Falcon-X: A Time Series Foundation Model for Heterogeneous Multivariate Modeling

Yiding Liu*, **Yifan Hu***, **Hongjie Xia***, **Peiyuan Liu***, **Hongzhou Chen**, **Xilin Dai**, **Zewei Dong**†, **Jiang-Ming Yang**
Ant International
{yiding.lyd, zewei.dong}@ant-intl.com

Abstract

Time series foundation models (TSFMs) are transforming the forecasting paradigm through large-scale cross-domain pretraining. However, most existing TSFMs remain univariate, and recent efforts to enable cross-variate modeling still operate directly within the raw variate space. This design introduces fundamental limitations in semantic alignment and relational expressivity. Specifically, raw-space group mixing lacks a dedicated mechanism to align heterogeneous physical quantities, while standard non-negative attention does not explicitly provide signed aggregation for synergistic and antagonistic interactions. To address these challenges, we propose **Falcon-X**, which decouples variates from the raw space and maps them into a unified latent prototype space. **Falcon-X** employs a Unified Prototype Diff-Attention mechanism that explicitly evaluates both positive and negative semantic affinities to explicitly align heterogeneous variates. Cross-variate interactions are then efficiently performed within this shared space via Latent Entity Attention, naturally facilitating zero-shot structural transfer. Finally, a Variate Reassembly Router robustly reconstructs variate-specific trajectories via a request-and-dispatch mechanism. Extensive evaluations on the GIFT-Eval and fev-bench benchmarks demonstrate that **Falcon-X** achieves state-of-the-art forecasting performance, supporting this architectural approach to heterogeneous multivariate forecasting. An API client and GIFT-Eval evaluation code are available at https://github.com/ant-intl/Falcon-TST.

1 Introduction

Time series forecasting is a fundamental task for understanding dynamic systems and supporting future-oriented decision making. Conventional deep forecasting models are typically trained at the dataset level [1, 2], making them difficult to reuse across domains, sampling frequencies, and variate structures that vary in the real world. Time series foundation models (TSFM) are reshaping this paradigm by pretraining on large-scale cross-domain data and transferring directly to new forecasting tasks [3], thereby substantially reducing the cost of repeated training and tuning. However, most existing models still take univariate series as the basic modeling unit, extrapolating the future solely from each series' own history. This formulation disconnects the co-evolving relationships that are ubiquitous in real systems, limiting the ability of TSFMs to capture complex multivariate dynamics.

To enable cross-variate modeling, recent TSFMs explore how foundation models can accommodate varying numbers of variates. As shown in Tab. 1, Moirai-1.0 [4] flattens multivariate series into a single sequence, enabling joint attention across time and variates. While straightforward, this design scales poorly with the number of variates, as attention cost increases rapidly in high-dimensional settings. A more effective alternative is group attention [5, 6], which organizes variates

*Equal contribution. † Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

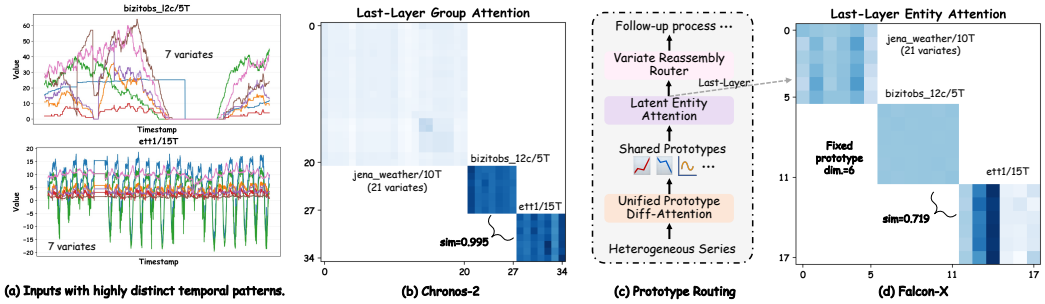

Figure 1: Conceptual comparison of multivariate modeling paradigms. (a) Heterogeneous inputs with highly distinct temporal patterns. (b) Group attention produces almost identical attention maps for completely dissimilar inputs, illustrating limited differentiation in these raw-space attention examples. (c-d) In contrast, **Falcon-X** projects variates into a latent prototype space, yielding highly discriminative attention maps that accurately capture the underlying dynamics.

into dataset-, entity-, or task-level groups and confines attention to variates within the same group. This avoids indiscriminate mixing across unrelated samples while allowing a shared backbone to process multivariate inputs with varying dimensionalities. By replacing global all-to-all interaction with structured within-group communication, group attention marks an important step toward scalable multivariate TSFMs.

However, group attention still exhibits fundamental limitations in ***semantic alignment*** and ***relational expressivity***. ❶ It operates directly in the raw variate space, controlling which variates can interact but not how their heterogeneous semantics are aligned [7]. In high-dimensional systems, only a small subset of variates typically exhibits strong dependencies, while many others are weakly related or noisy. As shown in Fig. 1, Dense attention over raw variates can therefore dilute meaningful signals and promote dataset-specific correlations as if they were transferable patterns. Moreover, different datasets share the same Transformer backbone, yet their variates often correspond to entirely different physical quantities and dynamics. Without a dedicated alignment space, the model must absorb such heterogeneity implicitly in its parameters, making cross-domain transfer a by-product of parameter sharing rather than a principled process of organizing reusable temporal structures. ❷ Existing attention mechanisms have limited relational expressivity. In real-world systems, cross-variate dependencies often involve both synergistic and antagonistic interactions. However, current attention formulations primarily capture aggregative effects, without explicitly assigning negative aggregation weights.

In this paper, we address these limitations by decoupling physical variates from the latent space used for cross-variate interaction. Instead of mixing raw variates, we map them into a shared fixed-dimensional latent prototype space, where interactions are mediated by pairs of learnable prototypes that explicitly capture both positive and negative semantic affinities. These prototypes absorb recurring temporal structures across datasets and filter out variate-specific noise, enabling the model to learn reusable cross-domain patterns in a unified aligned space. This alignment provides a common coordinate system for heterogeneous variates, making knowledge sharing more structured and transferable. It also replaces dense variate-level attention with lightweight variate-to-prototype interaction, improving scalability while preserving cross-variate modeling capacity.

Technically, we propose **Falcon-X**, a novel encoder-only TSFM with 591 million parameters implementing this latent prototype architecture for heterogeneous multivariate forecasting. At its core, the Unified Prototype Diff-Attention decouples heterogeneous variates into a fixed semantic space, utilizing a differential mechanism to explicitly capture both synergistic and antagonistic variate relationships. Once aligned, the Latent Entity Attention performs global cross-variate interactions entirely within this unified space, naturally facilitating seamless cross-domain transfer without coupling to raw variate dimensions. To project back to the original physical space, a dynamic Variate Reassembly Router robustly reconstructs variate-specific trajectories via a *request-and-dispatch* mechanism and gated residual connections. Furthermore, **Falcon-X** integrates essential instance-wise normalization, patching, and a probabilistic forecasting head to ensure end-to-end stability. Extensive experiments on the GIFT-Eval [8] and fev-bench [9] benchmarks validate that **Falcon-X** advances the state-of-the-art in scalable and zero-shot multivariate forecasting.

Table 1: Comparison of capabilities of TSFMs. *Cross Learning*: transferring of universal cross-variate interactive patterns across distinct datasets. *Signed Dependence*: computing both positive and negative affinities to capture synergistic and antagonistic dynamics. *Heterogeneous Unification*: projecting physical variates with varying dimensionalities into a dimension-agnostic latent space.

TSFMs	`Falcon-X`	Moirai-2	TimesFM-2.5	Chronos-2	Timer-S1	Toto	Sundial	Moirai-1.0	TabPFN-TS
Univariate	✓	✓	✓	✓	✓	✓	✓	✓	✓
Multivariate	✓	✗	✗	✓	✗	✓	✗	✓	✗
Cross Learning	✓	✗	✗	✓	✗	✗	✗	✗	✗
Signed Dependence	✓	✗	✗	✗	✗	✗	✗	✗	✗
Heterogeneous Unification	Prototype Routing	✗	✗	Group Mixing	✗	Fixed	✗	Concat.	✗

In a nutshell, our main contribution can be summarized as:

- ***Heterogeneous Modeling Framework.*** We introduce a shared latent prototype interface for multivariate time series foundation models. It maps heterogeneous physical variates into common interaction coordinates, enabling the same architecture to process different variate structures.
- ***Architectural Design.*** We propose `Falcon-X`, an encoder-only foundation model combining Unified Prototype Diff-Attention for signed aggregation, Latent Entity Attention for interaction after alignment, and a gated Variate Reassembly Router for variate-specific reconstruction.
- ***Empirical Evaluation.*** Evaluations on GIFT-Eval and fev-bench demonstrate strong zero-shot forecasting performance. Ablations assess the contributions of the architectural components and training strategies.

2 Related Works

In recent years, the emergence of pre-training techniques has shifted the paradigm of time series forecasting from domain-specific models to the era of foundation models. Early explorations in this domain, such as Time-LLM [10], primarily focused on tuning third-party large language models to adapt time series tasks. Subsequently, with the accumulation of massive time series data, the trend has transitioned toward learning from scratch, where models are pre-trained directly on large-scale time series data to capture inherent temporal dynamics. Chronos [11] frames time series forecasting as a language modeling task by tokenizing real-valued observations into a discrete vocabulary. MOMENT [12] introduces a family of foundation models using a masked prediction task, which allows for generalization across time series tasks through fine-tuning. Inspired by the success of large language models, the generative paradigm built upon decoder-only architectures has become a prominent choice for many mainstream works. These methods typically partition time series into non-overlapping patches, treating each patch as a single token [13], and leverage decoder-only structures to generate one patch at each step [14, 15, 16, 17, 18]. Despite achieving competitive zero-shot performance, these models primarily rely on channel independence, which limits them to univariate forecasting and ignores the rich context of multivariate dependencies within time series data.

Despite these advances, a key challenge of multivariate forecasting remains the unified modeling of heterogeneous time series. Moirai-1.0 [4] flattens variates into a single sequence to capture joint interactions. Toto [5] introduces proportional factorized space-time attention to efficiently model cross-variate dependency. Furthermore, Chronos-2 [6] further adopts group attention, facilitating in-context learning by sharing information across related series within flexible groups. However, these approaches still operate directly within the raw variate space, fundamentally limiting their semantic alignment and relational expressivity. Specifically, they primarily capture aggregative effects and struggle to represent the antagonistic dynamics ubiquitous in real-world physical systems. To bridge this gap, our `Falcon-X` introduces a differential mechanism within an explicitly aligned latent prototype space, enabling the foundation model to systematically capture complex dual dynamics and facilitate robust cross-domain transfer.

3 `Falcon-X`

In this section, we present the architecture of `Falcon-X`. As shown in Fig. 2, `Falcon-X` consists of four parts: pre-process of Normalization and Tokenization, Time Attention, Variate Attention and

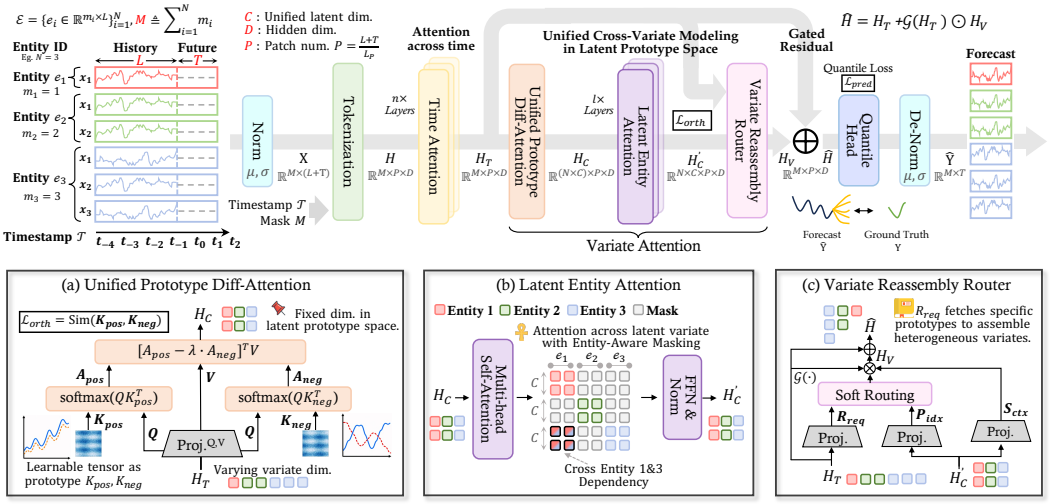

Figure 2: The overall architecture of **Falcon-X**. The raw inputs are normalized, tokenized, and processed by Time Attention to extract independent temporal features. The Unified Prototype Diff-Attention (UPDA) then projects these features into a shared prototype space, enabling Latent Entity Attention (LEA) to capture global cross-variate dependencies explicitly. The Variate Reassembly Router (VRR) then dynamically reconstructs variate-specific representations. These are fused with the temporal context and fed into a quantile head to generate the final probabilistic forecasts.

Forecasting Head. To maintain a concise exposition of the mathematical formulations, we present extended discussions of our *underlying design philosophies* in App. A.

3.1 Problem Formulation

Let $\mathcal{E} = \{\mathbf{e}_i \in \mathbb{R}^{m_i \times L}\}_{i=1}^{N}$ be a collection of N entities, where each entity $\mathbf{e}_i$ represents a multivariate time series with a specific variate dimensionality m_i and a historical look-back window L. A key challenge in TSFM is the heterogeneity of these dimensions, with m_i varying significantly across diverse entities and datasets. We define the total aggregated dimensionality M across all entities as $\sum_{i=1}^{N} m_i$. The objective of **Falcon-X** is to learn a dimension-agnostic mapping $\mathcal{F}_\theta$, parameterized by θ, that transforms the heterogeneous input space into the target predictive space:

$$\hat{\mathbf{Y}} = \mathcal{F}_\theta(\mathbf{X}), \quad \text{where } \mathbf{X} \in \mathbb{R}^{M \times L}, \hat{\mathbf{Y}} \in \mathbb{R}^{M \times T}. \tag{1}$$

3.2 Normalization and Tokenization

Normalization. We formulate the forecasting task as a unified masked reconstruction paradigm. Given the raw input $\mathbf{X} \in \mathbb{R}^{M \times (L+T)}$ spanning the historical window L and future horizon T, we first replace the target future steps with placeholder tokens. To accommodate different units and magnitudes across domains, we apply instance standardization followed by an inverse hyperbolic sine transformation, as in Chronos-2 [6]:

$$\hat{\mathbf{X}} = \operatorname{asinh}\left(\frac{\mathbf{X} - \mu}{\sigma}\right). \tag{2}$$

Crucially, instead of zero-filling or truncating sequences at missing positions [19], the instance-wise mean μ and standard deviation σ are computed exclusively from observed values, preserving missing entries for explicit downstream modeling. For non-degenerate series, standardization makes the representation invariant to positive rescaling and shifts, $X' = aX + b$ with $a > 0$. The subsequent asinh is approximately linear near zero and compresses large standardized magnitudes while remaining monotone and invertible. RevIN [20] also removes and restores instance statistics, but does not include this tail-compressing transform; a controlled comparison is provided in App. D.

Tokenization. To obtain robust representations, we augment $\hat{\mathbf{X}}$ by concatenating it with relative timestamps $\mathcal{T}$ and a binary observation mask $\mathcal{M}$. To generalize across heterogeneous sampling frequencies, $\mathcal{T} = \{-\frac{L}{L+T}, \ldots, 0, \ldots, \frac{T-1}{L+T}\}$ injects a normalized sequential ordering anchored at 0 for the first forecasting step: Meanwhile, the explicit inclusion of $\mathcal{M}$ enables the model to dynamically distinguish genuine observations from missing entries or masked future targets. Subsequently, the augmented sequence is partitioned into $P = \frac{L+T}{L_p}$ non-overlapping patches of length L_p and projected into the hidden dimension D via a residual patch embedding mechanism:

$$\mathbf{H} = \text{ResPatchEmbed}(\text{Concat}(\hat{\mathbf{X}}, \mathcal{T}, \mathcal{M})) \in \mathbb{R}^{M \times P \times D}. \tag{3}$$

Unlike standard linear projections, the residual patch embedding is designed to seamlessly integrate linear properties with the complex non-linear semantics extracted from the augmented inputs.

3.3 Time Attention

To capture the intrinsic evolutionary patterns of each individual variate, `Falcon-X` utilizes an encoder-only Transformer architecture (see App. A.2). Specifically, the Time Attention module consists of n identical encoder layers, each applied independently along the patch dimension P for all M variates. Let $\mathbf{H}^{(0)} = \mathbf{H}$ be the input from the Tokenization layer. For each layer $i \in \{1, \ldots, n\}$, the hidden state $\mathbf{H}^{(i)}$ is computed as follows:

$$\mathbf{H}^{(i)} = \text{LayerNorm}\left(\mathbf{H}^{(i-1)} + \text{MHA}\left(\mathbf{H}^{(i-1)}\right)\right), \tag{4}$$

where MHA denotes the multi-head attention mechanism [21]. We apply LayerNorm [22] to every layer to stabilize the internal activations and facilitate the training of deep foundation architectures. After n successive transformations, the final output of this module is denoted as $\mathbf{H}^{(n)} = \mathbf{H}_T \in \mathbb{R}^{M \times P \times D}$, encapsulating the comprehensive temporal dynamics of each variate.

3.4 Variate Attention

A core requirement of time-series foundation models is ***the ability to transcend rigid, dataset-specific dimensional constraints and learn a unified representation of dependencies across multivariate series***. However, treating different physical variates homogeneously or simply concatenating them leads to severe semantic misalignment. To overcome this, `Falcon-X` introduces a unified latent space architecture. As shown in Fig. 2(a–c), these modules progressively align the heterogeneous patch embeddings $\mathbf{H}_T$ into a shared prototype space, model both intra- and cross-dataset dependencies, and dynamically reassemble the global context back to the original variate dimensions.

3.4.1 Unified Prototype Diff-Attention

To overcome the semantic misalignment inherent in raw-space interactions (detailed discussion in App. A.3.1), `Falcon-X` projects the heterogeneous temporal embeddings H_T into a fixed-C-dimensional latent prototype space. We employ a Differential Attention mechanism to explicitly capture both positive and negative semantic affinities.

Concretely, we define two globally shared learnable parameter matrices, $\mathbf{K}_{\text{pos}}, \mathbf{K}_{\text{neg}} \in \mathbb{R}^{C \times D}$, representing the synergistic and antagonistic temporal prototypes, respectively. For each entity $\mathbf{e}_i$, given its temporal embeddings $\mathbf{h}_T^i \in \mathbb{R}^{m_i \times P \times D}$, we generate the query $\mathbf{Q}^i$ and value $\mathbf{V}^i$ via linear projections. The dual-dependency attention maps, quantifying the affinity between the m_i heterogeneous variates of the i-th entity and the C unified prototypes, are computed as:

$$\mathbf{A}_{\text{pos}}^i = \text{softmax}\left(\frac{\mathbf{Q}^i \mathbf{K}_{\text{pos}}^\top}{\sqrt{D}}\right), \quad \mathbf{A}_{\text{neg}}^i = \text{softmax}\left(\frac{\mathbf{Q}^i \mathbf{K}_{\text{neg}}^\top}{\sqrt{D}}\right), \tag{5}$$

where $\mathbf{A}_{\text{pos}}^i, \mathbf{A}_{\text{neg}}^i \in \mathbb{R}^{m_i \times P \times C}$. The unified representation $\mathbf{h}_C^i \in \mathbb{R}^{C \times P \times D}$ for entity e_i is then derived by aggregating the value features based on the differential attention score:

$$\mathbf{h}_C^i = \left[\mathbf{A}_{\text{pos}}^i - \lambda \cdot \mathbf{A}_{\text{neg}}^i\right]^\top \mathbf{V}^i, \tag{6}$$

where λ is a learnable scaling factor and $\mathbf{V}^i = \text{Linear}(\mathbf{h}_T^i)$. Notably, while this projection is logically defined at the individual entity level, we implement it concurrently across all N entities to

maximize computational efficiency. By employing an entity-aware masking strategy, we completely bypass inefficient explicit loops, seamlessly and parallelly transforming the heterogeneous inputs into a perfectly unified latent space $\mathbf{H}_C \in \mathbb{R}^{N\times C\times P\times D}$. To ensure semantic distinctiveness, we apply an orthogonality loss $\mathcal{L}_{\text{orth}} = \text{Sim}(\mathbf{K}_{\text{pos}}, \mathbf{K}_{\text{neg}})$ to constrain the relationship between positive and negative prototypes.

3.4.2 Latent Entity Attention

With all representations $\mathbf{H}_C \in \mathbb{R}^{(N\times C)\times P\times D}$ aligned into a shared, dimension-agnostic semantic space, Latent Entity Attention naturally facilitates cross-learning, leveraging shared structural patterns across entirely different domains (see App. A.3.2). To be specific, we treat the combined entity and prototype dimensions as the spatial sequence for interaction, and then apply l layers of the standard MHA mechanism [21] to capture the global cross-variate dependencies:

$$\mathbf{H}'_C = \text{LayerNorm}\left(\mathbf{H}_C + \text{MHA}(\mathbf{H}_C)\right), \tag{7}$$

where $\mathbf{H}'_C \in \mathbb{R}^{N\times C\times P\times D}$ denotes the refined global context matrix. Similar to the Time Attention, this straightforward yet highly effective operation utilizes residual connections and layer normalization to ensure stable representations. By allowing all aligned entities to interact fully within this latent space, $\mathbf{H}'_C$ successfully captures the holistic dynamics necessary for accurate forecasting. UPDA constructs the shared prototype interface, whereas LEA refines interactions after this mapping; the contribution lies in this organization with UPDA and VRR, not in a new MHA operator.

3.4.3 Variate Reassembly Router

To accurately reconstruct variate-specific trajectories, `Falcon-X` orchestrates a targeted retrieval from the unified prototype space back to individual physical dimensions (m_i) via a ***request-and-dispatch*** mechanism (see App. A.3.3). Formally, for each entity $\mathbf{e}_i$, we generate the routing components through independent linear projections based on their respective source tensors:

$$\mathbf{R}^i_{\text{req}} = \text{Linear}(\mathbf{h}^i_T), \quad \mathbf{P}^i_{\text{idx}} = \text{Linear}(\mathbf{h}'^i_C), \quad \mathbf{S}^i_{\text{ctx}} = \text{Linear}(\mathbf{h}'^i_C). \tag{8}$$

Here, the Routing Request ($\mathbf{R}^i_{req}$) acts as an *entity identity tag* conveying the unique temporal trajectory of the original variate. It is matched against the Prototype Index ($\mathbf{P}^i_{idx}$), an addressable map of the global prototype library, to selectively retrieve refined semantic payloads from the Source Context ($\mathbf{S}^i_{ctx}$). The reconstruction is then executed via a scaled dot-product soft-routing operation:

$$\mathbf{h}^i_V = \text{softmax}\left(\frac{\mathbf{R}^i_{\text{req}}(\mathbf{P}^i_{\text{idx}})^\top}{\sqrt{D}}\right)\mathbf{S}^i_{\text{ctx}}. \tag{9}$$

This retrieval paradigm successfully reconstructs variate-specific $\mathbf{h}^i_V \in \mathbb{R}^{m_i\times P\times D}$ with high fidelity, smoothly restoring the physical dimensionality to yield $\mathbf{H}_V \in \mathbb{R}^{M\times P\times D}$. Similar to the initial prototype projection, we apply an entity-aware masking strategy during routing, enabling concurrent soft routing across all N entities without explicit loops.

Finally, to maintain robust cross-dataset performance regardless of varying dependency strengths, we introduce an explicit gated residual connection to dynamically fuse the temporal embeddings $\mathbf{H}_T$ with the cross-variate representations $\mathbf{H}_V$. The final representation $\hat{\mathbf{H}} \in \mathbb{R}^{M\times P\times D}$ is computed as:

$$\hat{\mathbf{H}} = \mathbf{H}_T + \mathcal{G}(\mathbf{H}_T) \odot \mathbf{H}_V, \tag{10}$$

where $\mathcal{G}(\cdot)$ is a gating mechanism with a linear projection followed by a sigmoid activation, and $\odot$ is element-wise multiplication. Thus, `Falcon-X` effectively prevents semantic interference in weakly correlated systems while making full use of cross-variate dependencies in strongly correlated ones.

3.5 Forecasting Head

Following Chronos-2 [6] and Timer-S1 [23], `Falcon-X` adopts a probabilistic forecasting paradigm. Given the reconstructed representation $\hat{\mathbf{H}}$, we extract the embeddings corresponding to the masked future horizon and apply a linear projection to generate forecasts across a predefined set of quantiles $\mathcal{Q}$. The model is end-to-end optimized using the standard Quantile Loss:

$$\mathcal{L}_{\text{pred}} = \frac{1}{|\mathcal{Q}|\cdot M\cdot T}\sum_{q\in\mathcal{Q}}\sum_{i=1}^{M}\sum_{t=1}^{T}\max\left(q(\mathbf{Y}_{i,t} - \hat{\mathbf{Y}}^{(q)}_{i,t}), (q-1)(\mathbf{Y}_{i,t} - \hat{\mathbf{Y}}^{(q)}_{i,t})\right), \tag{11}$$

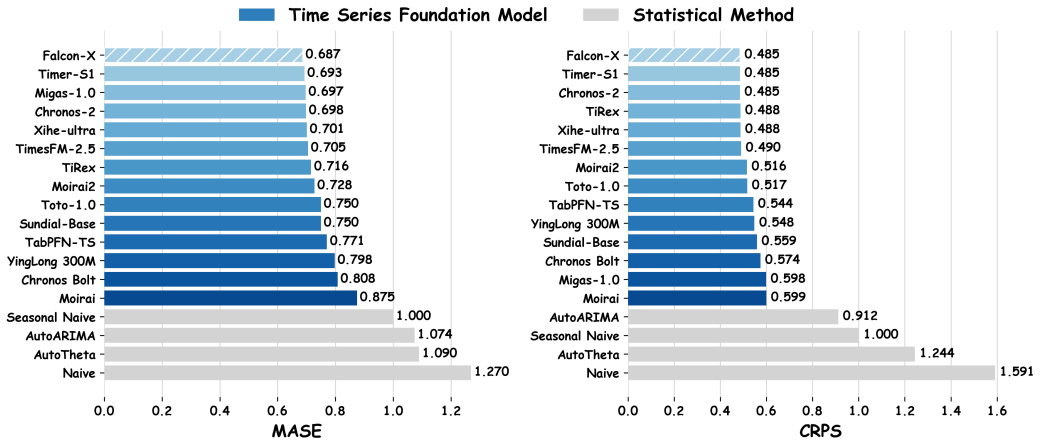

Figure 3: Performance of `Falcon-X` on the GIFT-Eval leaderboard.

where $\mathbf{Y}_{i,t}$ represents the ground truth and $\hat{\mathbf{Y}}_{i,t}^{(q)}$ denotes the model's prediction at the q-th quantile. The overall training objective combines this with the prototype orthogonality loss: $\mathcal{L} = \mathcal{L}_{\text{pred}} + \alpha\mathcal{L}_{\text{orth}}$, where α is a hyper-parameter balancing the forecasting and orthogonality objectives.

During inference, the predictions are mapped back to their original physical scale $\tilde{\mathbf{Y}}^{(q)}$ via a straightforward de-normalization process, applying the hyperbolic sine transformation followed by DeNorm using the preserved instance-wise statistics.

$$\tilde{\mathbf{Y}}^{(q)} = \sigma \cdot \sinh(\hat{\mathbf{Y}}^{(q)}) + \mu. \tag{12}$$

4 Training Details

4.1 Pre-Training Corpus

Our pre-training corpus combines large-scale real-world and synthetic time series data spanning several domains. In addition to public datasets from GIFT-EVAL [8] , CHRONOS [11] and QUITOBENCH [24], it includes synthetic univariate and multivariate series designed to increase diversity in temporal patterns and dependency structures. Specifically, synthetic univariate data are generated through data mixing and stochastic process sampling, while multivariate series are constructed by grouping related univariate signals and injecting explicit cross-variate dependencies, including both instantaneous and temporal interactions. The details can be found in App. B.1.

4.2 Training Infrastructure and Config

To enable scalable pretraining on massive, heterogeneous time-series corpora, we build `Falcon-X` upon Megatron-LM [25] and design a custom sampling pipeline to balance data distribution across diverse domains [26]. Furthermore, to address the heterogeneity in variate dimensionality, we implement a runtime multivariate sampling strategy that dynamically balances the number of consuming variates per batch, thereby improving GPU utilization and training stability. Further implementation details are provided in App. C.

`Falcon-X` features a hidden dimension of $D = 1024$, a patch length of $L_p = 16$, and utilizes $n = 16$ Time Attention layers alongside $l = 16$ Entity Attention layers (16 heads per layer). By accommodating up to 512 input tokens and 30 output tokens, it achieves a maximum context length of $L = 8192$ and a prediction length of $T = 480$ in a single inference pass. The model is pre-trained on a cluster of NVIDIA B200-180GB GPUs for one million iterations with a global batch size of 384 using `bf16` precision, optimized by a joint quantile and orthogonality loss. We adopt the AdamW optimizer [27] and weight decay regularization. The learning rate warms up linearly to 6×10^{-5} over the first 0.1% of steps, followed by a cosine decay to 6×10^{-6}. As shown in Fig. 4, the training process is highly stable, with the loss curve converging smoothly and robustly.

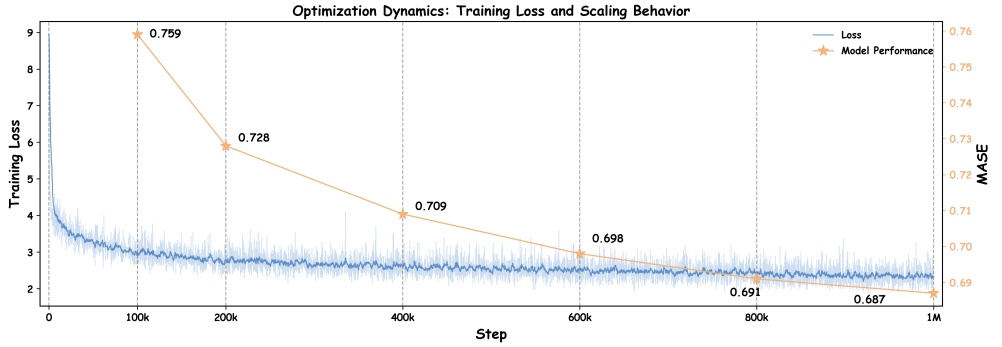

Figure 4: Stable training dynamics and performance scaling over one million iterations.

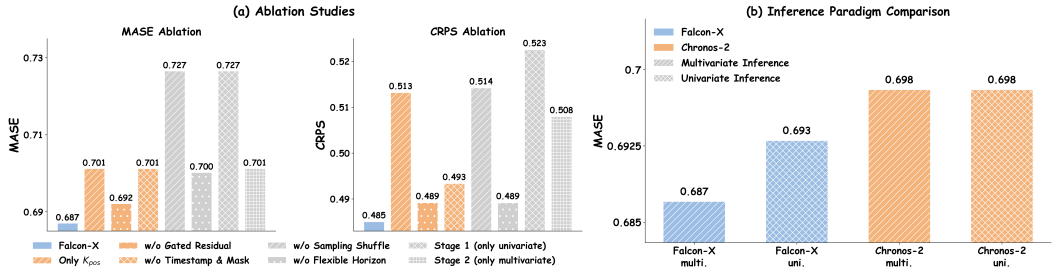

Figure 5: Comprehensive analysis of **Falcon-X**. (a) Ablation studies validating the necessity of key architectural components and training strategies. (b) Inference paradigm comparison, highlighting our robust multivariate modeling against Chronos-2.

5 Experiments

5.1 Main Results

We evaluate **Falcon-X** on the comprehensive GIFT-Eval [8] and fev-bench [9] benchmarks (Fig. 3 and App. F), reporting MASE and CRPS for point and probabilistic forecasting. On GIFT-Eval, **Falcon-X** achieves the lowest MASE and joint-lowest CRPS among the compared models. Relative to Timer-S1 [23], MASE decreases from 0.693 to 0.687 (approximately 0.9%), while CRPS is tied at 0.485 at the reported precision. **Falcon-X** also improves both metrics over Toto [5] and achieves lower MASE than Chronos-2 [6], whose CRPS is likewise 0.485. On fev-bench, **Falcon-X** ranks second among the compared models in both MASE (0.689) and CRPS (0.513), behind Chronos-2 [6] (0.645 and 0.485, respectively). Chronos-2 uses additional past-only and future-known covariates, whereas **Falcon-X** uses only endogenous target series; these results therefore compare different input settings.

5.2 Ablation Studies

As shown in Fig. 5(a), to isolate the contributions of the architecture and optimization pipeline, we conduct comprehensive ablation studies on both model components and training strategies. Additional controlled ablations are reported in App. D.

Module Ablation. We first evaluate the structural designs of **Falcon-X**. (i) **Only $\mathbf{K}_{\text{pos}}$:** Removing the negative prototype key ($\mathbf{K}_{\text{neg}}$) degrades performance, although this removal changes both the signed mechanism and branch capacity. (ii) **w/o gated residual:** Removing the gated residual connection degrades performance, showing its importance in context-aware filtering cross-variate information in various datasets. (iii) **w/o timestamp & mask:** Excluding the relative timestamp $\mathcal{T}$ and mask $\mathcal{M}$ reduces robustness to irregular sampling and missing values.

Strategy Ablation. We further analyze the impact of our data processing and training pipeline. (i) **w/o sampling shuffle:** Disabling variate shuffling significantly hurts performance, indicating that random permutation is crucial for learning content-driven rather than index-dependent relationships.

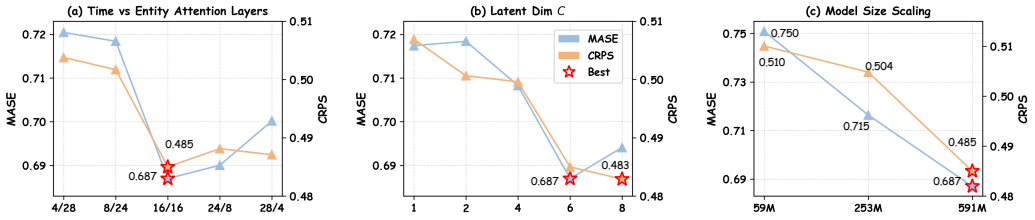

Figure 6: Sensitivity and scaling analysis. (a) Performance impact of layer allocation between Time Attention n and Entity Attention l. (b) Sensitivity to the latent prototype dimension C. (c) Consistent performance scaling across increasing model parameter sizes (from 59M to 591M).

(ii) **w/o flexible horizon:** Replacing flexible horizon sampling with fixed-length prediction weakens generalization across unseen forecasting horizons. (iii) **Two-stage vs. Joint training:** We compare direct joint training with a two-stage curriculum consisting of: (Stage 1) univariate pre-training for temporal modeling initialization, and (Stage 2) multivariate fine-tuning with univariate replay. Joint training consistently performs better, suggesting that `Falcon-X` can naturally learn unified temporal and cross-variate representations within a single optimization process.

5.3 Inference Setting Analysis

We compare inference paradigms against Chronos-2 in Fig. 5(b). On GIFT-Eval, Chronos-2 reports the same aggregate MASE (0.698) with and without group attention. For `Falcon-X`, multivariate inference lowers aggregate MASE from 0.693 to 0.687, supporting the benefit of cross-variate modeling in this evaluation. Equal aggregate scores alone do not establish semantic collapse, and the figure does not separately report performance on univariate tasks.

5.4 Influence of Key Parameter

We analyze the sensitivity of `Falcon-X` to two key architectural parameters.

Depth Distribution (n vs. l). We investigate the layer allocation between Time Attention (n) and Latent Entity Attention (l) under a fixed depth budget. As shown in Fig. 6(a), allocating sufficient capacity to temporal modeling is essential, while excessive cross-variate routing significantly degrades performance. This indicates that robust temporal modeling is the foundation of forecasting, while cross-variate interaction provides complementary gains. `Falcon-X` achieves the best trade-off with a balanced $16/16$ configuration.

Latent Prototype Dimension (C). We evaluate the representational capacity of the unified semantic space by varying C. Fig. 6(b) demonstrates that a severely restricted dimension ($C \leq 2$) induces an information bottleneck, leading to semantic over-compression. Conversely, expanding the dimension enhances relational expressivity. The lowest MASE (0.687) occurs at $C = 6$, while the lowest CRPS (0.483) occurs at $C = 8$, indicating a trade-off between the two metrics.

5.5 Scaling Analysis

We evaluate scaling across training iterations and parameter sizes. In Fig. 4, training loss decreases and MASE improves from 0.759 at 100,000 steps to 0.687 at 1,000,000 steps. Scaling from 59M ($l = n = 8, D = 512$) to 253M ($l = n = 12, D = 768$) and 591M parameters reduces MASE from 0.750 to 0.715 and 0.687, and CRPS from 0.510 to 0.504 and 0.485 (Fig. 6(c)). These results demonstrate improvements over the tested model sizes and training budgets; they do not establish a scaling law beyond this range.

5.6 Case Study

To qualitatively analyze `Falcon-X`, we present representative forecasting cases from GIFT-Eval.

Multivariate vs. Univariate Inference. We compare multivariate inference with channel-independent inference on highly correlated sequences from ETT1/15T (see Fig. 7). Without cross-variate interactions, the channel-independent setting gradually drifts from the ground truth under

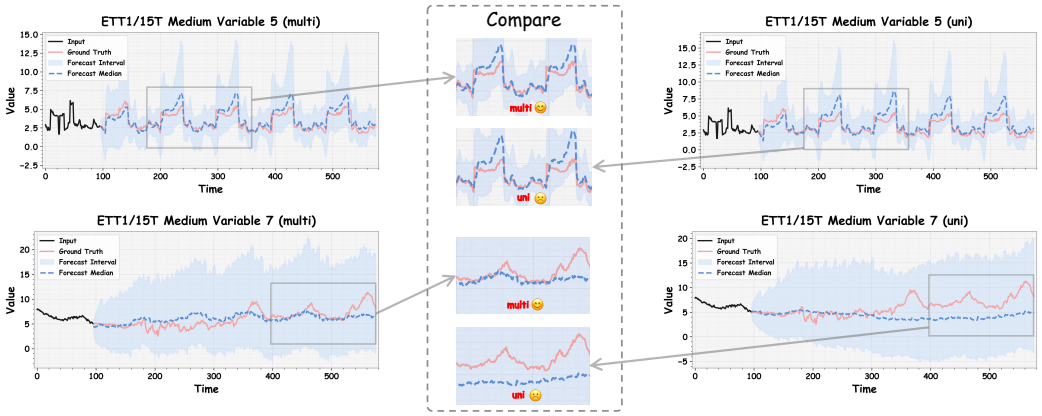

Figure 7: Case study on ETT1/15T dataset. In univariate inference mode, forecasts gradually deviate from the ground truth due to the absence of global context. In contrast, **Falcon-X**'s multivariate inference effectively leverages cross-variate signals to calibrate trajectories.

complex temporal shifts. In contrast, **Falcon-X** leverages its unified latent prototype space to aggregate complementary signals across variates, producing substantially more accurate trajectories.

Positive and Negative Dependency Modeling. We further examine the ability of **Falcon-X** to capture dual dependencies. As shown in Fig. 13, **Falcon-X** accurately models positive correlations with synchronized trends. More importantly, in Fig. 14, it successfully captures negative correlations with opposing dynamics. These results validate the effectiveness of our Unified Prototype Diff-Attention for modeling both synergistic and antagonistic relationships.

6 Limitations

Falcon-X is evaluated on zero-shot forecasting using endogenous time series; broader adaptation, multimodal inputs, and integration with agentic systems [28, 29] remain future work. The learned prototypes are latent interaction coordinates, and their interpretability, resistance to collapse, and robustness under substantial distribution and frequency shifts require further validation. Partitioned inference may miss cross-group dependencies, while temporal and latent entity attention retain quadratic costs despite the compact prototype interface (see App. E).

7 Broader Impacts

Time series forecasting supports decision-making in domains such as energy [30], finance [31, 32], healthcare [33], transportation [34], and cloud infrastructure. By introducing **Falcon-X**, we could help reduce the need for task-specific model design and make heterogeneous multivariate forecasting more scalable. However, forecasting models can still fail under distribution shifts or noisy data, especially in high-stakes scenarios. We therefore recommend careful validation, uncertainty-aware evaluation, and domain expert oversight before deployment, alongside continuous monitoring to detect performance degradation and limit the propagation of forecasting errors.

8 Conclusion

In this paper, we identify two fundamental challenges of existing multivariate time series foundation models: semantic alignment and relational expressivity. To address these issues, we propose **Falcon-X**, an architectural framework that decouples physical variables into a unified latent prototype space. By introducing the Unified Prototype Diff-Attention, our architecture effectively captures both synergistic and antagonistic correlations. Additionally, a Variate Reassembly Router ensures robust global context fusion across diverse domains. Evaluations on GIFT-Eval and fev-bench demonstrate strong zero-shot forecasting performance, while the scaling experiments show improvements over the tested model sizes and training budgets. We hope that this work will contribute to the development of more unified and expressive foundation models for time series.

References

[1] Yifan Hu, Peiyuan Liu, Peng Zhu, Dawei Cheng, and Tao Dai. Adaptive multi-scale decomposition framework for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17359–17367, 2025.

[2] Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. TimeFilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *International Conference on Machine Learning*, 2025.

[3] Siva Rama Krishna Kottapalli, Karthik Hubli, Sandeep Chandrashekhara, Garima Jain, Sunayana Hubli, Gayathri Botla, and Ramesh Doddaiah. Foundation models for time series: A survey. *arXiv preprint arXiv:2504.04011*, 2025.

[4] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, 2024.

[5] Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, et al. This time is different: An observability perspective on time series foundation models. In *Neural Information Processing Systems*, 2025.

[6] Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

[7] Yifan Hu, Jie Yang, Tian Zhou, Peiyuan Liu, Yujin Tang, Rong Jin, and Liang Sun. Bridging past and future: Distribution-aware alignment for time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026.

[8] 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.

[9] 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.

[10] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2023.

[11] 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. *Transactions on Machine Learning Research*, 2024.

[12] 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*, 2024.

[13] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[14] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: generative pre-trained transformers are large time series models. In *International Conference on Machine Learning*, pages 32369–32399, 2024.

[15] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, pages 10148–10167, 2024.

[16] 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.

[17] Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *International Conference on Machine Learning*, pages 39295–39317. PMLR, 2025.

[18] Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. In *Neural Information Processing Systems*, 2025.

[19] Shi Xiaoming, Wang Shiyu, Nie Yuqi, Li Dianqi, Ye Zhou, Wen Qingsong, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025.

[20] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[22] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. In *arXiv preprint arXiv:1607.06450*, 2016.

[23] Yong Liu, Xingjian Su, Shiyu Wang, Haoran Zhang, Haixuan Liu, Yuxuan Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Timer-S1: A billion-scale time series foundation model with serial scaling. *arXiv preprint arXiv:2603.04791*, 2026.

[24] Siqiao Xue, Zhaoyang Zhu, Wei Zhang, Rongyao Cai, Rui Wang, Yixiang Mu, Fan Zhou, Jianguo Li, Peng Di, and Hang Yu. QuitoBench: A high-quality open time series forecasting benchmark. *arXiv preprint arXiv:2603.26017*, 2026.

[25] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[26] Hongjie Xia, Yiding Liu, Yifan Hu, Peiyuan Liu, and Zewei Dong. Into the ORBIT for time series: Training regimes for foundation models. *arXiv preprint arXiv:2608.13262*, 2026.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[28] Yifan Hu, Jie Yang, Xilin Dai, Wanxu Cai, Kuiye Ding, Yuante Li, Qinghua Liu, Enze Ma, Zhiyuan Qu, Yixin Wang, Binyan Xu, Kexin Zhang, Peiyuan Liu, Zhijian Xu, Guibin Zhang, Yujin Tang, Yanwei Yue, Kening Zheng, Chengze Li, Hanrong Zhang, Haoyan Xu, Naiqi Li, Tao Dai, Dawei Cheng, John Paparrizos, Kaize Ding, Tian Zhou, Qiang Xu, Shu-tao Xia, Shirui Pan, and Philip S. Yu. The landscape of agentic time series systems: Architectures, reliability, and frontiers. Preprint, 2026.

[29] Yifan Hu, Xilin Dai, Zhiyuan Qu, Yiding Liu, Zewei Dong, Jiang-ming Yang, and Qiang Xu. When tomorrow becomes today: Self-evolving policies for agentic time-series forecasting. *arXiv preprint arXiv:2609.24862*, 2026.

[30] Yifan Hu, Hongzhou Chen, Peiyuan Liu, Yiding Liu, Zewei Dong, and Jiang-Ming Yang. Existence precedes value: Joint modeling of observational existence and evolving states in time series forecasting. *arXiv preprint arXiv:2606.13571*, 2026.

[31] Yifan Hu, Peiyuan Liu, Yuante Li, Dawei Cheng, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. Finmamba: Market-aware graph enhanced multi-level mamba for stock movement prediction. *arXiv preprint arXiv:2502.06707*, 2025.

[32] Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shutao Xia, Dawei Cheng, and Changjun Jiang. A comprehensive and practical benchmark for financial time series forecasting. *Frontiers of Computer Science*, 20(10):2010629, 2026.

[33] Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A. Pickett, and Varun Dutt. Ai in healthcare: Time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in Big Data*, Volume 3 - 2020, 2020.

[34] Zijian Zhang, Yujie Sun, Zepu Wang, Yuqi Nie, Xiaobo Ma, Ruolin Li, Peng Sun, and Xuegang Ban. Large language models for mobility analysis in transportation systems: A survey on forecasting tasks. *Transportation Research Record*, 2680(2):756–774, 2026.

[35] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. In *International Conference on Learning Representations*, 2025.

[36] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 2018.

[37] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Neural Information Processing Systems*, 2021.

[38] Santosh Palaskar, Vijay Ekambaram, Arindam Jati, Neelamadhav Gantayat, Avirup Saha, Seema Nagar, Nam Nguyen, Pankaj Dayama, Renuka Sindhgatta, Prateeti Mohapatra, Harshit Kumar, Jayant Kalagnanam, Nandyala Hemachandra, and Narayan Rangaraj. Automixer for improved multivariate time-series forecasting on business and it observability data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:22962–22968, 2024.

[39] Siqi Shen, Vincent Van Beek, and Alexandru Iosup. Statistical characterization of business-critical workloads hosted in cloud datacenters. In *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pages 465–474. IEEE, 2015.

[40] Addison Howard, Haruka Yui, Mark McDonald, and Will Cukierski. Recruit restaurant visitor forecasting. https://kaggle.com/competitions/recruit-restaurant-visitor-forecasting, 2017. Kaggle.

[41] Paolo Mancuso, Veronica Piccialli, and Antonio M Sudoso. A machine learning approach for forecasting hierarchical time series. *Expert Systems with Applications*, 182:115102, 2021.

[42] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[43] Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C58C86.

[44] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2017.

[45] Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chengkai Han, and Wayne Xin Zhao. Towards efficient and comprehensive urban spatial-temporal prediction: A unified library and performance benchmark. *arXiv preprint arXiv:2304.14343*, 2023.

[46] Rakshitha Wathsadini Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *The Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[47] Jiawei Jiang, Chengkai Han, Wenjun Jiang, Wayne Xin Zhao, and Jingyuan Wang. Libcity: A unified library towards efficient and comprehensive urban spatial-temporal prediction. *arXiv preprint arXiv:2304.14343*, 2023.

[48] Kamiar Mohaddes and Mehdi Raissi. Compilation, revision and updating of the global var (gvar) database. Mendeley Data, Version 1, 2024.

[49] Ines Wilms and Christophe Croux. Forecasting using sparse cointegration. *International Journal of Forecasting*, 32(4):1256–1267, 2016.

[50] George Athanasopoulos, Roman A. Ahmed, and Rob J. Hyndman. Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, 25(1):146–166, January 2009.

[51] Michael W. McCracken and Serena Ng. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

[52] Michael W. McCracken and Serena Ng. FRED-QD: A quarterly database for macroeconomic research. *Review*, 103(1):1–44, January 2021.

[53] Ulrik Thyge Pedersen. CO2 emissions by country. https://www.kaggle.com/datasets/ulrikthygepedersen/co2-emissions-by-country, 2025. Kaggle.

[54] Nafay Un Noor. Global life expectancy data (1950–2023). https://www.kaggle.com/datasets/nafayunnoor/global-life-expectancy-data-1950-2023, 2025. Kaggle.

[55] Bushra Qurban. Tourism and economic impact. https://www.kaggle.com/datasets/bushraqurban/tourism-and-economic-impact, 2025. Kaggle.

[56] Lawrence J. Christiano, Martin Eichenbaum, and Charles L. Evans. Monetary policy shocks: What have we learned and to what end? In *Handbook of Macroeconomics*, volume 1 of *Handbook of Macroeconomics*, pages 65–148. Elsevier, 1999.

[57] Philip J Fleming and John J Wallace. How not to lie with statistics: the correct way to summarize benchmark results. *Communications of the ACM*, 29(3):218–221, 1986.

[58] Open Power System Data. Data package time series. version 2020-10-06, 2020. https://doi.org/10.25832/time_series/2020-10-06.

[59] Iain Staffell, Stefan Pfenninger, and Nathan Johnson. A global model of hourly space heating and cooling demand at multiple spatial scales. *Nature Energy*, 8(12):1328–1344, 2023.

[60] AI Maverick. Renewable energy and weather conditions. https://www.kaggle.com/datasets/samanemami/renewable-energy-and-weather-conditions, 2025. Kaggle.

[61] lexis Cook, DanB, inversion, and Ryan Holbrook. Store sales – time series forecasting. https://www.kaggle.com/competitions/store-sales-time-series-forecasting, 2020. Kaggle.

[62] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022. Special Issue: M5 competition.

[63] FlorianKnauer and Will Cukierski. Rossmann store sales. https://kaggle.com/competitions/rossmann-store-sales, 2015. Kaggle.

[64] Walmart Competition Admin and Will Cukierski. Walmart recruiting - store sales forecasting. https://kaggle.com/competitions/walmart-recruiting-store-sales-forecasting, 2014. Kaggle.

[65] MichalKecera. Rohlik sales forecasting challenge. https://kaggle.com/competitions/rohlik-sales-forecasting-challenge-v2, 2024. Kaggle.

[66] Jingbo Zhou, Xinjiang Lu, Yixiong Xiao, Jian Tang, Jiantao Su, Yu Li, Ji Liu, Junfu Lyu, Yanjun Ma, and Dejing Dou. SDWPF: A dataset for spatial dynamic wind power forecasting over a large turbine array. *Scientific Data*, 11(1):649, 2024.

[67] Tao Hong, Pierre Pinson, and Shu Fan. Global energy forecasting competition 2012. *International Journal of Forecasting*, 30(2):357–363, 2014.

[68] ECDC. Respiratory viruses weekly data. https://github.com/EU-ECDC/Respiratory_viruses_weekly_data/tree/main, 2025. Open data repository; weekly respiratory virus surveillance in the EU/EEA.

[69] Etienne David, Jean Bellot, and Sylvain Le Corff. HERMES: Hybrid error-corrector model with inclusion of external signals for nonstationary fashion time series. *arXiv preprint arXiv:2202.03224*, 2022.

[70] General Directorate of Health Affairs and Saudi Arabia Ministry of Health. Riyadh hospital admissions dataset (2020–2024). https://www.kaggle.com/dsv/9992619, 2024.

[71] Alexander van Renen, Dominik Horn, Pascal Pfeil, Kapil Vaidya, Wenjian Dong, Murali Narayanaswamy, Zhengchun Liu, Gaurav Saxena, Andreas Kipf, and Tim Kraska. Why TPC is not enough: An analysis of the amazon redshift fleet. *Proc. VLDB Endow.*, 17(11):3694–3706, July 2024.

[72] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757, 2008.

[73] UK COVID-19 data from official UK government sources. UK COVID-19 dashboard data. https://www.kaggle.com/datasets/happyadam73/uk-covid19-dashboard-data-sqlite-compressed, 2022. Kaggle.

[74] Jiyuan Xu, Wenyu Zhang, Xin Jing, Jiahao Nie, Shuai Chen, and Shuai Zhang. CPiRi: Channel permutation-invariant relational interaction for multivariate time series forecasting. In *The International Conference on Learning Representations*, 2026.

[75] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[76] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending tabpfn-v2 to time series forecasting. *arXiv preprint arXiv:2501.02945*, 2025.

A Methodology Design Philosophies

A.1 Overall Architecture

As shown in Fig. 2, the architecture of `Falcon-X` follows a hierarchical transformation pipeline that progressively aligns heterogeneous variates into a unified latent space, and ultimately reconstructs them for accurate forecasting. We formulate the forecasting task as a unified masked reconstruction paradigm. Formally, the input heterogeneous series $\mathbf{X} \in \mathbb{R}^{M\times(L+T)}$ is first subjected to instance normalization and tokenization to generate the time tokens $\mathbf{H} \in \mathbb{R}^{M\times P\times D}$, where $P = \frac{L+T}{L_p}$ is the number of patches and L_p is the length of patches. These tokens are then processed through time attention layers, yielding temporal representations $\mathbf{H}_T \in \mathbb{R}^{M\times P\times D}$.

To bridge the dimensionality gap, `Falcon-X` employs the Unified Prototype Diff-Attention (UPDA) to project the disparate N entities into a compact, unified prototype space $\mathbf{H}_C \in \mathbb{R}^{N\times C\times P\times D}$. Following cross-variate interactions via Latent Entity Attention (LEA), which yields the refined context $\mathbf{H}'_C$, the Variate Reassembly Router (VRR) performs a soft-routing operation to retrieve and reassemble the latent representations back into the entity-specific space $\mathbf{H}_V \in \mathbb{R}^{M\times P\times D}$ by matching the routing request in $\mathbf{H}_T$ with the prototype index in $\mathbf{H}'_C$.

Finally, the reassembled entities $\mathbf{H}_V$ are fused with the temporal representations $\mathbf{H}_T$ and mapped through a quantile forecasting head to produce the final predictive output $\hat{\mathbf{Y}} \in \mathbb{R}^{M\times T}$, completing the end-to-end flow from raw heterogeneous inputs to unified latent representations and back to structured forecasts.

A.2 Time Attention

To capture the intrinsic evolutionary patterns of each individual variate, `Falcon-X` utilizes an encoder-only Transformer architecture. A critical design choice in `Falcon-X` is the deliberate decoupling of temporal and cross-variate modeling. Unlike many existing multivariate models, which interleave temporal and spatial mixing, leading to semantic entanglement whereby a variate's subtle temporal signal is prematurely confounded by the noisy dependencies of its heterogeneous neighbors, `Falcon-X` prioritizes establishing a robust temporal module. Stacking n layers of Time Attention before any cross-variate interaction ensures that the temporal evolutive state of each variate is fully distilled and stabilised. This provides a clean, time-aware foundation for the subsequent Variate Attention process.

A.3 Variate Attention

A core requirement of time-series foundation models is ***the ability to transcend rigid, dataset-specific dimensional constraints and learn a unified representation of dependencies across multivariate series***. However, treating different physical variates homogeneously or simply concatenating them leads to severe semantic misalignment. To overcome this, `Falcon-X` introduces a unified latent space architecture. As shown in Fig. 2(a–c), these modules progressively align the heterogeneous patch embeddings $\mathbf{H}_T$ into a shared prototype space, model both intra- and cross-dataset dependencies, and dynamically reassemble the global context back to the original variate dimensions.

A.3.1 Unified Prototype Diff-Attention

The primary challenge in modeling the foundations of time series lies in reconciling diverse physical variates within a unified semantic space. Dataset-level group mixing [6] relies on dense intra-dataset attention, lacking structural abstraction and incurring quadratic complexity $\mathcal{O}(M^2)$. To address this, `Falcon-X` introduces prototype alignment, projecting heterogeneous variates into a fixed set of learnable latent temporal prototypes of dimension C.

Instead of relying on arbitrary physical indexing, this paradigm dynamically allocates the full representation of each independent variate across the C universal prototypes based on their intrinsic semantic affinity, thereby achieving explicit semantic unification. This explicit mapping resolves the issue of semantic misalignment by aligning variates with similar temporal dynamics to the same semantic anchors, regardless of their original dataset or spatial proximity. Furthermore, by forcing the representations through this fixed-dimensional prototype space, the model inherently performs

structural denoising, filtering out localized noise and isolating the most salient temporal patterns. Importantly, this projection replaces dense intra-dataset attention with variate-to-prototype cross-attention. Entity-local routing requires $\mathcal{O}(MPCD)$ operations, linear in M for fixed P, C, and D. This statement concerns routing rather than the entire model; App. E includes temporal attention, LEA, and projection costs.

Furthermore, we observe that negative correlations among heterogeneous variates, which are critical for capturing counteracting interactions across diverse time series, are difficult to exploit in standard Transformer architectures. This limitation stems from the non-negative nature of the softmax attention function, which restricts attention scores to the range $[0, 1]$ and thus prevents the explicit modeling of opposing trends. To address this, we draw inspiration from differential attention [35]. While originally proposed for attention noise suppression, `Falcon-X` repurposes this mechanism to capture dual-dependency dynamics. By introducing positive and negative learnable keys, the model explicitly represents both synergistic and antagonistic relationships, yielding improved expressiveness of the cross-variate latent space.

A.3.2 Latent Entity Attention

Following the alignment of heterogeneous variates into the unified prototype space, this module models the comprehensive interactions among different variates. As all representations now reside in a shared, dimension-agnostic semantic space rather than their original disparate physical dimensions, Latent Entity Attention naturally facilitates cross-learning. This enables `Falcon-X` to leverage and transfer shared structural patterns across entirely different domains, thereby significantly enhancing zero-shot cross-dataset generalization.

A.3.3 Variate Reassembly Router

After capturing comprehensive dependencies in the unified prototype space, the model must reassemble this global context back into the original heterogeneous dimensions (m_i). `Falcon-X` formulates this reassembly as a targeted retrieval from the abstract prototype space to individual variate trajectories. The aim is to reconstruct the heterogeneous variates, each with distinct temporal patterns, by retrieving relevant information from the unified latent prototypes.

This is orchestrated via a ***request-and-dispatch*** mechanism: the Routing Request ($\mathbf{R}_{\text{req}}$), derived from $\mathbf{h}_T^i$, acts as a structural query conveying the specific physical dimensionality and unique temporal trajectory of the original variate, effectively serving as *entity identity tag*. The request is then matched against the Prototype Index ($\mathbf{P}_{\text{idx}}$), which is an addressable map of the global prototype library. Meanwhile, the Source Context ($\mathbf{S}_{\text{ctx}}$) delivers the refined semantic payloads. Using local, specific trajectories to selectively retrieve unified global prototypes enables this routing paradigm to reconstruct variate-specific patterns with high fidelity.

Finally, considering the significant variance in cross-variate dependencies across diverse datasets, forcing a uniform integration of the global context could introduce detrimental noise to datasets with inherently weak variate correlations. To maintain strict cross-dataset robustness, we introduce an explicit gated residual connection to dynamically fuse the temporal embeddings $\mathbf{H}_T$ with the cross-variate representations $\mathbf{H}_V$. Consequently, it effectively prevents semantic interference in weakly correlated systems while making full use of cross-variate dependencies in strongly correlated ones.

B Dataset Statistics

This section summarizes the datasets used in our experiments. Specifically, App. B.1 describes the corpus used for model pre-training, while App. B.2 and App. B.3 present the benchmarks used for downstream evaluation.

B.1 Pre-training Corpus

Our pre-training corpus comprises both **real-world** and **synthetic** time series datasets, covering a broad range of domains and data-generation characteristics.

Table 2: Summary statistics of univariate pre-training datasets.

Dataset Name	Freq.	Time Series	Variates	Time Points	Domain	Source
BDG-2	H	611	1	9,454,968	Energy	GIFT-Eval
BEIJING_SUBWAY_30MIN	30T	276	2	433,872	Transport	GIFT-Eval
CIF 2016	M	72	1	6,334	Finance	GIFT-Eval
CMIP6	6H	270,336	53	1,973,452,800	Nature	GIFT-Eval
ERA5	H	245,760	45	2,146,959,360	Nature	GIFT-Eval
Electricity	H, W	642	1	8,493,660	Energy	Chronos
HZMETRO	15T	80	2	190,160	Transport	GIFT-Eval
LOS_LOOP	5T	207	1	7,094,304	Transport	GIFT-Eval
LargeST	5T	42,333	1	4,452,510,528	Transport	GIFT-Eval
M1	A, M, Q	921	1	57,882	Finance	GIFT-Eval
M3	A, M, Q	3,003	1	209,114	Finance	GIFT-Eval
NN5	D, W	222	1	93,240	Finance	GIFT-Eval
PEMS03	5T	358	1	9,382,464	Transport	GIFT-Eval
PEMS04	5T	307	3	5,216,544	Transport	GIFT-Eval
PEMS07	5T	883	1	24,921,792	Transport	GIFT-Eval
PEMS08	5T	170	3	3,035,520	Transport	GIFT-Eval
PEMS_BAY	5T	325	1	16,941,600	Transport	GIFT-Eval
Q-TRAFFIC	15T	45,148	1	264,386,688	Transport	GIFT-Eval
Quito	10T, H	33,806	5	313,269,828	Various	QuitoBench
Residential Power	T	504	3	271,333,509	Energy	GIFT-Eval
SHMETRO	15T	288	2	2,536,992	Transport	GIFT-Eval
Solar	5T, H	10,332	1	588,304,080	Energy	Chronos
Taxi	30T, H	70,412	1	56,793,348	Transport	Chronos
Tourism	A, M, Q	1,212	1	150,822	Finance	GIFT-Eval
Traffic	H, W	1,724	1	15,060,864	Transport	GIFT-Eval
Uber TLC	D, H	524	1	1,176,531	Transport	GIFT-Eval
Weatherbench	D, H, W	675,840	1	82,753,646,592	Nature	Chronos
Wind Farms	D, H, T	1,011	1	175,154,333	Energy	Chronos
alibaba_cluster_trace_2018	5T	58,409	2	95,192,530	Web	GIFT-Eval
australian_electricity_demand	30T	5	1	1,153,584	Energy	GIFT-Eval
azure_vm_traces_2017	5T	159,472	1	885,522,908	Web	GIFT-Eval
beijing_air_quality	H	12	11	420,768	Nature	GIFT-Eval
bitcoin_with_missing	D	18	1	81,918	Finance	GIFT-Eval
borealis	H	15	1	83,269	Energy	GIFT-Eval
borg_cluster_data_2011	5T	143,386	2	537,552,854	Web	GIFT-Eval
buildings_900k	H	1,792,328	1	15,702,585,608	Energy	GIFT-Eval
bull	H	41	1	719,304	Energy	GIFT-Eval
cdc_fluview_ilinet	W	75	5	63,903	Healthcare	GIFT-Eval
cdc_fluview_who_nrevss	W	74	4	41,760	Healthcare	GIFT-Eval
china_air_quality	H	437	6	5,739,234	Nature	GIFT-Eval
cockatoo	H	1	1	17,544	Energy	GIFT-Eval
covid19_energy	H	1	1	31,912	Energy	GIFT-Eval
covid_mobility	D	362	1	148,602	Transport	GIFT-Eval
dominick	W	100,014	1	29,652,492	Sales	Chronos
elecdemand	30T	1	1	17,520	Energy	GIFT-Eval
elf	H	1	1	21,792	Energy	GIFT-Eval
exchange_rate	D	8	1	84,976	Finance	Chronos
extended_web_traffic_with_missing	D	145,063	1	370,926,091	Web	GIFT-Eval
godaddy	M	3,135	2	128,535	Finance	GIFT-Eval
hog	H	24	1	421,056	Energy	GIFT-Eval
ideal	H	217	1	1,255,253	Energy	GIFT-Eval
kaggle_web_traffic_weekly	W	145,063	1	16,537,182	Web	GIFT-Eval
lcl	H	713	1	9,543,553	Energy	GIFT-Eval
london_smart_meters_with_missing	30T	5,520	1	166,238,880	Energy	GIFT-Eval
mexico_city_bikes	H	494	1	38,687,004	Transport	Chronos
oikolab_weather	H	8	1	800,456	Nature	GIFT-Eval
pdb	H	1	1	17,520	Energy	GIFT-Eval
pedestrian_counts	H	66	1	3,130,762	Transport	GIFT-Eval
project_tycho	W	1,258	1	1,377,707	Healthcare	GIFT-Eval
rideshare_with_missing	H	2,304	1	859,392	Transport	GIFT-Eval
sceaux	H	1	1	34,223	Energy	GIFT-Eval
smart	H	5	1	95,709	Energy	GIFT-Eval
solar_power	4S	1	1	7,397,222	Energy	GIFT-Eval
spain	H	1	1	35,064	Energy	GIFT-Eval
subseasonal	D	862	4	14,197,140	Nature	GIFT-Eval
subseasonal_precip	D	862	1	9,760,426	Nature	GIFT-Eval
sunspot_with_missing	D	1	1	73,894	Nature	GIFT-Eval
ushcn_daily	D	1,218	5	47,080,115	Nature	Chronos
vehicle_trips_with_missing	D	329	1	32,512	Transport	GIFT-Eval
weather	D	3,010	1	42,941,700	Nature	GIFT-Eval
wiki-rolling_nips	D	47,675	1	40,619,100	Web	GIFT-Eval
wiki_daily_100k	D	100,000	1	274,100,000	Web	Chronos
wind_power	4S	1	1	7,397,147	Energy	GIFT-Eval

Note. Frequency aliases follow common time-series conventions: S = second, T = minute, H = hourly, D = daily, W = weekly, M = monthly, Q = quarterly, and A = annual.

Real-world datasets. We aggregate several large-scale time series collections, including the GIFT-EVAL [8] pre-training dataset[2], the CHRONOS [11] training corpus[3], and the QUITOBENCH [24] training dataset[4], as detailed in Tab. 2. Collectively, these resources span seven major domains: *nature*, *energy*, *transport*, *finance*, *healthcare*, *web*, and *sales*. The resulting corpus contains a large number of *univariate* time series datasets, as well as a small set of *multivariate* time series datasets.

Synthetic univariate datasets. We also incorporate the synthetic *univariate* datasets introduced by CHRONOS [11]: TSMIXUP and KERNELSYNTH. TSMIXUP synthesizes new time series by taking random convex combinations of samples drawn from different real-world datasets, thereby increasing diversity while preserving realistic temporal characteristics. KERNELSYNTH, in contrast, generates synthetic series by randomly composing Gaussian Process (GP) kernels and sampling from the resulting GP priors, producing time series with diverse trends, periodicities, and stochastic patterns.

Synthetic multivariate datasets. High-quality *multivariate* time series datasets remain relatively scarce in existing public resources. To address this limitation, we construct a large amount of synthetic *multivariate* data through two complementary strategies:

1. **Similarity-based multivariate construction from real univariate series.** Drawing from the real univariate time series presented in Tab. 2, we compute pairwise similarities to group related sequences into cohesive multivariate datasets. This process enables us to derive multivariate structures from naturally occurring signals while preserving semantic coherence across dimensions.
2. **Dependency injection over synthetic univariate generators.** Inspired by CHRONOS-2 [6], we transform multiple independently sampled univariate series from base generators (e.g., KERNELSYNTH) into multivariate synthetic time series by imposing explicit dependency structures. These multivariatization procedures include:
   - **Cotemporaneous multivariatizers**, which introduce instantaneous cross-variate dependencies through linear or nonlinear transformations at the same time step;
   - **Sequential multivariatizers**, which impose temporal cross-series relations across time, such as *lead–lag* dependencies and *cointegration*.

Through this combination of real-world corpora, *univariate* synthetic generators, and large-scale *multivariate* synthesis, our dataset collection supports training and evaluation across diverse domains and temporal dependency structures.

B.2 GIFT-Eval Benchmark

GIFT-Eval is constructed from 15 univariate and 8 multivariate datasets, spanning 7 domains and 10 frequencies. The benchmark is reported to contain approximately 144,000 time series and 177 million observations [8]. To support evaluation across forecasting horizons, prediction lengths are determined in two ways. For widely used benchmarks such as M4 [36], established prediction lengths are retained. For the remaining datasets, short horizons depend on frequency and domain, with 48 steps used in many configurations. Where evaluated, medium and long horizons are $10\times$ and $15\times$ the corresponding short horizon, respectively. This results in 97 unique combinations of dataset, frequency, and prediction length, with model performance reported as the geometric mean across these configurations.

The benchmark is curated from 10 publicly available sources covering a diverse set of application domains. Below, the included datasets are grouped by domain and described together with their original sources.

- **Nature.** The benchmark includes the Jena Weather dataset[5], following the preprocessing protocol used in **Autoformer** [37].

[2] https://huggingface.co/datasets/Salesforce/GiftEvalPretrain
[3] https://huggingface.co/datasets/autogluon/chronos_datasets
[4] https://huggingface.co/datasets/hq-bench/quito-corpus
[5] https://www.bgc-jena.mpg.de/wetter/

Table 3: Statistics of the GIFT-Eval benchmark across seven domains. Entries under Short-term/Med-term/Long-term are reported as Pred/Win, denoting the prediction length and the number of rolling windows, respectively.

Domain	Dataset	Freq	#Series	Avg. length	#Vars	Short-term	Med-term	Long-term
Nature	Jena Weather	10T	1	52,704	21	48 / 20	480 / 11	720 / 8
		H	1	8,784	21	48 / 19	480 / 2	720 / 2
		D	1	366	21	30 / 2	–	–
	Saugeen	D	1	23,741	1	30 / 20	–	–
		W-THU	1	3,391	1	8 / 20	–	–
		M	1	780	1	12 / 7	–	–
	Temperature Rain	D	32,072	725	1	30 / 3	–	–
	KDD Cup 2018	H	270	10,898	1	48 / 20	480 / 2	720 / 2
		D	270	455	1	30 / 2	–	–
Web/CloudOps	BizITObs - Application	10S	1	8,834	2	60 / 15	600 / 2	900 / 1
	BizITObs - Service	10S	21	8,835	2	60 / 15	600 / 2	900 / 1
	BizITObs - L2C	5T	1	31,968	7	48 / 20	480 / 7	720 / 5
		H	1	2,664	7	48 / 6	480 / 1	720 / 1
	Bitbrains - Fast Storage	5T	1,250	8,640	2	48 / 18	480 / 2	720 / 2
		H	1,250	721	2	48 / 2	–	–
	Bitbrains - rnd	5T	500	8,640	2	48 / 18	480 / 2	720 / 2
		H	500	720	2	48 / 2	–	–
Energy	ETT1	15T	1	69,680	7	48 / 20	480 / 15	720 / 10
		H	1	17,420	7	48 / 20	480 / 4	720 / 3
		D	1	725	7	30 / 3	–	–
		W-THU	1	103	7	8 / 2	–	–
	ETT2	15T	1	69,680	7	48 / 20	480 / 15	720 / 10
		H	1	17,420	7	48 / 20	480 / 4	720 / 3
		D	1	725	7	30 / 3	–	–
		W-THU	1	103	7	8 / 2	–	–
	Solar	10T	137	52,560	1	48 / 20	480 / 11	720 / 8
		H	137	8,760	1	48 / 19	480 / 2	720 / 2
		D	137	365	1	30 / 2	–	–
		W-FRI	137	52	1	8 / 1	–	–
	Electricity	15T	370	140,256	1	48 / 20	480 / 20	720 / 20
		H	370	35,064	1	48 / 20	480 / 8	720 / 5
		D	370	1,461	1	30 / 5	–	–
		W-FRI	370	208	1	8 / 3	–	–
Transport	Loop Seattle	5T	323	105,120	1	48 / 20	480 / 20	720 / 15
		H	323	8,760	1	48 / 19	480 / 2	720 / 2
		D	323	365	1	30 / 2	–	–
	SZ-Taxi	15T	156	2,976	1	48 / 7	480 / 1	720 / 1
		H	156	744	1	48 / 2	–	–
	M_DENSE	H	30	17,520	1	48 / 20	480 / 4	720 / 3
		D	30	730	1	30 / 3	–	–
Sales	Restaurant	D	807	358	1	30 / 1	–	–
	Hierarchical Sales	D	118	1,825	1	30 / 7	–	–
		W-WED	118	260	1	8 / 4	–	–
	Car Parts	M	2,674	51	1	12 / 1	–	–
Econ/Fin	M4 Yearly	A	22,974	37	1	6 / 1	–	–
	M4 Quarterly	Q	24,000	100	1	8 / 1	–	–
	M4 Monthly	M	48,000	234	1	18 / 1	–	–
	M4 Weekly	W	359	1,035	1	13 / 1	–	–
	M4 Daily	D	4,227	2,371	1	14 / 1	–	–
	M4 Hourly	H	414	902	1	48 / 2	–	–
Healthcare	Hospital	M	767	84	1	12 / 1	–	–
	COVID Deaths	D	266	212	1	30 / 1	–	–
	US Births	D	1	7,305	1	30 / 20	–	–
		W-TUE	1	1,043	1	8 / 14	–	–
		M	1	240	1	12 / 2	–	–

- **Web/CloudOps.** This domain contains the BizITObs Application, Service, and L2C datasets[6], processed according to the pipeline introduced in **AutoMixer** [38]. These datasets combine business KPIs with IT event channels, forming multivariate time series for observability-related forecasting tasks. In addition, Bitbrains datasets from the **Grid Workloads Archive** [39] are included in the same domain.
- **Sales.** For the sales domain, the Restaurant dataset is adopted from the **Recruit Restaurant Forecasting Competition** [40], where the objective is to predict future customer visits using reservation and visitation records. Another sales dataset is included from Mancuso et al. [41].
- **Energy.** The energy domain includes ETT1 and ETT2 from **Informer** [42], which represent electricity transformer temperature and are widely used in long-horizon forecasting. It also in-

[6] https://github.com/BizITObs/BizITObservabilityData/tree/main

cludes the Electricity dataset from the **UCI ML Archive** [43], containing electricity consumption records for 370 clients, and the Solar dataset from **LSTNet** [44], which focuses on forecasting solar plant power output.

- **Transport.** Transport datasets are drawn from **LibCity** [45], a benchmark collection of urban spatio-temporal and time series datasets.
- **Econ/Fin & Healthcare.** A subset of datasets is selected from the **Monash** repository [46], which provides a broad collection of time series from multiple domains. The selected datasets are chosen to avoid any leakage between pretraining and test data.

Detailed dataset statistics are provided in Tab. 3, including frequency, prediction length, variate setting, number of series, and average series length. Evaluation uses the benchmark-specific forecast horizons and rolling-window counts listed in Tab. 3; the resulting test fraction varies across datasets.

B.3 fev-bench Benchmark

The fev-bench benchmark comprises a total of 100 time series forecasting tasks. Detailed dataset statistics are provided in Tab. 4. This section summarizes the main characteristics of these tasks and provides citations for the corresponding data sources. For datasets originating from forecasting competitions, the benchmark adopts the fixed forecast horizon T specified by the original competition setup. For all other datasets, the forecast horizon is determined according to a frequency–horizon mapping. An exception is made for a subset of hourly datasets, for which $T = 168$ is used in order to support long-range forecasting over a one-week period. The number of evaluation windows W is then selected so as to split each series as evenly as possible while ensuring that sufficient historical context remains available for every forecast of length T. Dataset frequencies are reported using `pandas` frequency aliases, namely minu**T**ely, **H**ourly, **D**aily, **W**eekly, **M**onthly, **Q**uarterly, and **Y**early.

The benchmark is constructed from a diverse collection of domains, including macroeconomics, energy systems, retail and sales forecasting, epidemiology, public health, environmental monitoring, and database operations. The included datasets can be grouped into the following source categories.

- **GIFT-Eval.** The benchmark includes datasets from the GIFT-Eval corpus [8], which contains a mixture of univariate and multivariate forecasting tasks. The original GIFT-Eval collection draws on data sources compiled from prior benchmark and application papers [46, 47, 41, 37, 38].
- **Macroeconomic datasets.** A broad set of macroeconomic and socioeconomic datasets is included, such as GVAR [48], US Consumption [49], Australian Tourism [50], FRED-MD [51], FRED-QD [52], world CO_2 emissions [53], life expectancy [54], and global tourism [55]. For both FRED-MD and FRED-QD, two separate forecasting tasks are defined. The first task follows the CEE model [56] and focuses on forecasting employment, inflation, and federal funds rate indicators. The second task considers the joint forecasting of 51 core macroeconomic indicators. It should be noted that the benchmark uses the August 2025 snapshot of FRED-MD, which differs from the snapshot used in Monash repository [46].
- **Energy datasets.** The energy-related portion of the benchmark includes several forecasting settings of practical relevance. These datasets cover the electricity price forecasting (EPF) benchmark [57], ERCOT generation data [11], ENTSO-e load data [58] paired with weather variates obtained from `Renewables.ninja` [59], and solar generation data [60]. Together, these datasets provide a mix of load, price, and renewable generation forecasting tasks.
- **BOOMLET.** The benchmark also includes multivariate observability datasets from BOOMLET [5], which is itself a subset of the larger BOOM benchmark curated by the original authors. To maintain diversity across data sources and prevent overrepresentation from a single benchmark family, only BOOMLET datasets with a sampling frequency of at least one minute are retained.
- **Forecasting competitions.** A substantial portion of the benchmark is drawn from forecasting competitions, many of which were hosted on `kaggle.com`. These include Favorita store sales and transactions [61], the M5 competition [62], restaurant visitor and reservation forecasting [40], Rossmann store sales [63], Walmart sales forecasting [64], and Rohlik sales forecasting [65]. In addition, the benchmark includes the KDD Cup 2022 dataset for wind power forecasting [66], as well as datasets from the Global Energy Forecasting Competitions held in 2012, 2014, and 2017

[67]. These competition datasets typically come with standardized train–test setups and fixed forecast horizons, making them especially useful for controlled model comparison.

- **Other sources.** To further broaden domain coverage, the benchmark incorporates datasets from several additional sources:
  - Influenza-like illness case counts collected by the European Centre for Disease Prevention and Control [68].
  - Fashion trend data from Hermes [69].
  - Hospital admissions data from Riyadh [70].
  - Query count data for Amazon Redshift database servers [71].
  - Solar energy generation data with associated weather covariates [60].
  - Air quality measurements from an Italian city together with weather variates [72].
  - COVID-19 cases, hospital admissions, and deaths in the United Kingdom across multiple administrative levels [73].

These additional datasets complement the benchmark by introducing forecasting tasks from healthcare, epidemiology, fashion, environmental sensing, and cloud/database system monitoring, thereby increasing the breadth of real-world scenarios represented in fev-bench.

Table 4: Individual statistics of the fev-bench benchmark across all datasets.

Task	Domain	Freq.	T	W	Median length	# series	# targets
GIFT-Eval							
BizITObs-L2C	cloud	5T	288	20	31,968	1	7
BizITObs-L2C	cloud	H	24	20	2,664	1	7
ETT	energy	15T	96	20	69,680	2	7
ETT	energy	H	168	20	17,420	2	7
ETT	energy	D	28	20	724	2	7
ETT	energy	W	13	5	103	2	7
Hierarchical Sales	retail	D	28	10	1,825	118	1
Hierarchical Sales	retail	W	13	10	260	118	1
Hospital	healthcare	M	12	4	84	767	1
Jena Weather	nature	10T	144	20	52,704	1	21
Jena Weather	nature	D	28	11	366	1	21
Jena Weather	nature	H	24	20	8,784	1	21
Loop Seattle	mobility	D	28	10	365	323	1
Loop Seattle	mobility	5T	288	10	105,120	323	1
Loop Seattle	mobility	H	168	10	8,760	323	1
M-DENSE	mobility	D	28	10	730	30	1
M-DENSE	mobility	H	168	10	17,520	30	1
SZ Taxi	mobility	15T	96	10	2,976	156	1
SZ Taxi	mobility	H	168	2	744	156	1
Solar	energy	W	13	1	52	137	1
Solar	energy	D	28	10	365	137	1
Macroeconomic datasets							
Australian Tourism	econ	Q	8	2	36	89	1
FRED-MD-CEE	econ	M	12	20	798	1	3
FRED-MD-Macro	econ	M	12	20	798	1	51
FRED-QD-CEE	econ	Q	8	20	266	1	3
FRED-QD-Macro	econ	Q	8	20	266	1	51
GVAR	econ	Q	8	10	178	33	6
US Consumption	econ	M	12	10	792	31	1
US Consumption	econ	Q	8	10	262	31	1
US Consumption	econ	Y	5	10	64	31	1
World CO2 Emissions	econ	Y	5	9	60	191	1
World Life Expectancy	econ	Y	5	10	74	237	1
World Tourism	econ	Y	5	2	21	178	1
Energy datasets							
ENTSO-e Load	energy	15T	96	20	175,292	6	1
ENTSO-e Load	energy	30T	96	20	87,645	6	1
ENTSO-e Load	energy	H	168	20	43,822	6	1
EPF-BE	energy	H	24	20	52,416	1	1
EPF-DE	energy	H	24	20	52,416	1	1

Continued on next page

Table 4: Individual statistics of the fev-bench benchmark across all datasets. *(continued)*

Task	Domain	Freq.	T	W	Median length	# series	# targets
EPF-FR	energy	H	24	20	52,416	1	1
EPF-NP	energy	H	24	20	52,416	1	1
EPF-PJM	energy	H	24	20	52,416	1	1
ERCOT	energy	D	28	20	6,452	8	1
ERCOT	energy	H	168	20	154,872	8	1
ERCOT	energy	M	12	15	211	8	1
ERCOT	energy	W	13	20	921	8	1
GFC12	energy	H	168	10	39,414	11	1
GFC14	energy	H	168	20	17,520	1	1
GFC17	energy	H	168	20	17,544	8	1
Solar with Weather	energy	15T	96	20	198,600	1	1
Solar with Weather	energy	H	24	20	49,648	1	1
BOOMLET							
BOOMLET-1062	cloud	5T	288	20	16,384	1	21
BOOMLET-1209	cloud	5T	288	20	16,384	1	53
BOOMLET-1225	cloud	T	60	20	16,384	1	49
BOOMLET-1230	cloud	5T	288	20	16,384	1	23
BOOMLET-1282	cloud	T	60	20	16,384	1	35
BOOMLET-1487	cloud	5T	288	20	16,384	1	54
BOOMLET-1631	cloud	30T	96	20	10,463	1	40
BOOMLET-1676	cloud	30T	96	20	10,463	1	100
BOOMLET-1855	cloud	H	24	20	5,231	1	52
BOOMLET-1975	cloud	H	24	20	5,231	1	75
BOOMLET-2187	cloud	H	24	20	5,231	1	100
BOOMLET-285	cloud	T	60	20	16,384	1	75
BOOMLET-619	cloud	T	60	20	16,384	1	52
BOOMLET-772	cloud	T	60	20	16,384	1	67
BOOMLET-963	cloud	T	60	20	16,384	1	28
Forecasting competitions							
Favorita Store Sales	retail	M	12	2	54	1,579	1
Favorita Store Sales	retail	W	13	10	240	1,579	1
Favorita Store Sales	retail	D	28	10	1,688	1,579	1
Favorita Transactions	retail	M	12	2	54	51	1
Favorita Transactions	retail	W	13	10	240	51	1
Favorita Transactions	retail	D	28	10	1,688	51	1
KDD Cup 2022	energy	D	14	10	243	134	1
KDD Cup 2022	energy	10T	288	10	35,279	134	1
KDD Cup 2022	energy	30T	96	10	11,758	134	1
M5	retail	M	12	1	58	30,490	1
M5	retail	W	13	1	257	30,490	1
M5	retail	D	28	1	1,810	30,490	1
Restaurant	retail	D	28	8	296	817	1
Rohlik Orders	retail	W	8	5	170	7	1
Rohlik Orders	retail	D	61	5	1,197	7	1
Rohlik Sales	retail	W	8	1	150	5,243	1
Rohlik Sales	retail	D	14	1	1,046	5,390	1
Rossmann	retail	W	13	8	133	1,115	1
Rossmann	retail	D	48	10	942	1,115	1
Walmart	retail	W	39	1	143	2,936	1
Other datasets							
ECDC ILI	healthcare	W	13	10	201	25	1
Hermes	retail	W	52	1	261	10,000	1
Hospital Admissions	healthcare	D	28	20	1,731	8	1
Hospital Admissions	healthcare	W	13	16	246	8	1
Redset	cloud	5T	288	10	25,920	118	1
Redset	cloud	15T	96	10	8,640	126	1
Redset	cloud	H	24	10	2,160	138	1
UCI Air Quality	nature	H	168	20	9,357	1	4
UCI Air Quality	nature	D	28	11	389	1	4
UK COVID-Nation-Cumulative	healthcare	D	28	20	729	4	3
UK COVID-Nation-Cumulative	healthcare	W	8	4	105	4	3
UK COVID-Nation-New	healthcare	D	28	20	729	4	3
UK COVID-Nation-New	healthcare	W	8	4	105	4	3
UK COVID-UTLA-Cumulative	healthcare	W	13	5	104	214	1
UK COVID-UTLA-New	healthcare	D	28	10	721	214	1

C Sampling Details

C.1 Flexible Context Length and Horizon Sampling

Unlike decoder-only transformers that inherently support variable context and prediction lengths during pre-training, encoder-only architectures face significant challenges in achieving such flexible length settings, which is quite crucial for model generalization. Theoretically, while models pre-trained with fixed target horizon can perform arbitrary-length inference via auto-regression, they are susceptible to the cumulative error propagation common in decoder-only structures. To address this, we implement flexible context length and horizon during the sampling stage. Specifically, for each sampled entity e_i with context length L_i and target horizon T_i, we construct each batch by left-padding input contexts with `NaN` values to the batch-wise maximum input length $L_{\text{max}} = \max\{L_i\}_{i=1}^{N}$. Meanwhile, target outputs are right-padded to a pre-defined maximum horizon $T_{\text{max}} = 480$, corresponding to 30 output tokens with patch length $L_p = 16$. The padded context positions, together with invalid observations, constitute the binary observation mask $\mathcal{M}$, while the padded target segments are excluded from the final loss calculation. Such flexible sampling strategy effectively enhances the predictive generalization of `Falcon-X`.

C.2 Runtime Multivariate Sampling

To mitigate potential GPU memory bottlenecks arising from multivariate sampling, we adopt a dynamic variate sampling strategy at runtime. After flexible context length and horizon sampling, for each entity e_i with v_i variates, we first randomly permute the variate order to reduce order bias and encourage permutation-robust, content-driven cross-variate modeling, consistent with prior findings that channel shuffling improves robustness to channel ordering in multivariate forecasting [74]. Then we traverse variates in the permuted order. Subsequently, we iteratively append valid variates to the training sample until all candidates are processed or a pre-defined per-sample limit, $M_{\max}$, is reached. Finally, entity e_i contributes $m_i = \min(v_i, M_{\max})$ variates. This runtime design naturally supports heterogeneous variate dimensionalities and prevents dimensionality-related computational bottlenecks while facing excessively large v_i.

After variate selection, multivariate samples may still vary in variate count. To batch them efficiently, we enforce a batch-level variate budget and accumulate samples until the total number of retained variates reaches a preset threshold, stabilizing memory usage across training steps. This variate-wise batching strategy substantially reduces channel-padding waste and enables efficient training on multivariate data with highly variable variate dimensionality.

D Additional Controlled Ablations

We report additional ablations using the GIFT-Eval aggregate MASE and CRPS metrics, for which lower values are better. Within each comparison, the data, backbone scale, training steps, optimization, and model-selection protocol are held fixed, with only the indicated component changed.

D.1 Interaction after Prototype Alignment

The w/o LEA variant retains UPDA and VRR but removes interaction among the aligned entity–prototype tokens. As shown in Tab. 5, the full model reduces MASE by approximately 1.2% (0.695 to 0.687) and CRPS by approximately 0.8% (0.489 to 0.485) relative to this variant. This supports the contribution of post-alignment interaction beyond UPDA aggregation alone. A literal w/o UPDA removal is not a defined single-module ablation, because it removes the shared prototype interface required by LEA and VRR; replacing that interface would require an alternative architecture.

Table 5: Controlled LEA ablation. Both variants retain UPDA and VRR.

Variant	MASE ↓	CRPS ↓
w/o LEA	0.695	0.489
`Falcon-X`	**0.687**	**0.485**

D.2 Signed Aggregation and Projection Capacity

Removing the negative branch changes both the signed mechanism and branch capacity. To reduce this confound, we compare against a widened positive-only branch while retaining the backbone, prototype count, LEA, and VRR. The capacity-dominant branch projection scales quadratically with its internal width: two width-D branches have relative projection capacity $2D^2$, matching the proxy $(\sqrt{2}D)^2$ of a single widened branch.

Table 6: Projection-capacity control for signed aggregation. The capacity column is a branch-projection proxy, not the total parameter count.

Variant	Signed subtraction	Projection capacity	MASE ↓	CRPS ↓
Only-K_{pos}, width D	No	D^2	0.701	0.513
Positive-only, width $\sqrt{2}D$	No	$2D^2$	0.694	0.492
`Falcon-X`, two width-D branches	Yes	$2D^2$	**0.687**	**0.485**

The widened positive-only control recovers part of the degradation but does not close the gap to the signed model (Tab. 6). This supports a contribution from signed aggregation beyond the tested increase in projection capacity. The control matches this projection-capacity proxy, not every model parameter or FLOP, and does not establish a causal interpretation of the learned negative affinities.

D.3 Normalization

We replace only the normalization/de-normalization pipeline with RevIN [20]. As shown in Tab. 7, observed-only standardization followed by asinh, inverted with $\sinh$ and the stored statistics, improves both metrics. This comparison supports the complete normalization pipeline; it does not separately isolate each element of that pipeline.

Table 7: Controlled comparison of normalization pipelines.

Normalization	MASE ↓	CRPS ↓
RevIN	0.703	0.493
Observed-only standardization + asinh / $\sinh$	**0.687**	**0.485**

D.4 Positional Encoding

`Falcon-X` uses rotary position embedding (RoPE) [75] in attention. RoPE introduces relative displacement into query–key interactions, whereas the appended scalar timestamp anchors the history/forecast boundary. The observation mask separately identifies observed versus missing or masked positions. In Tab. 8, all variants retain the same timestamp and mask; only the positional encoding is changed. RoPE achieves the lowest MASE, while learnable positional encoding achieves the lowest CRPS among the tested alternatives.

Table 8: Positional-encoding ablation with the same timestamp and mask.

Positional encoding	MASE ↓	CRPS ↓
Sinusoidal	0.719	0.485
Learnable	0.702	**0.473**
RoPE	**0.687**	0.485

E Computational Complexity

Let N denote the number of entities, $M = \sum_{i=1}^{N} m_i$ the total number of physical variates, P the patch count, C the prototype count per entity, and D the hidden width. The Time Attention and LEA depths are n and l, respectively. We count forward-pass operations with a fixed number of attention heads and fixed feed-forward expansion ratios.

Attention and routing. Temporal attention independently processes P patches for each of the M variates, costing $\mathcal{O}(nMP^2D)$. Entity-local UPDA and VRR route between m_i variates and C prototypes at each patch, costing $\mathcal{O}(MPCD)$ in total; two signed branches and the two routing directions change constant factors. LEA performs attention over NC aligned tokens at each patch, costing $\mathcal{O}(lP(NC)^2D)$. These terms together are

$$\mathcal{O}\big(nMP^2D + MPCD + lP(NC)^2D\big)\,. \tag{13}$$

Projections and feed-forward operations. The temporal and latent stacks add $\mathcal{O}(nMPD^2 + lPNCD^2)$ for their linear projections and any fixed-expansion feed-forward sublayers. UPDA, VRR, and the residual gate additionally require $\mathcal{O}((M+NC)PD^2)$. The resulting total is

$$\begin{aligned}\mathcal{O}\big(&nMP^2D + MPCD + lP(NC)^2D \\ &+ nMPD^2 + lPNCD^2 + (M+NC)PD^2\big).\end{aligned} \tag{14}$$

For fixed patch length and quantile count, input embedding and the forecasting head are linear in the number of patches and introduce no additional quadratic interaction term.

Scope of the efficiency claim. Global raw-variate attention would instead contain $\mathcal{O}(lPM^2D)$; grouped raw-variate attention contains $\mathcal{O}(lP\sum_g M_g^2 D)$, where M_g is the number of variates in group g. The benefit of prototype interaction therefore depends on entity size, prototype count, and grouping. Only routing is linear in M with P, C, D fixed; LEA remains quadratic in NC. The routing count assumes computation is restricted to the allowed entity-local pairs. Materializing dense routing scores between all M variates and all NC prototypes would instead cost $\mathcal{O}(MPNCD)$, even if disallowed pairs are subsequently masked. Thus, masks alone do not guarantee arithmetic or memory savings. These operation counts characterize the architecture and do not constitute a measured latency or memory bound.

F Additional Results

To examine performance across prediction lengths, we stratify the GIFT-Eval MASE results into short, medium, and long horizons (Fig. 8). `Falcon-X` obtains 0.70, 0.66, and 0.67, respectively, compared with 0.68, 0.70, and 0.71 for Timer-S1 [23] and 0.67, 0.73, and 0.76 for Chronos-2 [6]. Thus, `Falcon-X` performs best among the compared models on medium and long horizons, while several baselines perform better on short horizons.

On fev-bench (Fig. 9), `Falcon-X` achieves MASE of 0.689 and CRPS of 0.513. Both metrics are lower than those of TiRex [18] (0.699, 0.533), TabPFN-TS [76] (0.719, 0.535), and Toto-1.0 [5] (0.715, 0.547). These aggregate comparisons show lower reported errors; statistical significance has not been established. Chronos-2 [6] obtains lower MASE (0.645) and CRPS (0.485) than `Falcon-X`, using additional past-only and future-known covariates. `Falcon-X` uses only endogenous target series. The comparison therefore reflects both model and input-setting differences.

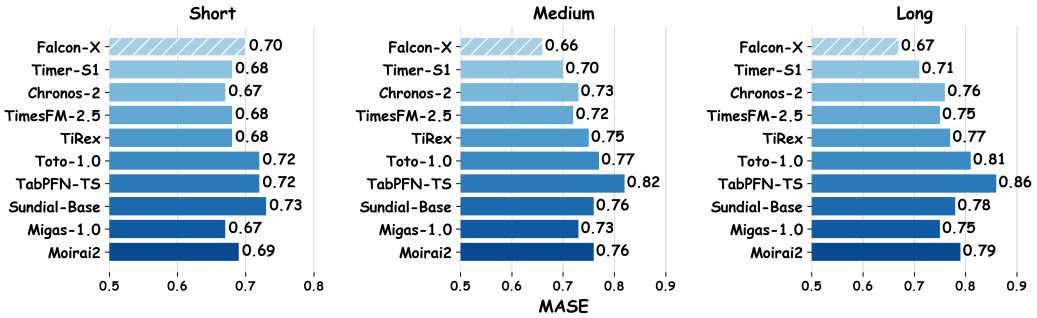

Figure 8: GIFT-Eval MASE grouped by prediction horizon. `Falcon-X` achieves the lowest reported MASE on medium and long horizons among the compared models.

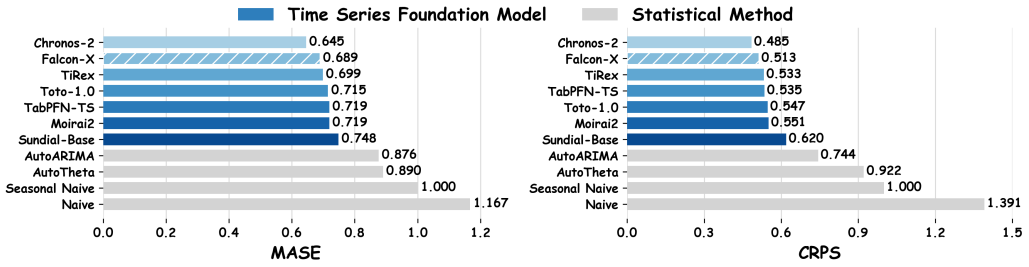

Figure 9: Performance of **Falcon-X** on the fev-bench leaderboard.

G Additional Visualization

We provide visualizations of **Falcon-X**'s quantile forecasts across representative datasets at different frequencies {5T, 15T, 10S, H} and prediction horizons $T = \{48, 60, 480, 720\}$. Specifically, Fig. 10 shows medium-horizon ($T = 480$) forecasts for Loop Seattle/H, Fig. 11 for KDD Cup 2018/H, and Fig. 12 for Electricity/H at $T = 720$. Fig. 13 presents forecasts on Bitbrains Fast Storage/5T at $T = 48$, where the two channels exhibit strong positive correlation, which **Falcon-X** accurately captures. In contrast, Fig. 14 shows forecasts on Bizitobs Application/10s at $T = 60$, where two channels are negatively correlated, and **Falcon-X** successfully models the inverse relationship. These results demonstrate that the variable-to-prototype design effectively captures both positive and negative inter-variable dependencies, and preserves temporal consistency across diverse datasets. Moreover, the visualizations highlight **Falcon-X**'s robustness in handling different sampling frequencies and prediction horizons without manual adjustment.

Moreover, we visualize **Falcon-X**'s medium-horizon ($T = 480$) quantile forecasts on ETTh1/15T and ETTh1/H, comparing results with and without multivariate inference across all seven channels. As shown in Fig. 15 and Fig. 16, enabling multivariate inference allows **Falcon-X** to more accurately capture complex inter-variable relationships, including both strong and subtle dependencies, which results in tighter predictive intervals and improved alignment with observed dynamics across channels.

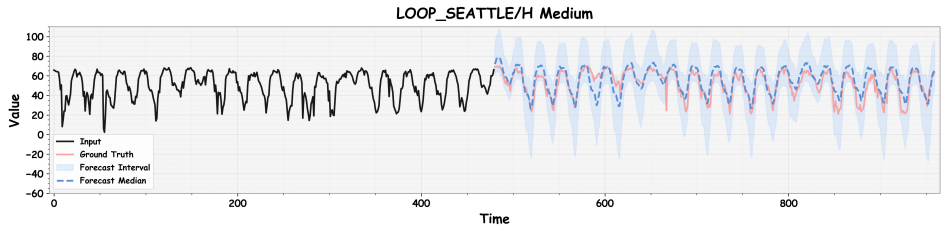

Figure 10: Medium-horizon ($T = 480$) quantile forecasts on Loop Seattle.

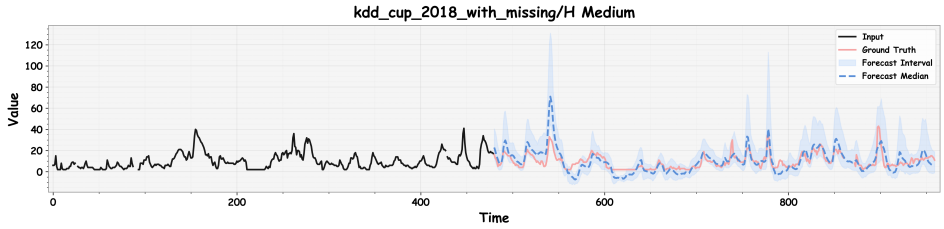

Figure 11: Medium-horizon ($T = 480$) quantile forecasts on kdd cup 2018.

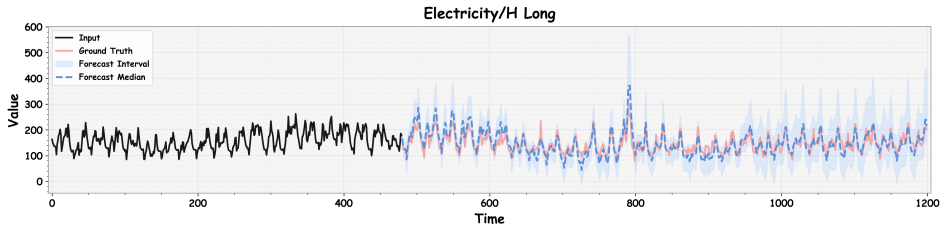

Figure 12: Long-horizon ($T = 720$) quantile forecasts on electricity.

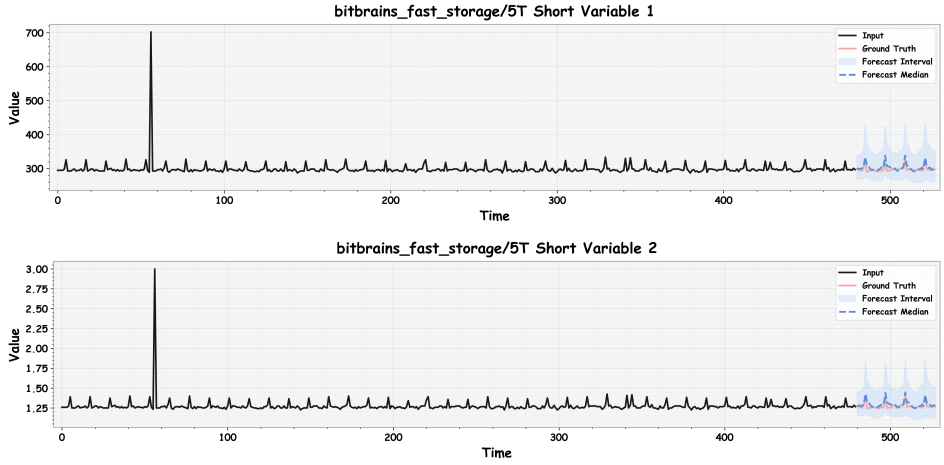

Figure 13: Short-horizon ($T = 48$) quantile forecasts on bitbrains fast storage.

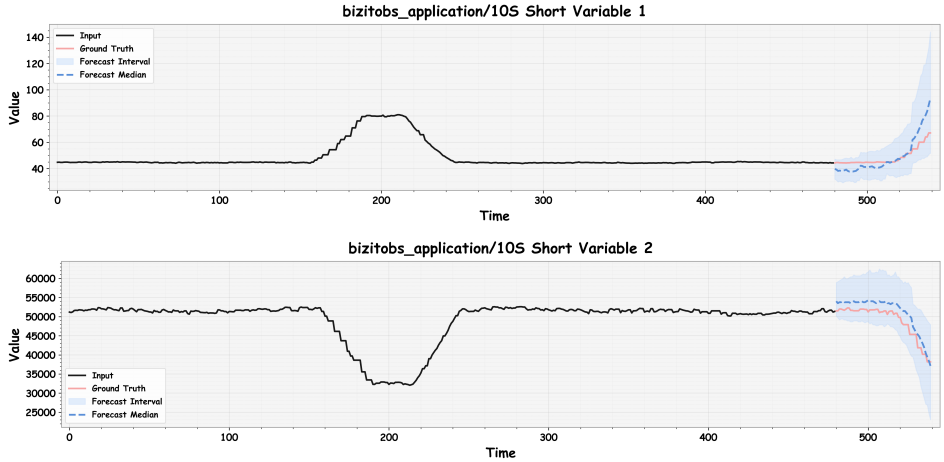

Figure 14: Short-horizon ($T = 60$) quantile forecasts on bizitobs application.

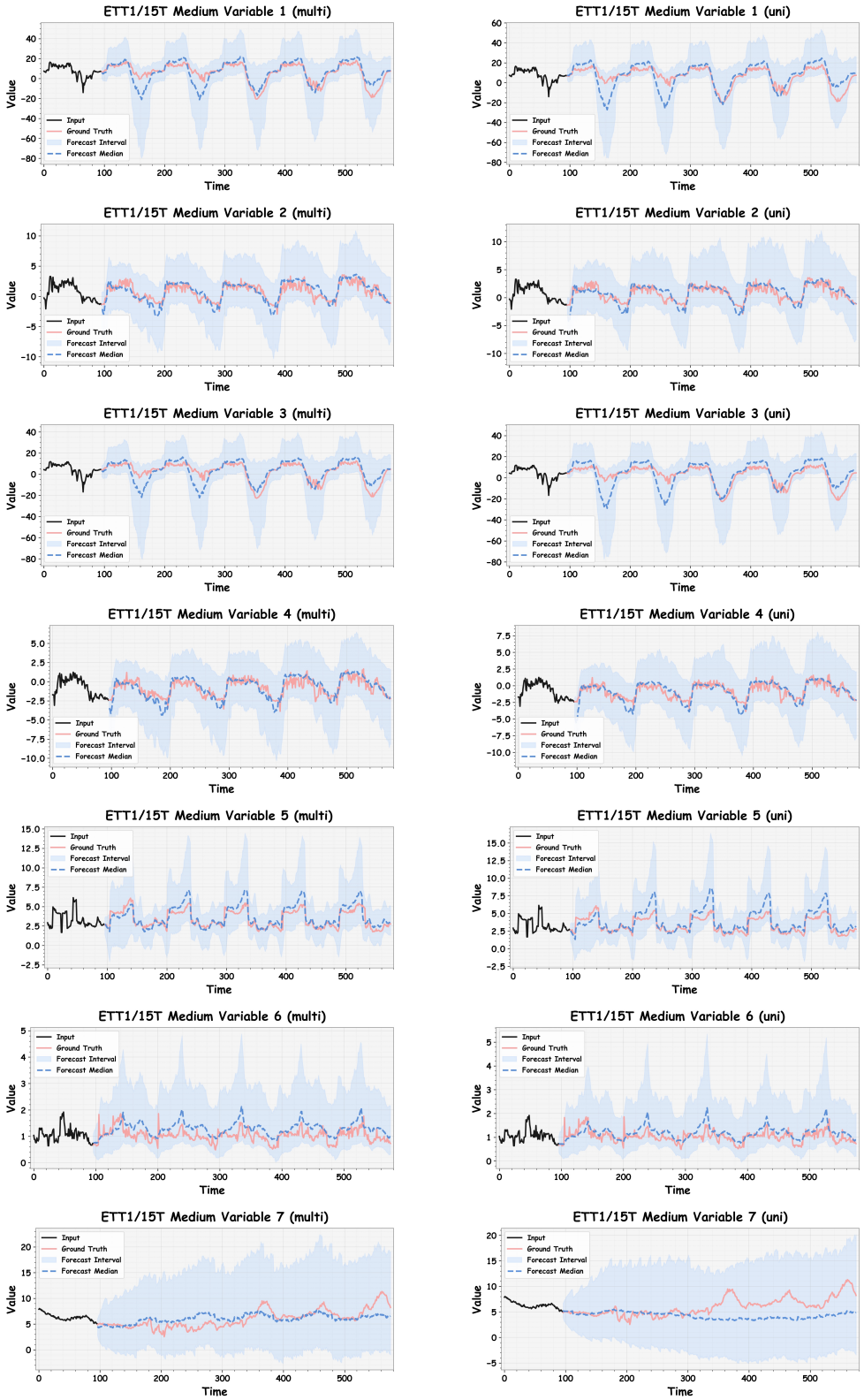

Figure 15: Medium-horizon ($T = 480$) quantile forecasts on ETT1/15T with 7 channels, comparing `Falcon-X` with and without multivariate inference.

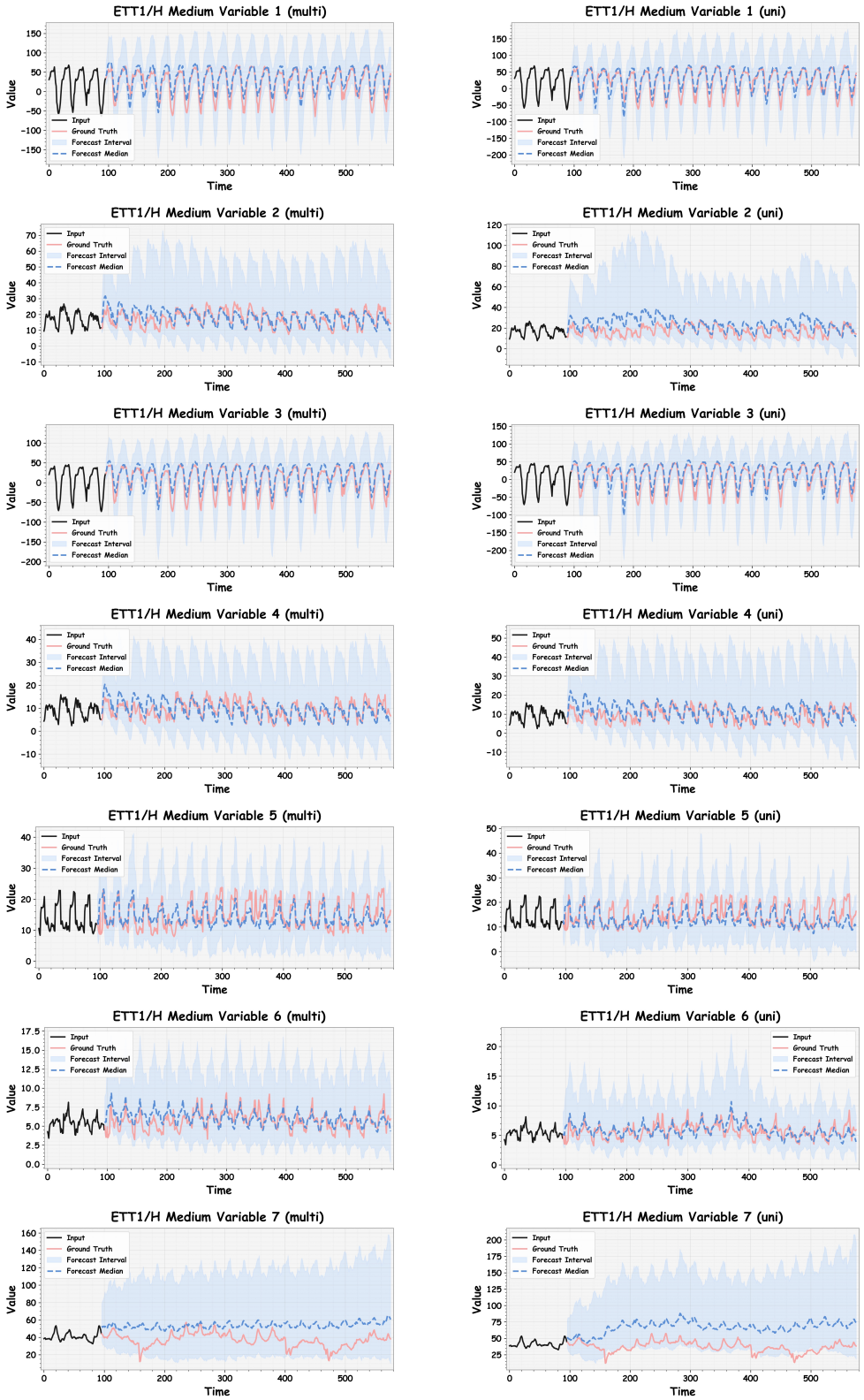

Figure 16: Medium-horizon ($T = 480$) quantile forecasts on ETT1/H with 7 channels, comparing `Falcon-X` with and without multivariate inference.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the proposed method, its intended scope, and the empirical claims evaluated in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Sec. 6 discusses the scope of evaluation, prototype interpretability and robustness, limitations of partitioned inference, and computational costs.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results; all findings are empirical.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The code are in the supplementary material

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: See Sec. 4.2, App. B.1, App. B.2, App. B.3, and the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Sec. 4.2, App. B.2, App. B.3, and App. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: As a probabilistic forecasting model, `Falcon-X` naturally provides uncertainty estimation, including confidence intervals and error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Sec. 4.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper complies with standard ethical guidelines for academic research.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Sec. 7 discusses potential societal benefits and risks arising from forecasting errors, distribution shifts, and high-stakes applications, together with mitigation measures including validation, continuous monitoring, and domain-expert oversight.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and codebases are properly cited, with licenses and terms respected.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This research does not use LLMs as part of its core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.