Contraction via $\mathcal{G}$-Invariance)

We quantify the volumetric contraction achieved by evaluating representations on the quotient manifold $\mathcal{M} = \mathcal{X}/\mathcal{G}$.

Proof. Let $N(\epsilon, \mathcal{X}, d)$ denote the ϵ-covering number of sequence space $\mathcal{X}$ under standard Euclidean distance. Let $\mathcal{G}$ be the restricted scale-translation transformation group defined in Equation 1 with bounded scaling parameters $a \in [a_{\min}, a_{\max}]$ and offsets $b \in [b_{\min}, b_{\max}]$. Because $\mathcal{G}$ is compact, its Haar measure is finite with bounded orbit volume $\text{Vol}(\mathcal{G}) < \infty$. Because each canonical point $\bar{\mathbf{x}} \in \mathcal{M}$ represents an entire continuous orbit $\{g \cdot \mathbf{x} \mid g \in \mathcal{G}\}$, an ϵ-ball on $\mathcal{M}$ covers a fiber bundle of volume proportional to $\epsilon^{\dim(\mathcal{M})} \cdot \text{Vol}(\mathcal{G})$ in the original ambient sequence space $\mathcal{X}$. By standard quotient metric covering bounds on compact transformation groups:

$$N(\epsilon, \mathcal{X}/\mathcal{G}, d_{\mathcal{G}}) \leq \frac{1}{\text{Vol}(\mathcal{G})} \cdot N(\epsilon, \mathcal{X}, d). \tag{28}$$

This volume contraction reduces the sample complexity required to reach a given retrieval radius ϵ and lowers retrieval bias at fixed knowledge base size N. This completes the proof of Proposition 4. □

F.6 Proof of Proposition 5 (Wasserstein-2 and CRPS Risk Bounds for Quantile Barycenters)

We bound the 2-Wasserstein discrepancy and expected CRPS excess error of the quantile Vincentization estimator.

Proof. For one-dimensional continuous distributions, the quadratic Wasserstein distance $\mathcal{W}_2$ between a predicted distribution $\hat{P}$ with quantile function $\hat{Q}$ and the true target distribution P with quantile function Q admits the exact closed-form characterization:

$$\mathcal{W}_2^2(\hat{P}, P) = \int_0^1 \big|\hat{Q}(\tau) - Q(\tau)\big|^2 d\tau = \|\hat{Q} - Q\|^2_{L_2([0,1])}. \tag{29}$$

The retrieved quantile barycenter under non-negative weights $\sum_{i=1}^k w_i(\mathbf{x}_q) = 1$ is $\hat{Q}_{\text{ret}}(\tau \mid \mathbf{x}_q) = \sum_{i=1}^k w_i(\mathbf{x}_q) Q_i(\tau)$, where $Q_i(\tau) = Q(\tau \mid \mathbf{x}_i)$ denotes the conditional quantile function at neighbor $\mathbf{x}_i$. Applying Jensen's inequality to the convex function $u \mapsto u^2$:

$$\big|\hat{Q}_{\text{ret}}(\tau \mid \mathbf{x}_q) - Q(\tau \mid \mathbf{x}_q)\big|^2 = \left|\sum_{i=1}^k w_i(\mathbf{x}_q)\big(Q(\tau \mid \mathbf{x}_i) - Q(\tau \mid \mathbf{x}_q)\big)\right|^2 \tag{30}$$

$$\leq \sum_{i=1}^k w_i(\mathbf{x}_q)\big|Q(\tau \mid \mathbf{x}_i) - Q(\tau \mid \mathbf{x}_q)\big|^2. \tag{31}$$

Integrating over $\tau \in [0, 1]$ and using Fubini's theorem:

$$\mathcal{W}_2^2\big(\hat{P}_{\text{ret}}(\cdot \mid \mathbf{x}_q),\, P(\cdot \mid \mathbf{x}_q)\big) \leq \sum_{i=1}^k w_i(\mathbf{x}_q) \int_0^1 \big|Q(\tau \mid \mathbf{x}_i) - Q(\tau \mid \mathbf{x}_q)\big|^2 d\tau \tag{32}$$

$$= \sum_{i=1}^k w_i(\mathbf{x}_q) \|Q(\cdot \mid \mathbf{x}_i) - Q(\cdot \mid \mathbf{x}_q)\|^2_{L_2}. \tag{33}$$

By the L_Q-Lipschitz continuity of the conditional quantile map on $(\mathcal{M}, d_{\mathcal{G}})$, $\|Q(\cdot \mid \mathbf{x}_i) - Q(\cdot \mid \mathbf{x}_q)\|_{L_2} \leq L_Q\, d_{\mathcal{G}}(\mathbf{x}_q, \mathbf{x}_i)$, which proves Equation 7:

$$\mathcal{W}_2^2\big(\hat{P}_{\text{ret}}(\cdot \mid \mathbf{x}_q),\, P(\cdot \mid \mathbf{x}_q)\big) \leq L_Q^2 \sum_{i=1}^k w_i(\mathbf{x}_q) d_{\mathcal{G}}(\mathbf{x}_q, \mathbf{x}_i)^2. \tag{34}$$

To connect this result to the Continuous Ranked Probability Score, we utilize the quantile decomposition of CRPS (Gneiting & Raftery, 2007; Laio & Tamea, 2007):

$$\text{CRPS}(\hat{P}, y) = \int_0^1 \text{QS}_\tau\big(\hat{Q}(\tau), y\big) d\tau, \tag{35}$$

where $\text{QS}_\tau(\hat{q}, y) = (\tau - \mathbb{I}(y < \hat{q}))(y - \hat{q})$ is the quantile pinball loss. Since the pinball loss is 1-Lipschitz in its first argument, $|\text{QS}_\tau(\hat{Q}(\tau), y) - \text{QS}_\tau(Q(\tau), y)| \leq |\hat{Q}(\tau) - Q(\tau)|$. Taking the expectation over $y \sim P(\cdot \mid \mathbf{x}_q)$ and integrating over $\tau \in [0, 1]$:

$$\mathbb{E}_y\big[\text{CRPS}(\hat{P}_{\text{ret}}, y) - \text{CRPS}(P, y)\big] \leq \int_0^1 \big|\hat{Q}_{\text{ret}}(\tau \mid \mathbf{x}_q) - Q(\tau \mid \mathbf{x}_q)\big| d\tau. \tag{36}$$

Applying the Cauchy-Schwarz inequality on $L_2([0, 1])$:

$$\int_0^1 \big|\hat{Q}_{\text{ret}}(\tau \mid \mathbf{x}_q) - Q(\tau \mid \mathbf{x}_q)\big| d\tau \leq \sqrt{\int_0^1 \big|\hat{Q}_{\text{ret}}(\tau \mid \mathbf{x}_q) - Q(\tau \mid \mathbf{x}_q)\big|^2 d\tau} = \mathcal{W}_2\big(\hat{P}_{\text{ret}}, P\big). \tag{37}$$

Combining this with the bound on $\mathcal{W}_2$ yields Equation 8, which bounds expected CRPS error by the weighted distance of retrieved neighbors. □