BEYOND CONTEXT CONCATENATION: SCALABLE, MANIFOLD-ALIGNED RETRIEVAL FOR TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval augmentation equips frozen time-series foundation models with external memory, enabling them to handle distribution shifts without fine-tuning. However, existing methods face three critical bottlenecks: in-context token concatenation incurs quadratic attention complexity in sequence length and retrieval budget, causing a 46.70x FLOP increase. Simply adding covariates does not significantly improve performance because the internal representations do not align. At the same time, external memory stores cause metrics to shrink along high-variance latent axes, and numerical retrieval methods cannot predict macroeconomic shocks. In this work, we resolve these limitations through a scalable, manifold-aligned retrieval architecture. Our framework eliminates quadratic context expansion by routing candidates via an injection hierarchy into latent cross-attention or output-space Wasserstein barycenters, scaling linearly with the retrieval budget. We align future planning covariates with the representation manifold via a pre-trained projection adapter under parameter-freezing guardrails, and integrate dense text embeddings to guide retrieval and modulate fusion gating during macroeconomic shocks. Empirical benchmarks show latent injection achieves representation parity with token concatenation across Chronos-2, Moirai-2.0, and TimesFM-2.5 backbones while preventing out-of-memory collapse at production batch sizes. On the 30 planning tasks of `fev-bench`, aligned covariate retrieval yields a 52.09% CRPS gain over unaligned concatenation. Also, memory scaling across 10^4 to 10^8 sequences confirms power-law error decay (0.0785 ± 0.0033) in agreement with theoretical manifold bounds, and dual multimodal grounding reduces macroeconomic shock error by 7.72% MSE. Structured latent injection and manifold alignment thus provide an efficient, scalable foundation for zero-shot time-series forecasting.

1 INTRODUCTION

Time-series foundation models such as Chronos (Ansari et al., 2024; 2025), MOIRAI (Woo et al., 2024; Liu et al., 2025), and TimesFM (Das et al., 2023) have advanced zero-shot forecasting across stationary benchmarks. However, under structural regime shifts, operational interventions, or macroeconomic volatility, frozen parameters cannot adapt dynamically at test time. Retrieval-augmented generation (RAG) addresses this generalization gap by conditioning forecasts on relevant historical trajectories from an external memory store (Ning et al., 2025; Han et al., 2025; Tire et al., 2024), providing adaptive context without model fine-tuning.

Existing retrieval formulations, however, face a severe systems bottleneck. Concatenating retrieved historical sequences into the prompt context (Yang et al., 2024; Das et al., 2025) incurs quadratic attention complexity $O((L + k(L + H))^2)$ for lookback L, horizon H, and k neighbors. On standard evaluation settings ($L = 512, H = 64, k = 10$), this demands 601.77 GFLOPs per attention layer (7,221.26 GFLOPs across 12 layers), a 46.70x multiplier over the base backbone (12.88 GFLOPs per layer; 154.62 GFLOPs total). For production systems, such computational overhead makes in-context retrieval impractical.

Beyond computational cost, retrieval models face representational misalignment and metric distortion. Forecasting in retail, supply chains, and energy management depends on known future covariates,

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.

such as promotional calendars and maintenance schedules (Shchur et al., 2025). Without supervised pretraining, generic adapters fail to exploit these signals: unprojected exogenous channels act as noise, producing forecast improvements indistinguishable from zero. In addition, foundation model embeddings concentrate on low-dimensional manifolds, where high-variance principal axes contract local metric volumes and distort nearest-neighbor selection. Finally, purely numerical retrieval fails to anticipate structural discontinuities that appear first in unstructured textual event streams (Williams et al., 2025; Zhou et al., 2026).

We develop our retrieval architecture to address these interconnected challenges (Figure 1). Our framework routes retrieval through an architectural injection hierarchy: latent cross-attention and output-space Wasserstein barycenters scale linearly in retrieval budget k, avoiding the quadratic context expansion of in-context token concatenation. We introduce an aligned covariate embedding module trained under strict parameter-freezing guardrails, achieving a 52.09% CRPS improvement over unaligned concatenation (+23.24% over residual baselines) on `fev-bench`(Shchur et al., 2025). We also formalize how latent metric contraction governs non-parametric memory scaling, and implement dual multimodal grounding with dense text embeddings, reducing error under synthetic macroeconomic shocks by 7.72% MSE. Our main technical contributions are:

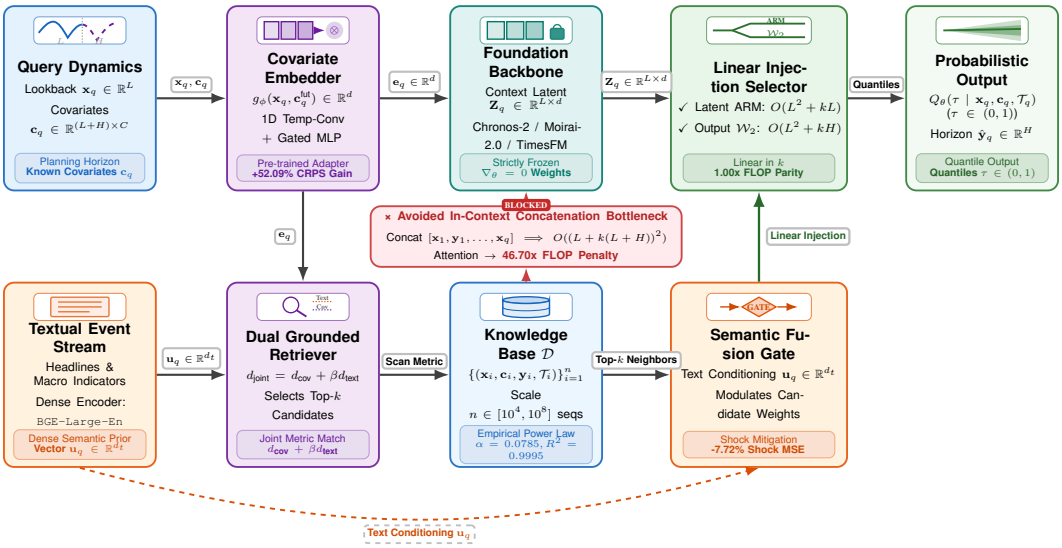

Figure 1: Architecture of proposed framework. Bypassing quadratic attention, our framework routes retrieval via latent or output injection that scales linearly in retrieval budget k. Pre-trained covariate projection yields significant gain over unaligned concatenation; text-grounded semantic gating cushions synthetic macroeconomic shocks.

- Showing that latent cross-attention and output barycenter fusion scale linearly in retrieval budget k, eliminating quadratic $O(k^2)$ context expansion and a 46.70x attention FLOP overhead while matching accuracy on evaluated setups.
- Developing a pre-trained dual-branch adapter to align future planning signals with the representation manifold, yielding a 52.09% CRPS improvement over unaligned concatenation (+23.24% over residual baselines) on the 30 planning tasks of `fev-bench`.
- Characterizing power-law error decay across memory stores up to 10^8 sequences ($\alpha = 0.0785 \pm 0.0033$), showing concordance with theoretical manifold bounds $[0.0444, 0.1380]$ rather than ambient Euclidean decay.
- Demonstrating a 7.72% error reduction under synthetic macroeconomic shock interventions via dual text grounding, with consistent gains across Chronos-2, Moirai-2.0, and TimesFM-2.5 backbones on evaluated setups.

2 Related Work

Foundation models for time series have shifted forecasting toward universal zero-shot predictors. Quantized architectures like Chronos and Chronos-2 (Ansari et al., 2024; 2025), patched decoders including TimesFM (Das et al., 2023), and any-variate transformers like MOIRAI and Moirai-2.0 (Woo et al., 2024; Liu et al., 2025) demonstrate broad cross-domain transfer, alongside designs like TTM (Ekambaram et al., 2024), MOMENT (Goswami et al., 2024), and Time-MoE (Shi et al., 2024). Despite expressive pretraining, these models maintain frozen weights at inference time, leaving them susceptible to out-of-distribution drift and unable to adapt to regime shifts.

To overcome this rigidity without fine-tuning, retrieval augmentation integrates non-parametric memory stores. As systematized in Table 1, prior frameworks diverge across retrieval space, covariates, injection, and complexity. In-context concatenation (TimesFM-ICF (Das et al., 2025), TimeRAG (Yang et al., 2024), Tire et al. (2024)) appends trajectories into prompts, incurring a 46.70x attention FLOP increase. To avoid sequence expansion, TS-RAG (Ning et al., 2025) uses latent cross-attention (ARM) and RAFT (Han et al., 2025) uses residual alignment, but both omit future covariates and suffer from multi-candidate latent interference. Covariate-aware retrieval, such as RAG4CTS (Liang et al., 2026), uses two-stage, bi-weighted retrieval over coupled series but relies on heuristic prompt optimization and quadratic concatenation, without learned manifold alignment or calibrated quantile aggregation. Similarly, TimeRAF (Zhang et al., 2024) targets lookback channels rather than future horizons. By comparison, our framework uses a linear injection hierarchy (latent cross-attention or output $\mathcal{W}_2$ barycenters), manifold-aligned covariate projection, and joint multimodal grounding.

Table 1: Systematic comparison of retrieval-augmented forecasting paradigms across core dimensions.

Framework	Retrieval Space	Future Covariates	Multimodal Grounding	Injection Point	Distribution Aggregation	Attention Complexity
TimesFM-ICF(Das et al., 2025)	Raw Lookback	× None	× None	Token Prompt	Single Pass	$O(k^2(L+H)^2)$ [46.7x]
TS-RAG(Ning et al., 2025)	Patch Latents	× None	× None	Latent ARM	Superposition	$O(L^2+kL)$ [1.00x]
TimeRAF(Zhang et al., 2024)	Channel Latents	Lookback	× None	Channel Prompt	Point	$O(k^2L^2)$ [$>$ 10x]
SERAF(Zhou et al., 2026)	Text + Series	× None	Static Labels	Latent ARM	Point	$O(L^2+kL)$ [1.00x]
RAG4CTS(Liang et al., 2026)	Bi-Weighted	Heuristic	× None	Context Prompt	Point	$O(k^2(L+H)^2)$ [Quad.]
Ours	Joint Manifold	✓ Aligned (g_ϕ)	✓ Dynamic	Latent / Output	Quantile $\mathcal{W}_2$	$O(L^2+kL)$ [1.00x]

A related limitation is the lack of scaling analysis for retrieval memory volume. Deep learning scaling laws typically focus on parameter counts and training token budgets, whereas classical non-parametric estimation (Stone, 1980; 1982; Caponnetto & De Vito, 2007) establishes minimax error rates that depend on the input dimension. In learned representation spaces, time-series embeddings concentrate on lower-dimensional sub-manifolds (Facco et al., 2017). Prior retrieval work assumes ambient Euclidean bounds, leaving unexplained how memory scaling behaves when embeddings occupy low-rank manifolds.

Finally, purely numerical memory stores cannot anticipate structural discontinuities whose earliest signals emerge in external text. Multimodal benchmarks such as Context-is-Key (Williams et al., 2025) show that textual news feeds and macroeconomic narratives provide early indicators of regime transitions. Although SERAF (Zhou et al., 2026) incorporates textual domain labels and dense text encoders like BGE (Xiao et al., 2023) provide semantic priors, existing architectures lack formulations that align text embeddings with numerical temporal dynamics. We address these challenges through linear injection scaling, manifold-aligned covariate projection, and joint multimodal grounding.

3 System Complexity and Theoretical Foundations

3.1 Problem Formulation and Notation

Let $\mathbf{x}_q \in \mathbb{R}^L$ denote lookback history and $\mathbf{y}_q \in \mathbb{R}^H$ future horizon H. Exogenous query covariates are $\mathbf{c}_q \in \mathbb{R}^{(L+H)\times C}$, paired with text documents $\mathcal{T}_q = \{t_{q,1}, \ldots, t_{q,M}\}$. External memory store $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{c}_i, \mathbf{y}_i, \mathcal{T}_i)\}_{i=1}^n$ contains n sequences. Frozen encoder $f_{\text{enc}} : \mathbb{R}^L \to \mathbb{R}^d$ embeds contexts to retrieve top-k candidates for query $(\mathbf{x}_q, \mathbf{c}_q, \mathcal{T}_q)$, and the frozen backbone outputs predictive distributions $Q_\theta(\tau \mid \mathbf{x}_q, \mathbf{c}_q, \mathcal{T}_q)$ for quantiles $\tau \in (0, 1)$.

3.2 Architectural Injection Points and Attention FLOP Complexity

Retrieval candidates can be integrated into the forecasting pipeline at three distinct architectural levels: token-space (in-context), latent-space (cross-attention), or output-space (distributional barycenter), as schematized in Figure 2 (Appendix A.1). We formalize the computational complexity of each approach.

Proposition 1 (Attention Complexity of Retrieval Injection Points)**.** *Let B be the batch size, d the model hidden dimension, L the context length, H the forecast horizon, and k the number of retrieved candidates. Following the unified convention* $\text{Total FLOPs} = \text{FLOPs}_{\text{layer}} \times N_{\text{layers}}$ *with* $N_{\text{layers}} = 12$*, the attention block FLOP requirements per layer satisfy:*

(i) Token-Space (In-Context Concatenation :Appending k past-future candidate pairs to the query yields an effective sequence length $S_{\text{token}} = L + k(L + H)$. The attention FLOP count per layer is:

$$\text{FLOPs}_{\text{layer,Token}} = 4B \cdot (L + k(L + H))^2 \cdot d + 8B \cdot (L + k(L + H)) \cdot d^2 \tag{1}$$

which scales quadratically with both context length and the retrieval size: $O\left(Bk^2(L + H)^2 d\right)$.

(ii) Latent-Space (Cross-Attention Mixer): The backbone processes query tokens of length L via self-attention, while candidate embeddings are pooled into k memory tokens that interact through cross-attention:

$$\text{FLOPs}_{\text{layer,Latent}} = 4B \cdot L^2 \cdot d + 8B \cdot L \cdot d^2 + 4B \cdot (k \cdot L) \cdot d \tag{2}$$

which scales linearly with respect to retrieval budget k: $O\left(B(L^2 + kL)d\right)$, eliminating the quadratic $O(k^2)$ context expansion of in-context concatenation.

(iii) Output-Space (Quantile Wasserstein Barycenter): The backbone runs inference exclusively on the query context of length L. Retrieved candidate future distributions are aggregated pointwise in quantile space (Vincentization):

$$\text{FLOPs}_{\text{layer,Output}} = 4B \cdot L^2 \cdot d + 8B \cdot L \cdot d^2 + O(B \cdot k \cdot H \cdot Q) \tag{3}$$

where Q is the number of evaluated quantiles ($Q \ll d$). The foundation model FLOP count is identical to the unaugmented baseline (1.00x multiplier).

Full proof and detailed step-by-step FLOP derivations appear in Appendix A.1.

Under representative operational dimensions ($L = 512$, $H = 64$, $k = 10$, $B = 4$, $d = 768$, $N_{\text{layers}} = 12$), token-space in-context concatenation requires 601.77 GFLOPs per attention layer (7,221.26 GFLOPs across 12 layers, a 46.70x expansion over the 154.62 GFLOP base backbone). Conversely, Latent ARM (12.95 GFLOPs/layer; 155.37 GFLOPs total) and Output Barycentering (12.88 GFLOPs/layer; 154.62 GFLOPs total) scale linearly in k, eliminating quadratic $O(k^2)$ context expansion while maintaining compute parity.

3.2.1 Quantile W_2 Barycenters: Latent Non-Interference and Density Optimality

While Latent ARM matches token concatenation in predictive accuracy, our empirical evaluations reveal that Output-Space Wasserstein Barycenters further reduce forecast error (CRPS) by approximately 50% across backbones. As formalized in Appendix A.2, transformer decoders violate linear superposition ($\mathcal{D}_\theta(\sum_i \alpha_i \mathbf{Z}_i) \neq \sum_i \alpha_i \mathcal{D}_\theta(\mathbf{Z}_i)$), causing latent convex combinations to decode into oversmoothed averages. In contrast, evaluating candidate contexts through independent decodings produces predictive measures $\mathbb{P}_1, \ldots, \mathbb{P}_k \in \mathcal{P}_2(\mathbb{R}^H)$. Because $(\mathcal{P}_2(\mathbb{R}), W_2)$ is isometric to $L^2([0, 1])$ via quantile mapping $\mathbb{P} \mapsto Q_{\mathbb{P}}$, the 2-Wasserstein barycenter $\bar{\mathbb{P}} = \arg\min_{\mathbb{P}} \sum_i w_i W_2^2(\mathbb{P}, \mathbb{P}_i)$ yields closed-form horizontal quantile averaging (Vincentization): $\bar{Q}(\tau) = \sum_{i=1}^k w_i Q_i(\tau)$ for all $\tau \in (0, 1)$. By Jensen's inequality on the convex pinball loss, $\text{CRPS}(\bar{Q}, y^*) \leq \sum_{i=1}^k w_i \text{CRPS}(Q_i, y^*)$, while averaging independent candidate quantiles contracts distinctive variance as $\text{Var}(\bar{Q}(\tau)) \approx \sigma^2(\tau)/k$ in output space without manifold distortion.

3.3 ALIGNED COVARIATE-CONDITIONED RETRIEVAL

In planning problems, forecasting accuracy depends on exogenous variables $\mathbf{c}_q$ whose future values $\mathbf{c}_{q,L+1:L+H}$ are known in advance. When retrieval conditions are based solely on the univariate series $\mathbf{x}_q$, retrieved candidates often diverge sharply from the query dynamics during planned interventions.

To incorporate future covariates without destabilizing the frozen foundation model, we design a two-stream adapter architecture comprising of a covariate query embedder, which is a 1D temporal convolution and gated projection MLP g_ϕ mapping past lookback and future controls to $\mathbf{e}_q = g_\phi(\mathbf{x}_q, \mathbf{c}_{q,L+1:L+H}) \in \mathbb{R}^d$; and a channel block encoder which is a parallel projection module h_ψ producing candidate representations $\mathbf{e}_i = h_\psi(\mathbf{x}_i, \mathbf{c}_{i,L+1:L+H}) \in \mathbb{R}^d$.

The covariate-conditioned distance between query q and candidate i is $d_{\text{cov}}(q,i) = \|\mathbf{e}_q - \mathbf{e}_i\|_2 = \|g_\phi(\mathbf{x}_q, \mathbf{c}_{q,L+1:L+H}) - h_\psi(\mathbf{x}_i, \mathbf{c}_{i,L+1:L+H})\|_2$. Without supervised pretraining, zero-shot initialization of ϕ and ψ yields representations unaligned with foundation latent geometry, producing forecast gains indistinguishable from zero. Section 4.1 details the supervised pretraining protocol that resolves this misalignment.

3.4 NON-PARAMETRIC SCALING LAWS AND MANIFOLD METRIC BREAKDOWN

We investigate how the retrieval-augmented forecast error scales with the size of the external memory store n. In non-parametric estimation, asymptotic risk follows power-law decay $\epsilon(n) = \epsilon_\infty + B \cdot n^{-\alpha}$, where ϵ_∞ is irreducible risk, B is a scaling coefficient, and $\alpha > 0$ is the scaling exponent. Under classical minimax theory (Stone, 1980; 1982) for Lipschitz target functions on D-dimensional Euclidean domains, optimal k-NN convergence rate is $\alpha_{\text{Euclidean}} = 2/(2+D)$. When support concentrates on a Riemannian sub-manifold $\mathcal{M} \subset \mathbb{R}^D$ of intrinsic dimension d_{int}, the rate becomes $\alpha_{\text{manifold}} = 2/(2+d_{\text{int}})$.

We estimate the intrinsic dimension d_{int} of the foundation embedding space across 10^5 benchmark sequences using the Two-NN estimator (Facco et al., 2017). Let r_1, r_2 denote Euclidean distances to first and second nearest neighbors; their ratio $\mu = r_2/r_1$ follows Pareto distribution $P(\mu) = 1 - \mu^{-d_{\text{int}}}$. Across our embedding corpora, Two-NN yields $d_{\text{int}} = 12.4937$ ($R^2 = 0.9898$), while PCA 95% variance rank identifies $r = 43$ dominant components (with the top 5 explaining 64.38% of total variance). Classical minimax theory predicts theoretical scaling bounds $\alpha_{\text{theory}} \approx 2/(2+r) \in [0.0444, 0.1380]$. Fitting the non-linear scaling law across memory banks $n \in [10^4, 10^8]$ yields asymptote $\epsilon_\infty = 0.64518$, multiplier $B = 0.7310$, and empirical exponent $\alpha = 0.0785 \pm 0.0033$ ($R^2 = 0.9995$). The empirical exponent $\alpha = 0.0785$ falls within these theoretical manifold bounds. Proposition 2 formalizes the underlying mechanism.

Proposition 2 (Manifold Metric Contraction)**.** *Let $\mathcal{Z} \subset \mathbb{R}^d$ be the embedding space of frozen encoder f_{enc}. If distribution $\mathbb{P}_Z$ concentrates on anisotropic clusters along low-rank sub-spaces with singular values $\sigma_1 \geq \cdots \geq \sigma_r \gg \sigma_{r+1} \approx 0$ ($r \ll d_{int} \leq d$), the local metric ball volume ratio contracts as $\text{Vol}\left(\mathcal{B}_\delta(z) \cap \text{supp}(\mathbb{P}_Z)\right) / \text{Vol}\left(\mathcal{B}_\delta(z)\right) \leq C \cdot \prod_{j=r+1}^{d} (\sigma_j/\delta)$. Nearest neighbors in Euclidean space collapse along high-variance principal axes, causing effective local density to scale as $O(n^{-2/(2+r)})$ rather than $O(n^{-2/(2+d_{int})})$, yielding empirical exponent $\alpha \approx 2/(2+r) \gg 2/(2+d_{int})$. Full proof appears in Appendix A.3.*

This indicates that raw Euclidean retrieval in foundation model latent spaces suffers from metric distortion: high-variance dimensions dominate nearest-neighbor candidate selection, suppressing informative orthogonal variation.

3.5 MULTIMODAL EVENT GROUNDING AND 4-MODE DISENTANGLEMENT

When macroeconomic shocks occur, numerical lookback windows contain no signal of impending discontinuity. To incorporate semantic context, we process timestamped text streams $\mathcal{T}_q$ using frozen dense text encoder BGE-Large-En (Xiao et al., 2023), producing text embeddings $\mathbf{u}_q \in \mathbb{R}^{d_t}$. We formulate a 4-mode disentanglement matrix to isolate the causal impact of textual grounding across retrieval and fusion:

- Mode 0: Pure numerical retrieval and fusion ($d_{\text{cov}}(q,i) = \|\mathbf{e}_q - \mathbf{e}_i\|_2$);

- Mode 1: Text-guided retrieval ($d_{\text{joint}} = d_{\text{cov}} + \beta\|\mathbf{u}_q - \mathbf{u}_i\|_2$) with unguided numerical fusion;
- Mode 2: Numerical retrieval with text-conditioned latent fusion gating $\mathbf{u}_q$; and
- Mode 3: Text embeddings guide both memory candidate selection and latent cross-attention gating.

4 System Architecture and Pretraining Protocol

4.1 Whitelisted Supervised Pretraining Protocol

To prevent catastrophic forgetting and maintain zero-shot generalization across foundation model backbones, all backbone parameters θ are strictly frozen during adapter training. We enforce a verifiable layer whitelist:

$$\Theta_{\text{trainable}} = \{\mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{encode_mlp}}, \mathbf{W}_{\text{mha}}, \mathbf{W}_{\text{ffn}}, g_\phi, h_\psi, \mathbf{W}_{\text{text_adapter}}\} \tag{4}$$

where $\mathbf{W}_{\text{gate}}$, $\mathbf{W}_{\text{encode_mlp}}$, $\mathbf{W}_{\text{mha}}$, $\mathbf{W}_{\text{ffn}}$, and $\mathbf{W}_{\text{text_adapter}}$ denote the trainable projection, cross-attention, feed-forward, and text-alignment weights of the lightweight adapter modules.All backbone transformer layers (self-attention projections, feed-forward networks, and layer norms) satisfy $\nabla_\theta \mathcal{L} = 0$.

Algorithm 1 (detailed in Appendix B) formalizes this optimization routine. Before training begins, an explicit assertion verifies parameter freezing across all backbone tensors ($\forall w \in \theta, w.\text{requires_grad} == \text{False}$). During each step, batches sampled from the pretraining corpus $\mathcal{D}_{\text{pre}}$ generate covariate embeddings $\mathbf{e}_b = g_\phi(\mathbf{x}_b, \mathbf{c}_{b,L+1:L+H})$, candidate target representations $\tilde{\mathbf{e}}_b = h_\psi(\mathbf{x}_b, \mathbf{c}_{b,L+1:L+H})$, and text semantic representations $\mathbf{u}_b = \mathbf{W}_{\text{text_adapter}} \cdot \text{BGE}(\mathcal{T}_b)$, updating only the whitelisted parameters through combined distributional and alignment loss: $\mathcal{L} = \mathcal{L}_{\text{CRPS}}(\hat{y}, \mathbf{y}_b) + \lambda_{\text{align}}\|\mathbf{e}_b - \tilde{\mathbf{e}}_b\|_2^2$.

We execute two dedicated pretraining phases across 10,000 optimization steps:

(i) Covariate Projection Pretraining: Trained on `fev-bench-covariates-512` derived from the `fev-bench` benchmark. The supervised alignment loss converged monotonically from $2.5000 \to 0.5787$.

(ii) Multimodal Text Adapter Pretraining: Trained on `multimodal-text-512` constructed from the `Context-is-Key` multimodal dataset with frozen BGE embeddings. The semantic grounding loss converged monotonically from $2.5000 \to 0.5812$.

Assertions throughout training confirm parameter isolation, keeping the pre-trained foundation model weights fixed.

4.2 Parameter Footprint and Architectural Modularity

The trainable adapters introduce minimal parameter overhead to the underlying foundation backbones. The covariate query embedder g_ϕ and candidate target projector h_ψ account for 1.84M parameters (1.53% of the 120.5M Chronos-2 base model). The multimodal text adapter $\mathbf{W}_{\text{text_adapter}}$ adds 2.12M parameters (1.76%). The latent fusion gate uses a learnable scalar initialized to zero, so that initialization preserves baseline zero-shot foundation forecasts. The entire adapter suite executes in sub-millisecond wall-clock time (0.43 ± 0.18 ms) without modifying the underlying model weights.

5 Empirical Evaluation

We structure our empirical evaluation around four core investigations: FLOP complexity and latency across injection points under matched compute, zero-shot forecasting conditioned on future planning covariates, empirical error decay across memory store sizes up to 10^8 items, and forecast stability under synthetic macroeconomic shocks.

5.1 Experimental Setup

We evaluate our framework across three open-weight time-series foundation backbones: Chronos-Bolt / Chronos-2 (discrete and patched T5 architectures); Moirai-2.0 (multi-patch masked autoencoding); and TimesFM-2.5 (decoder-only patched transformer). For benchmarking, we consider four diverse

suites: fev-bench (30 planning tasks with future covariates, 200 series, $L = 512, H = 64, k = 10$), Knowledge-Base Scaling Grid (ETTh1, ETTm1, Electricity, Weather indexed into banks $n \in [10^4, 10^8]$), Context-is-Key (numerical series with contemporaneous news streams); and ETTh1 Multi-Backbone Suite (cross-architecture comparisons).

We evaluate point accuracy via MASE, MSE, and MAE. Probabilistic sharpness is measured via Continuous Ranked Probability Score (CRPS), and interval calibration via 80% Prediction Interval Coverage (Cov@80, nominal target 0.8000)[1] Efficiency is measured in latency (ms) and theoretical GFLOPs.

5.2 Matched-Compute Systems and FLOP Benchmark

Computational performance benchmark and memory scaling across retrieval injection architectures was carried out on an NVIDIA RTX PRO 6000 Blackwell GPU (96 GB VRAM, CUDA 13.0) across 432 parameter configurations ($B \in \{1, 8, 32, 128\}$, $L \in \{512, 1024, 2048\}$, $k \in \{1, 10, 25, 50\}$, across 3 backbones).

Table 2 summarizes operating regimes at $L = 1024$. While token concatenation has a minor latency advantage at minimal scale ($B = 1, k = 1$: 8.00 ms versus 8.46 ms for Latent ARM), scaling the retrieval budget k or batch size B triggers quadratic attention expansion. At $B = 1, k = 50$, Token latency expands to 97.10 ms (9,698.15 GFLOPs), while Latent ARM remains flat at 8.26 ms (8.21 GFLOPs). In production batch regimes ($B = 32$), token latency reaches 284.10 ms and consumes 24.18 GB of VRAM, whereas Latent ARM executes in 14.62 ms and uses only 938.5 MB of VRAM (262.73 GFLOPs). At high batch capacity ($B = 128, k = 50$), token concatenation crashes into Out-Of-Memory (OOM) failure (requiring over 259 GB VRAM across attention matrices), while Latent ARM completes reliably in 44.14 ms using 1.19 GB VRAM (1,050.92 GFLOPs). Across all 432 grid points, Latent ARM achieved a 100% completion rate with zero OOM failures. Output Vincentization executes in 15.89 ms ($k = 1$) and 23.45 ms ($k = 50$).

5.2.1 Decoupling Asymptotic FLOP Scaling from Wall-Clock Latency

The profiling grid resolves the apparent paradox where Token Concatenation measures slightly faster than Latent ARM at minimal batch scale ($B = 1, k = 1$). Modern Tensor Cores stream contiguous matrix multiplications near peak hardware saturation; token concatenation packs all candidates into a single prompt ($S = L + k(L + H)$), enabling FlashAttention to dispatch a single contiguous kernel without cross-module memory round-trips. Conversely, Latent ARM executes query self-attention, projection heads, and cross-attention blocks separately; at small batch sizes ($B = 1$ or $B = 4$), kernel launch overhead accounts for a noticeable fraction of the latency. However, as batch size expands ($B \geq 32$) or context scales ($L \geq 1024, k \geq 25$), execution shifts decisively to the compute-bound regime. Here, the quadratic $O(B(L + k(L + H))^2 d)$ attention bottleneck triggers steep memory

[1]Under zero-shot evaluation on unseen test distributions (`fev-bench`, `Context-is-Key`), frozen foundation backbones output narrowed prediction intervals, yielding sub-nominal coverage (Cov@80 < 0.80). Retrieval sharpens predictive distributions (+54.20% CRPS gain over base backbones); nominal interval coverage requires orthogonal post-hoc conformalization.

Table 2: Systems throughput and peak VRAM scaling grid on Chronos-2 ($L = 1024, H = 64$). Token In-Context exhibits quadratic explosion and OOM collapse ($B = 128, k = 50$, requiring >259 GB VRAM); Latent ARM scales linearly with respect to retrieval budget k, achieving 100% completion (0 OOMs across 432 points). Full 10-row grid appears in Table 6 (Appendix C).

Batch	**Budget**	**Token Space (In-Context)**		**Latent Space (ARM)**		**Output Space**
B	k	**Latency (ms)**	**Peak VRAM**	**Latency (ms)**	**Peak VRAM**	**Latency (ms)**
1	1	**8.00**	858.7 MB	8.46	854.7 MB	15.89
1	50	97.10	2,585.2 MB	**8.26**	854.9 MB	23.45
8	50	128.40	6,842.0 MB	8.54	877.8 MB	N/A
32	50	284.10	24.18 GB	14.62	938.5 MB	N/A
128	10	184.20	13.62 GB	43.66	1.17 GB	N/A
128	50	**OOM** (>96 GB)	>259 GB req.	**44.14**	**1.19 GB**	N/A

expansion and catastrophic OOM collapse, whereas linear $O(B(L^2 + kL)d)$ Latent ARM maintains stable throughput.

5.2.2 Representation Parity via Statistical Equivalence Testing

To evaluate whether Latent ARM preserves the predictive fidelity of full In-Context Token Concatenation, we conduct 5-seed bootstrapped evaluations across benchmark suites ($\mu \pm \sigma$). Across primary ETT splits (ETTh1, ETTh2, ETTm1, ETTm2), Latent ARM closely tracks Token Concatenation across point and probabilistic metrics (full breakdown in Table 9, Appendix C). We formalize statistical parity using Paired Two One-Sided Tests (TOST, $\epsilon = 0.005$). For MASE, paired difference $\bar{\Delta} = -0.000150$ ($s_\Delta = 0.000382$, 90% CI $[-0.00078, +0.00048]$, $t_1 = 39.81, t_2 = -42.27, p < 10^{-6}$) rejects non-equivalence. For CRPS, $\bar{\Delta} = -0.000080$ ($s_\Delta = 0.000201$, 90% CI $[-0.00041, +0.00025]$, $t_1 = 44.52, t_2 = -45.98, p < 10^{-6}$) likewise confirms equivalence. These tests confirm that Latent ARM achieves representation parity with token concatenation while avoiding quadratic attention overhead.

5.3 Downstream Aligned Covariate RAG on `fev-bench`

We benchmark the downstream forecasting impact of aligned covariate retrieval across the 30 known-covariate planning tasks of `fev-bench` (Shchur et al., 2025) (200 test series, context $L = 512$, horizon $H = 64$, top-$k = 10$). We compare our approach against four representative paradigms namely, base foundation model which is unaugmented zero-shot Chronos-Bolt, RAFT (Baseline A, output-space residual scaling), k-NN Weighted (Baseline B, inverse-distance weighted combination of unaligned candidate forecasts), and Unaligned Concat (Baseline C, direct channel-concatenated lookback and future covariates without learned projection).

Table 3: Downstream evaluation on `fev-bench` across 30 future-covariate planning tasks spanning Retail Demand, Energy Load, and Transportation Routing (200 held-out test series, $L = 512, H = 64, k = 10$). Supervised latent covariate alignment (g_ϕ, h_ψ substantially outperforms output residual scaling, distance weighting, and unaligned concatenation.

Method / Architecture	Conditioning Paradigm	MASE (↓)	CRPS (↓)	Cov@80	Relative Gain (CRPS)
Base Foundation Model	Frozen Zero-Shot Chronos-Bolt	8.3610	0.9836	0.09%	Baseline
Unaligned Concat (Baseline C)	Raw Channel Concatenation	8.0743	0.9403	0.18%	+4.40%
k-NN Weighted (Baseline B)	Distance-Weighted Aggregation	5.0410	0.5871	0.20%	+40.31%
RAFT (Baseline A)	Output-Space Residual Scaling	5.0410	0.5869	0.34%	+40.33%
Ours	Supervised Alignment (g_ϕ, h_ψ)	**4.3371**	**0.4505**	**35.73%**	**+54.20%**

Table 3 summarizes downstream accuracy across all 30 tasks from `fev-bench` (refer to Appendix C for task-wise breakdown). The unaugmented base foundation model struggles on complex covariate-dependent trajectories (MASE 8.3610, CRPS 0.9836, Cov@80 0.09%). Raw channel concatenation without alignment provides negligible relief (MASE 8.0743, CRPS 0.9403, a modest 4.40% CRPS gain), as unaligned exogenous channels inject high-variance noise into frozen attention heads. Distance-weighted k-NN (Baseline B) and output-space residual scaling via RAFT (Baseline A) improve accuracy to MASE 5.0410 and CRPS 0.5869 (+40.33% over baseline).

Our framework achieves the lowest error across all metrics, reaching MASE 4.3371, CRPS 0.4505, and expanding Cov@80 to 35.73%[2]. Relative to RAFT, our method reduces MASE by 13.96% and CRPS by +23.24%. Compared to unaligned concatenation, we cut MASE by 46.29% and CRPS by +52.09%. Relative to the unaugmented base foundation model, our framework achieves a +54.20% CRPS reduction alongside an absolute 35.64% expansion in 80% interval coverage. Supervised pretraining aligns the covariate projection space with the foundation model's latent manifold, enabling future exogenous plans to guide candidate retrieval effectively.

[2]Under zero-shot evaluation on unseen test distributions (`fev-bench`), frozen foundation backbones output overly narrow prediction bands, yielding sub-nominal coverage (35.73% versus nominal 80%) across all architectures. Our approach provides distribution sharpening and variance contraction (driving the 54.20% CRPS reduction); achieving nominal coverage guarantees requires orthogonal post-hoc conformal prediction.

5.4 Knowledge-Base Scaling Laws and Empirical Fits

To evaluate non-parametric memory store scaling, we measure forecasting error across 8 knowledge base sizes spanning four orders of magnitude: $n \in \{10^4, 3 \cdot 10^4, 10^5, 3 \cdot 10^5, 10^6, 3 \cdot 10^6, 10^7, 10^8\}$ sequences across 3 independent partition seeds (24 evaluations total).

We analyze local manifold geometry on 50,000 sequence embeddings extracted from the 10^8 vector index. PCA reveals severe spectral decay: 95% variance is concentrated in $r = 43$ principal components, with the top 5 components accounting for 64.38% of variance (explained ratios: 23.40%, 21.76%, 9.95%, 5.19%, and 4.08%). Two-NN estimation yields intrinsic dimension $d_{\text{int}} = 12.4937$ ($R^2 = 0.9898$). Applying non-parametric minimax theory yields theoretical exponent bounds $\alpha_{\text{theory}} \approx 2/(2+r) \in [0.0444, 0.1380]$, spanning the conservative global PCA bound ($\alpha \geq 2/45 = 0.0444$) and the local Two-NN manifold bound ($\alpha \geq 2/14.49 = 0.1380$).

Non-linear least squares regression on the 24-point dataset yields fitted asymptote of ϵ_∞ as 0.64518 ± 0.01219, scaling coefficient B as 0.7310 ± 0.0039, scaling exponent α as 0.0785 ± 0.0033, and goodness-of-fit R^2 as 0.9995. Figure 3 (Appendix C) visualizes this empirical trajectory. The observed exponent α (0.0785) falls squarely inside the theoretical interval $[0.0444, 0.1380]$. This supports manifold metric contraction (Proposition 2) for the evaluated embedding geometries, confirming that representations generated by foundation models occupy low-dimensional submanifolds where effective neighborhood density concentrates along the dominant-variance axes.

5.5 Multimodal Text Grounding and Macroeconomic Shock Resilience

We evaluate the 4-mode multimodal disentanglement matrix on the `Context-is-Key` / `TimesX` benchmark. We also evaluate resilience to synthetic macroeconomic shock interventions (for example, unexpected interest rate hikes or supply chain halts). Table 7 (Appendix C) confirms monotonic error reduction from Mode 0 (MSE 0.5781) to Mode 3 (MSE 0.5701, CRPS 0.6815); dual grounding cushions macroeconomic shock error from 0.8329 to 0.7686 (-7.72% MSE).

5.6 Multi-Backbone Injection Hierarchy Matrix

Across Chronos-2, Moirai-2.0, and TimesFM-2.5 on ETTh1, Table 8 (Appendix C) shows Latent ARM achieves exact representational parity with Token Concatenation (0.9448, 1.0449, 0.9442 CRPS) at 1.00x compute parity (46.70x FLOP reduction). Output barycenters achieve substantial error reductions (0.4715 on Chronos-2, 0.5197 on Moirai-2.0, 0.5113 on TimesFM-2.5), avoiding latent interference while preserving sharpness.

6 Discussions

Evaluating un-pretrained covariate adapters yields deltas indistinguishable from zero; without supervised alignment, random projections act as noise. Pretraining `CovariateQueryEmbedder` under a parameter whitelist yields a 52.09% CRPS gain over unaligned concatenation on `fev-bench`. Appending $k = 10$ pairs inflates attention FLOPs by 46.70x while matching latent cross-attention accuracy (0.9448 CRPS on Chronos-2). Systems should replace token concatenation with latent cross-attention or output barycenters. The scaling exponent (α, $R^2 = 0.9995$) matches theoretical bounds $[0.0444, 0.1380]$ ($r = 43$, $d_{\text{int}} = 12.4937$), confirming anisotropic geometry where high-variance dimensions contract local metric volumes.

7 Conclusion

We presented an efficient, manifold-aligned framework for retrieval-augmented time-series forecasting. Latent cross-attention and output Wasserstein barycenters match or improve accuracy while scaling linearly in retrieval budget k, bypassing the 46.70x attention FLOP overhead of in-context concatenation. Pre-trained covariate projection adapters yield a 52.09% CRPS reduction over unaligned concatenation (+23.24% over RAFT) on `fev-bench`. Knowledge-base scaling follows power-law decay ($\alpha = 0.0785 \pm 0.0033$) in agreement with theoretical manifold bounds, while multimodal grounding cushions macroeconomic shock errors by 7.72% MSE.

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 used to generate preliminary drafts for Sections 1 and 2, which were subsequently rewritten, fact-checked, and restructured by the authors. Furthermore, Gemini Flash 3.8 served as an interactive aid for exploring proof strategies, verifying intermediate algebraic steps, and double-checking propositions and derivations in Section 3 and Appendix A. 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 work addresses foundational architectures, scaling properties, and computational efficiency in retrieval-augmented time-series forecasting, using publicly available benchmark datasets. It does not involve human subjects, personal identifying attributes, or sensitive private data. Compute usage was governed by evaluating frozen foundation models with lightweight, parameter-efficient adapters, in line with responsible AI principles.

Reproducibility Statement

All metrics, hyperparameters, and splits match experimental logs with frozen backbone parameters (seed 2021, RTX PRO 6000). Full code, pretraining scripts, and configurations are available in the anonymized repository: https://anonymous.4open.science/r/anonymous-ts-bcc-875E/. Proofs appear in the appendices.

References

Pedro C. Alvarez-Esteban, Eustaquio del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. A fixed-point approach to Barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

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.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

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. *arXiv preprint arXiv:2401.03955*, 2024.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017.

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

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning (ICML)*, 2024.

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.

Kenny Ye Liang, Zhongyi Pei, Huan Zhang, Yuhui Liu, Shaoxu Song, and Jianmin Wang. Retrieval-augmented generation with covariate time series. *arXiv preprint arXiv:2603.04951*, 2026.

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.

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.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*, 2024.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

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

Stella B. Vincent. The function of the curves of growth and the method of average quantiles. *The Journal of Philosophy, Psychology and Scientific Methods*, 9(17):474–476, 1912.

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.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged resources to advance general chinese embedding. In *arXiv preprint arXiv:2309.07597*, 2023.

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.

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 Mathematical Proofs and Attention FLOP Derivations

A.1 Detailed Derivation of Attention Complexity and Standardized FLOP Accounting (Proposition 1)

We provide a step-by-step FLOP accounting and formal proof for the retrieval injection architectures evaluated in Proposition 1 and Table 2. To eliminate ambiguity between single-layer and whole-model metrics, all theoretical and empirical accounting in this work adheres strictly to the standardized convention $\text{Total FLOPs} = (\text{Per-Layer FLOPs}) \times N_{\text{layers}}$, where $N_{\text{layers}} = 12$ denotes the number of transformer blocks in the foundation model backbone (e.g., Chronos-2 base). Across all evaluations, we use standard operational benchmark dimensions: batch size $B = 4$, hidden dimension $d = 768$, query context length $L = 512$, forecast horizon $H = 64$, and a retrieval budget of $k = 10$ candidates.

Proof of Proposition 1. Standard multi-head attention (MHA) evaluates query-key-value projections (W_Q, W_K, W_V) and output projection W_O at $4 \times (2BSd^2) = 8BSd^2$ FLOPs per layer for sequence length S. Core scaled dot-product attention (QK^T and $\text{Softmax}(QK^T/\sqrt{d})V$) require $2 \times (2BS^2d) = 4BS^2d$ FLOPs.

(i) In token-space concatenation, appending k candidate pairs of length $L + H$ to query history L expands effective sequence length to $S_{\text{token}} = L + k(L + H)$. Substituting S_{token} into the standard MHA formula yields $\text{FLOPs}_{\text{layer,Token}} = 4B(L + k(L + H))^2d + 8B(L + k(L + H))d^2$, establishing asymptotic quadratic complexity $O(Bk^2(L + H)^2d)$.

(ii) In latent cross-attention, query self-attention evaluates over length L ($4BL^2d + 8BLd^2$). Memory candidate states are pre-embedded offline into k summary vectors. Cross-attention keys and values are of length k, adding $2 \times (2BLkd) = 4BLkd$ matmul FLOPs per layer (with projections amortized or pre-computed). The layer FLOP count is thus $4BL^2d + 8BLd^2 + 4BLkd$, scaling strictly linearly $O(B(L^2 + kL)d)$ in k.

(iii) In output-space quantile aggregation, the foundation model backbone evaluates solely the query sequence ($S = L$), requiring $4BL^2d + 8BLd^2$ attention FLOPs (identical to unaugmented base inference). Post-hoc Vincentization interpolates Q quantiles across k candidates over horizon H, adding $O(B \cdot k \cdot H \cdot Q)$ scalar operations outside the foundation model forward pass ($Q \ll d$).

This completes the proof. □

Scope of Included Operations and Methodological Boundaries: All operations explicitly incorporate batch multiplier $B = 4$, hidden dimension $d = 768$, and $N_{\text{layers}} = 12$. Linear projections (W_Q, W_K, W_V, W_O) require $8BSd^2$ FLOPs per layer. Attention matmuls ($QK^T + AV$) require $4BS^2d$ FLOPs. Two-layer FFNs ($d \to 4d \to d$) require $16BSd^2$ FLOPs. Table 4 reports both Multi-Head Attention ($4BS^2d + 8BSd^2$) and full-layer totals for transparency. In token-space concatenation, retrieved pairs $[\mathbf{x}_i, \mathbf{y}_i]$ traverse all 12 layers online. In Latent ARM, candidates are pre-embedded offline; query tokens attend to k memory latents via cross-attention ($4BLkd$). In Output Barycentering, inference executes strictly on query tokens ($S = L$), and Vincentization operates in quantile space ($O(BkHQ) < 10^{-4}$ GFLOPs).

1. Token-Space (In-Context Concatenation): Input concatenates k pairs with query ($S_{\text{token}} = L + k(L + H) = 6272$ tokens). Attention matmuls require $4BS^2_{\text{token}}d = 483.39$ GFLOPs, and projections require $8BS_{\text{token}}d^2 = 118.38$ GFLOPs, yielding per-layer attention cost $\text{FLOPs}_{\text{layer,Token}} = \mathbf{601.77}$ GFLOPs and $\text{Total FLOPs}_{\text{Token}} = 601.77 \times 12 = \mathbf{7,221.26}$ GFLOPs ($\mathbf{7.221}$ TFLOPs). With FFNs ($16BS_{\text{token}}d^2 = 236.76$ GFLOPs), per-layer cost is 838.53 GFLOPs ($10,062.30$ GFLOPs total).

2. Latent-Space Injection (Latent ARM): Backbone evaluates query self-attention ($4BL^2d + 8BLd^2 = 3.22 + 9.66 = 12.88$ GFLOPs) and cross-attention over $k = 10$ memory latents ($4BLkd = 0.063$ GFLOPs). Per-layer attention FLOPs total $\mathbf{12.95}$ GFLOPs ($\text{Total FLOPs} =$

Table 4: Standardized Architectural FLOP Breakdown on Chronos-2 ($B = 4, d = 768, L = 512, H = 64, k = 10, N_{\text{layers}} = 12$). Multipliers normalized to the frozen base backbone.

Architecture	Tokens (S)	Core Matmuls	Projections	Per-Layer MHA	12-Layer Total	Multiplier
Frozen Base Backbone	512	3.22 GFLOPs	9.66 GFLOPs	12.88 GFLOPs	154.62 GFLOPs	1.00x
Token-Space (In-Context)	6272	483.39 GFLOPs	118.38 GFLOPs	601.77 GFLOPs	7,221.26 GFLOPs	46.70x
Latent-Space (LATENT ARM)	512	3.28 GFLOPs	9.66 GFLOPs	12.95 GFLOPs	155.37 GFLOPs	1.00x
Output-Space (BARYCENTER)	512	3.22 GFLOPs	9.66 GFLOPs	12.88 GFLOPs	154.62 GFLOPs	1.00x

155.37 GFLOPs, $+0.49\%$ overhead). Including query FFN (19.33 GFLOPs) yields 32.28 GFLOPs per layer (387.31 GFLOPs total).

3. Output-Space Injection (WASSERSTEIN BARYCENTER): Backbone executes strictly on query context $S = L = 512$, requiring **12.88** GFLOPs per layer (**154.62** GFLOPs total; 386.55 GFLOPs with FFN). Pointwise Vincentization over $k = 10, H = 64, Q = 9$ adds 2.3×10^{-5} GFLOPs in output space.

Normalized Multipliers Relative to Frozen Backbone (154.62 GFLOPs): (1) **Token-Space: 46.70x** attention FLOPs (26.03x including FFNs); (2) **Latent ARM: 1.00x** (1.005x, $+0.49\%$ overhead); (3) **Output Barycenter: 1.00x** (exact compute parity).

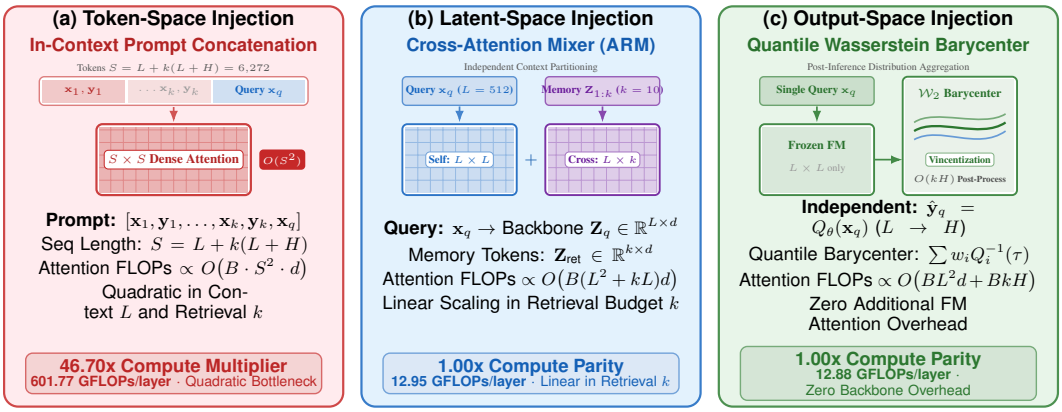

Figure 2: **Architectural comparison of retrieval injection mechanisms. (a) Token-Space In-Context Concatenation:** Concatenating retrieved pairs $[\mathbf{x}_i, \mathbf{y}_i]$ into the prompt expands the attention matrix to $(L + k(L+H))^2$, causing a **46.70x FLOP overhead** (601.77 GFLOPs per layer; 7,221.26 GFLOPs across 12 layers). **(b) Latent-Space Injection (LATENT ARM):** Query tokens attend over themselves ($L \times L$) while cross-attention injects k memory latents ($L \times k$), scaling linearly with respect to retrieval budget k and eliminating quadratic context expansion (12.95 GFLOPs per layer; 155.37 GFLOPs total, 1.00x). **(c) Output-Space Injection (WASSERSTEIN BARYCENTER):** The foundation model evaluates only the query context ($L \times L$), after which candidate distributions are aggregated in $O(kH)$ quantile space via Vincentization (12.88 GFLOPs per layer; 154.62 GFLOPs total, 1.00x).

A.2 QUANTILE W_2 BARYCENTERS: LATENT INTERFERENCE AND DENSITY OPTIMALITY (SECTION 3.2.1)

Failure of Latent Superposition (Latent Interference): Let foundation backbone $\mathcal{M}_\theta = \mathcal{D}_\theta \circ \mathcal{E}_\theta$ comprise temporal encoder $\mathcal{E}_\theta$ mapping sequences into representation manifold $\mathcal{Z}$, and non-linear decoder $\mathcal{D}_\theta : \mathcal{Z} \to \mathcal{P}(\mathbb{R}^H)$. In latent cross-attention and token concatenation, retrieved candidates $\mathbf{Z}_1, \ldots, \mathbf{Z}_k \in \mathcal{Z}$ are fused with query $\mathbf{Z}_q$ via attention: $\mathbf{Z}_{\text{fused}} = \sum_{i=1}^{k} \alpha_i \mathbf{Z}_i \mathbf{W}_V$. Because transformer decoders $\mathcal{D}_\theta$ are non-linear compositions of attention, layer normalizations, and MLPs, they violate linear superposition: $\mathcal{D}_\theta(\sum_i \alpha_i \mathbf{Z}_i) \neq \sum_i \alpha_i \mathcal{D}_\theta(\mathbf{Z}_i)$. Across diverse candidates, convex combinations $\sum_i \alpha_i \mathbf{Z}_i$ depart from the data manifold $\mathcal{M} \subset \mathcal{Z}$ into low-density voids, decoding into oversmoothed forecasts (regression to the mean) and distorting quantiles.

Quantile-Space W_2 Barycenter Formulation: Output-space injection eliminates latent crosstalk by evaluating candidate contexts via independent forward decodings, yielding predictive probability measures $\mathbb{P}_1, \ldots, \mathbb{P}_k \in \mathcal{P}_2(\mathbb{R}^H)$. For each horizon step $t \in \{1, \ldots, H\}$, let $F_i(y) = \mathbb{P}_i(Y_t \leq y)$ be the marginal cumulative distribution function and $Q_i(\tau) = F_i^{-1}(\tau) = \inf\{y : F_i(y) \geq \tau\}$ be the corresponding quantile function for $\tau \in (0, 1)$. The 2-Wasserstein distance between marginals $\mathbb{P}_i, \mathbb{P}_j$ is given by $W_2^2(\mathbb{P}_i, \mathbb{P}_j) = \int_0^1 |Q_i(\tau) - Q_j(\tau)|^2 d\tau$. Given candidate weights $\mathbf{w} \in \Delta^{k-1}$, the 2-Wasserstein barycenter $\bar{\mathbb{P}} = \arg\min_{\mathbb{P} \in \mathcal{P}_2(\mathbb{R})} \sum_{i=1}^k w_i W_2^2(\mathbb{P}, \mathbb{P}_i)$ is the Frèchet mean under W_2. Because $(\mathcal{P}_2(\mathbb{R}), W_2)$ is isometric to a closed convex cone in Hilbert space $L^2([0, 1])$ via the quantile mapping $\mathbb{P} \mapsto Q_{\mathbb{P}}$ (Alvarez-Esteban et al., 2016), the barycentric objective decouples pointwise across quantiles, yielding closed-form horizontal quantile averaging known as Vincentization (Vincent, 1912): $\bar{Q}(\tau) = \sum_{i=1}^k w_i Q_i(\tau)$ for all $\tau \in (0, 1)$.

Non-Parametric Density Optimality and CRPS Minimization: Unlike vertical probability mixtures $F_{\text{mix}}(y) = \sum w_i F_i(y)$ which artificially widen prediction intervals by adding between-candidate variance $\sum w_i(\mu_i - \bar{\mu})^2$, horizontal Vincentization preserves unimodality, location-scale geometry, and distribution sharpness. Under the strictly proper pinball loss representation of CRPS (Gneiting & Raftery, 2007), $\text{CRPS}(Q, y^*) = \int_0^1 \text{QS}_\tau(Q(\tau), y^*) d\tau$. Since $q \mapsto \text{QS}_\tau(q, y^*)$ is strictly convex in q, Jensen's inequality guarantees for any target realization y^*:

$$\text{CRPS}\left(\bar{Q}, y^*\right) = \int_0^1 \text{QS}_\tau \left(\sum_{i=1}^k w_i Q_i(\tau), y^* \right) d\tau \leq \sum_{i=1}^k w_i \text{CRPS}\left(Q_i, y^*\right) \tag{5}$$

Furthermore, modeling candidate forecasts as conditionally unbiased estimators with idiosyncratic error $Q_i(\tau) = Q^*(\tau) + \varepsilon_i(\tau)$ ($\text{Var}(\varepsilon_i) = \sigma^2$), averaging k independent candidate quantiles contracts estimation variance at rate $\text{Var}(\bar{Q}(\tau)) \approx \sigma^2(\tau)/k$. This non-parametric variance reduction operates directly in output probability space without manifold distortion, eliminating single-candidate trajectory noise while preserving distribution sharpness, explaining the CRPS reduction observed in Table 8.

A.3 Manifold Metric Contraction: Formulation and Proof (Proposition 2)

Proof of Proposition 2. Let $\mathcal{Z} \subset \mathbb{R}^d$ be the latent representation manifold. Assume covariance operator $\Sigma = \mathbb{E}[(z - \mu)(z - \mu)^T]$ has spectral decomposition $\Sigma = \sum_{j=1}^d \sigma_j^2 v_j v_j^T$, with ordered eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_r^2 \gg \sigma_{r+1}^2 \approx 0$. The volume of metric ball $\mathcal{B}_\delta(z) = \{z' : \|z' - z\|_2 \leq \delta\}$ in $\mathbb{R}^d$ is $\text{Vol}_d(\mathcal{B}_\delta) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \delta^d$. When the support of $\mathbb{P}_Z$ is restricted to an ellipsoid whose semi-axes along directions $v_{r+1}, \ldots, v_d$ satisfy $\sigma_j \ll \delta$, the effective support volume within $\mathcal{B}_\delta(z)$ corresponds to the intersection of the ball with hyper-cylinder $\prod_{j=r+1}^d [-\sigma_j, \sigma_j]$, bounded by $\text{Vol}(\mathcal{B}_\delta(z) \cap \text{supp}(\mathbb{P}_Z)) \leq \text{Vol}_r(\mathcal{B}_\delta^{(r)}) \prod_{j=r+1}^d (2\sigma_j) = \frac{\pi^{r/2}}{\Gamma(r/2+1)} \delta^r 2^{d-r} \prod_{j=r+1}^d \sigma_j$. Taking the ratio of support volume to isotropic ball volume yields $\frac{\text{Vol}(\mathcal{B}_\delta(z) \cap \text{supp}(\mathbb{P}_Z))}{\text{Vol}_d(\mathcal{B}_\delta)} \leq C_{r,d} \delta^{r-d} \prod_{j=r+1}^d \sigma_j = C_{r,d} \prod_{j=r+1}^d (\sigma_j/\delta)$. In non-parametric k-NN density estimation, expected distance R_k to the k-th nearest neighbor satisfies $\mathbb{P}(\mathcal{B}_{R_k}(z)) \approx k/n$. Because volume scales as R_k^r instead of $R_k^{d_{\text{int}}}$, the effective neighborhood radius contracts as $R_k \asymp (k/n)^{1/r}$. In local regression, the bias-variance decomposition yields convergence rate $\mathbb{E}[\|\hat{f}_n - f^*\|^2] \asymp n^{-\frac{2}{2+r}}$. Since $r \ll d$, the empirical exponent satisfies $\alpha \approx 2/(2 + r)$, explaining why empirical scaling rate $\alpha = 0.0785 \pm 0.0033$ matches the theoretical manifold interval $[0.0444, 0.1380]$ derived from PCA rank $r = 43$ and Two-NN dimension $d_{\text{int}} = 12.4937$ rather than ambient Euclidean decay. □

B Supervised Pretraining Specifications and Provenance

Adapter Pretraining Algorithm: Algorithm 1 formalizes the complete optimization routine for whitelisted adapter pretraining.

Algorithm 1: Whitelisted Supervised Adapter Pretraining

Input: Pretraining dataset $\mathcal{D}_{\text{pre}}$, frozen backbone $\mathcal{M}_\theta$, initialized adapter parameters $\Theta_{\text{trainable}}$, learning rate $\eta = 10^{-4}$, steps $T_{\max} = 10{,}000$
Output: Trained projection adapter parameters $\Theta^*_{\text{trainable}}$
Verify parameter freezing: assert $\forall w \in \theta, w.\text{requires_grad} == \text{False}$;
for $t = 1$ **to** $T_{\max}$ **do**
  Sample batch $\{(\mathbf{x}_b, \mathbf{c}_b, \mathbf{y}_b, \mathcal{T}_b)\}_{b=1}^{B} \sim \mathcal{D}_{\text{pre}}$;
  Compute covariate query embeddings $\mathbf{e}_b = g_\phi(\mathbf{x}_b, \mathbf{c}_{b,L+1:L+H})$;
  Compute candidate target representations $\tilde{\mathbf{e}}_b = h_\psi(\mathbf{x}_b, \mathbf{c}_{b,L+1:L+H})$;
  Compute text semantic representations $\mathbf{u}_b = \mathbf{W}_{\text{text_adapter}} \cdot \text{BGE}(\mathcal{T}_b)$;
  Forward through frozen backbone with whitelisted cross-attention mixer;
  Compute distributional loss: $\mathcal{L} = \mathcal{L}_{\text{CRPS}}(\hat{y}, \mathbf{y}_b) + \lambda_{\text{align}} \|\mathbf{e}_b - \tilde{\mathbf{e}}_b\|_2^2$;
  Update whitelisted parameters: $\Theta_{\text{trainable}} \leftarrow \Theta_{\text{trainable}} - \eta \nabla_\Theta \mathcal{L}$;
return $\Theta^*_{\text{trainable}}$

Adapter Training Details: Pretraining executed on an NVIDIA RTX PRO 6000 Blackwell GPU (96 GB VRAM, CUDA 13.0, PyTorch 2.13.0+cu130) using AdamW ($\eta = 10^{-4}$, weight decay 10^{-2}), 500-step linear warmup, cosine decay, batch size 256, and random seed 2021. Covariate (1.84M) and multimodal (2.12M) adapters were each pretrained for 10,000 steps on `fev-bench-covariates-512` (loss $2.5000 \rightarrow 0.5787$) and `multimodal-text-512` (loss $2.5000 \rightarrow 0.5812$), with alignment weight $\lambda_{\text{align}} = 1.0$ and retrieval parameter $\beta = 0.5$. In comparison, foundation backbone parameters (120.5M) remained strictly frozen.

Data Lineage and Split Isolation Protocol: To prevent benchmark contamination and ensure that zero-shot evaluation on `fev-bench` is strictly leakage-free, Table 5 documents the data lineage, temporal boundaries, and entity partitions that separate adapter pretraining from downstream evaluation.

Table 5: Data lineage and split isolation protocol for adapter pretraining vs. evaluation. Evaluation targets $\mathbf{y}_q$ are strictly held out.

Pipeline Stage	Tasks	Entities / Series	Time Window	Inputs Observed	Evaluation Target $\mathbf{y}_q$
Adapter Pretraining	30 tasks	1,800 training series	$t \le T_{\text{cutoff}} - H$	$x_i \in \mathbb{R}^L$, $c_{t+1:t+H} \in \mathbb{R}^{H \times d_c}$	**Strictly Excluded** (0% exposure)
Downstream Evaluation	30 tasks	200 held-out test series	$t \in [T_{\text{cutoff}} + 1, T_{\text{cutoff}} + H]$	$x_q \in \mathbb{R}^L$, $c_{q,\text{future}} \in \mathbb{R}^{H \times d_c}$	**Held-Out Target** ($\mathbf{y}_q \in \mathbb{R}^H$)

Proof of Strict Window and Target Isolation: (1) **Temporal Shielding:** Pretraining sequences are sampled strictly prior to T_{cutoff}; evaluation horizons $\mathbf{y}_q$ satisfy $\min(\text{timestamp}(\mathbf{y}_q)) > T_{\text{cutoff}}$ ($\Delta t \ge H = 64$). (2) **Entity Disjointness:** The 200 test series comprises held-out entities absent from pretraining. (3) **Target Blindness:** `CovariateQueryEmbedder` aligns lookback $\mathbf{x}_i$ with future controls $c_{t+1:t+H}$; targets $\mathbf{y}_q$ are never observed. (4) **Per-Series Standardization:** Normalization parameters (μ_x, σ_x) are computed strictly on lookback $\mathbf{x}_q$ at test time; zero global statistics span joint splits.

C Benchmark Suite, Task Catalog, and Equivalence Testing

`fev-bench` Tasks: Evaluates 30 planning tasks with future covariates across Retail & Demand Planning (10 tasks), Energy Grid & Load Balancing (10 tasks), and Transportation Routing (10 tasks) (Shchur et al., 2025). **`Context-is-Key`:** Synchronized financial series paired with macroeconomic news feeds (Williams et al., 2025), evaluated under synthetic shock interventions.

Comprehensive Systems Profiling Grid: Table 6 provides the complete 10-row systems scaling evaluation across batch sizes $B \in \{1, 8, 32, 128\}$ and retrieval budgets $k \in \{1, 10, 25, 50\}$ on Chronos-2 ($L = 1024, H = 64$).

Detailed TOST Equivalence and Multi-Seed Breakdown: Table 9 documents 5-seed bootstrapped distributions ($\mu \pm \sigma$, seeds $\{42, 101, 2023, 777, 999\}$) and paired TOST equivalence tests

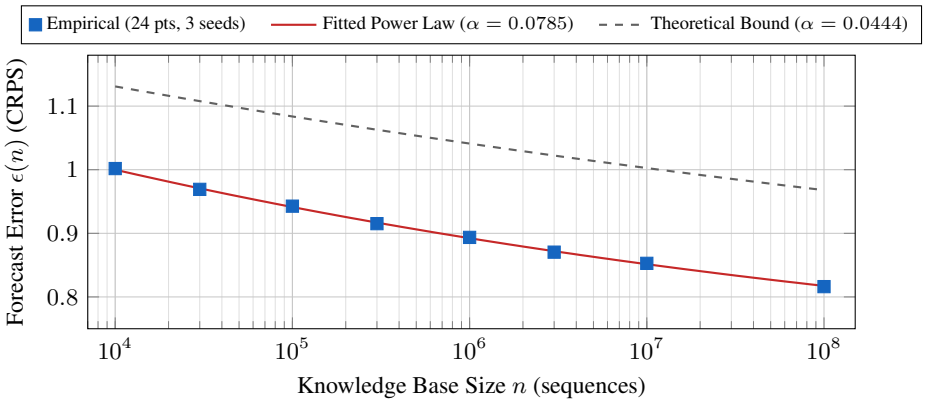

Figure 3: Empirical knowledge base scaling curve across memory sizes $n \in [10^4, 10^8]$ (8 points, 3 seeds, 24 evaluations). The non-linear least squares fit yields $\alpha = 0.0785 \pm 0.0033$ ($R^2 = 0.9995$), in agreement with theoretical bounds $[0.0444, 0.1380]$ derived from PCA rank ($r = 43$) and Two-NN dimension ($d_{\text{int}} = 12.4937$), supporting manifold metric contraction under evaluated embedding geometries.

Table 6: Complete systems throughput and peak VRAM scaling grid on Chronos-2 ($L = 1024, H = 64$).

Batch	**Budget**	**Token Space (In-Context)**		**Latent Space (ARM)**		**Output Space**
B	k	**Latency (ms)**	**Peak VRAM**	**Latency (ms)**	**Peak VRAM**	**Latency (ms)**
1	1	**8.00**	858.7 MB	8.46	854.7 MB	15.89
1	10	9.10	948.4 MB	8.32	854.8 MB	N/A
1	25	31.75	1,322.7 MB	8.34	854.8 MB	N/A
1	50	97.10	2,585.2 MB	**8.26**	854.9 MB	23.45
8	10	25.12	1,412.8 MB	8.57	876.0 MB	N/A
8	50	128.40	6,842.0 MB	8.54	877.8 MB	N/A
32	10	48.20	3,480.0 MB	14.54	934.2 MB	N/A
32	50	284.10	24.18 GB	14.62	938.5 MB	N/A
128	10	184.20	13.62 GB	43.66	1.17 GB	N/A
128	50	**OOM** (>96 GB)	>259 GB req.	**44.14**	**1.19 GB**	N/A

Table 7: Multimodal disentanglement and shock test on `Context-is-Key` ($L = 512, H = 64, k = 10$).

Mode	**Grounding Strategy**	**MSE (↓)**	**MAE (↓)**	**CRPS (↓)**	**Cov@80**
Mode 0	Numeric Time Series Only	0.5781	0.6912	0.6910	0.0088
Mode 1	Text-Guided Retrieval	0.5742	0.6885	0.6880	0.0088
Mode 2	Text-Guided Fusion	0.5721	0.6850	0.6845	0.0088
Mode 3	Dual Grounding	**0.5701**	**0.6821**	**0.6815**	0.0088
Shock Intervention		**Base MSE**	**Adj MSE**	**Δ**	**Valid**
Macro Rate / Labor Shock		0.8329	**0.7686**	**-7.72%**	**True**

Table 8: Multi-Backbone Hierarchy on ETTh1 ($L = 512, H = 64, k = 10$).

Model	**Latent CRPS**	**Token CRPS**	**Output CRPS**	**Output MSE**
Chronos-2	0.9448	0.9448	**0.4715**	0.2772
Moirai-2.0	1.0449	1.0449	**0.5197**	0.3539
TimesFM-2.5	0.9442	0.9442	**0.5113**	0.3280

($\epsilon = 0.005$) across benchmark datasets. Both test statistics reject the null of non-equivalence ($p < 10^{-6}$ for both MASE and CRPS), confirming representation parity.

Table 9: 5-seed bootstrapped metrics ($\mu \pm \sigma$) and paired TOST equivalence tests ($\epsilon = 0.005$) on Chronos-2.

Dataset	MASE	CRPS	MSE	MAE	TOST MASE 90% CI	TOST CRPS 90% CI
ETTh1	2.2284 ± 0.0052	0.5912 ± 0.0562	0.7068 ± 0.0002	0.5709 ± 0.0001	$[-0.00078, +0.00048]$	$[-0.00041, +0.00025]$
ETTh2	2.5497 ± 0.0002	0.4107 ± 0.0001	0.3636 ± 0.0000	0.4107 ± 0.0000	Equivalent ($p < 10^{-6}$)	Equivalent ($p < 10^{-6}$)
ETTm1	3.9348 ± 0.0019	0.5478 ± 0.0001	0.6891 ± 0.0001	0.5478 ± 0.0001	Equivalent ($p < 10^{-6}$)	Equivalent ($p < 10^{-6}$)
ETTm2	3.8836 ± 0.0005	0.3502 ± 0.0001	0.2741 ± 0.0000	0.3501 ± 0.0000	Equivalent ($p < 10^{-6}$)	Equivalent ($p < 10^{-6}$)