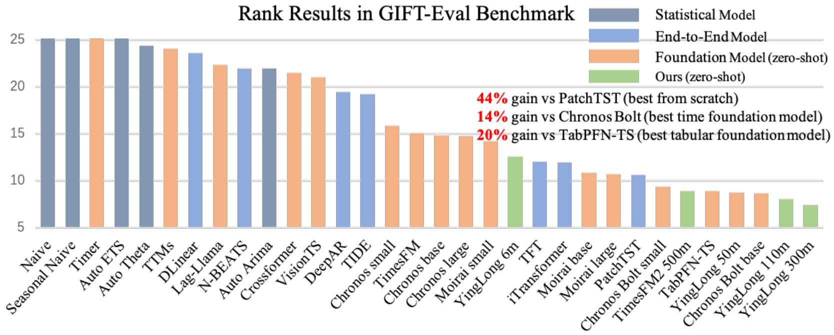
Output Scaling: YingLong Delayed Chain of Thought in a Large Pretrained Time Series Forecasting Model

Xue Wang[∗] **Tian Zhou**[∗] **Jinyang Gao** **Bolin Ding** **Jingren Zhou**
`{xue.w,tian.zt,jinyang.gjy,bolin.ding,jingren.zhou}@alibaba-inc.com`

Abstract

We present a joint forecasting framework for time series prediction that contrasts with traditional direct or recursive methods. This framework achieves state-of-the-art performance for our designed foundation model, YingLong, and reveals a novel scaling effect: longer outputs significantly enhance model accuracy due to delayed chain-of-thought reasoning in our non-causal approach. YingLong is a non-causal, bidirectional attention encoder-only transformer trained through masked token recovery, aligning more effectively with language understanding tasks than with generation tasks. Additionally, we boost performance by tackling output variance with a multi-input ensemble. We release four foundation models ranging from 6M to 300M parameters, demonstrating superior results in zero-shot tasks on the ETT and Weather datasets. YingLong achieves more than 60% best performance. To ensure generalizability, we assessed the models using the GIFT-Eval benchmark, which comprises 23 time series datasets across 7 domains. YingLong significantly outperformed the best time-series foundation models, end-to-end trained models by 14% and 44% in rank respectively.The pretrained 300M model is available at `https://huggingface.co/qcw1314/YingLong_300m`



1 Introduction

Time series data play a crucial role in dynamic real-world systems and applications across various domains (Box et al., 2015; Zhang et al., 2024; Liang et al., 2024). Analyzing such data is inherently challenging due to their complexity and distribution shifts, yet gaining insights from them is essential for enhancing predictive analytics and decision-making.

Both recursive and direct forecasting paradigms possess inherent limitations that necessitate novel approaches. The recursive forecasting method often operates under the assumption that time series

∗ Equal contribution

Preprint. Under review.

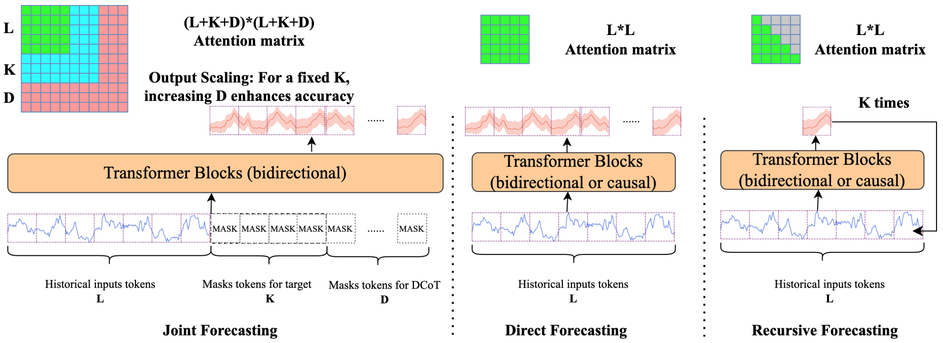


Figure 1: Joint Forecasting with Output Scaling via Delayed Chain of Thought (DCoT). In joint forecasting, each transformer block employs a fully dense attention of size $(L+K+D)\times(L+K+D)$. In the direct forecasting paradigm, a dense attention map of size $L \times L$ is used. Conversely, recursive forecasting utilizes either a half-dense or fully dense $L \times L$ attention map.

signals exhibit a causal and auto-regressive generative nature. However, this assumption frequently fails to account for the complexities inherent in many time series datasets, which are affected by latent driving factors that are not readily observable and do not exhibit self-autoregressive behavior. Moreover, even in recursive forecasting without a strict causal mask, error accumulation remains a significant challenge when using predicted values as inputs. So forecasting shares similarities with natural language understanding (NLU), wherein the integration of comprehensive input signals through bidirectional information flow is essential for accurate outcome prediction. In the well-established NLU METE Benchmark (Massive Text Embedding Benchmark) (Muennighoff et al., 2022), large language models (LLMs) have surpassed smaller BERT-based models, establishing themselves as the leading solutions within the benchmark. Notably, the top-performing LLM approaches have adopted non-causal bidirectional schemes, utilizing edited masks to enhance their performance (Lee et al., 2024; Li et al., 2023; Gao et al., 2023).

Conversely, the direct forecasting method, while often exhibiting superior numerical performance in various time series prediction tasks compared to recursive forecasting, faces its own challenges. In direct forecasting, each output is predicted independently, employing a large transformer backbone as a feature extraction mechanism with a single output layer for each prediction point. This approach assumes a complete non-causal and non-auto-regressive nature between outputs, which, similar to recursive forecasting, may not be representative of the intrinsic dynamics of numerous time series datasets.

Our objective is to integrate the strengths of both recursive and direct forecasting methods by harnessing correlated output modeling while alleviating the constraints of strict causal relationships. We introduce a novel approach termed the joint forecasting paradigm, which promises to be a more robust framework for time series prediction tasks. Within this paradigm, we have developed a large pre-trained time series forecasting model based on a non-causal bidirectional encoder-only transformer architecture. By employing mask token recovery during training, we aim to fully exploit the bidirectional flow of information, offering an innovative method that enhances forecasting accuracy.

With this joint forecasting paradigm, we not only achieve state-of-the-art performance but also uncover an intriguing output scaling effect: the longer the overall forecast duration, the greater the predictive accuracy for outputs of a fixed length. We refer to this phenomenon as the delayed chain-of-thought (DCoT). By incorporating a non-causal bidirectional approach, tokens on the right can influence those on the left. Unlike traditional causal chain-of-thought (CoT) methods, our model allows for thoughts and reasoning tokens to be positioned after the final answers or targets. This newly discovered scaling effect significantly enhances the model's predictive power, as demonstrated in Figure 4. Figure 1 compares the joint forecasting paradigm with direct and recursive forecasting approaches.

Furthermore, we propose a multi-input ensemble method to address challenges related to input sequence length in time series forecasting. Extended lookbacks may better capture low-frequency patterns, while shorter lookbacks might be more effective for high-frequency patterns. By employing multi-input ensembling during inference, we aim to enhance prediction accuracy and reduce forecasting variance.

In a nutshell, our contributions can be summarized as follows:

1. We introduce a novel joint forecasting framework for time series prediction, distinct from the conventional direct and recursive forecasting approaches. Our framework unveils an unexpected and innovative scaling phenomenon: output scaling. We further develop a delayed chain-of-thought (DCoT) method to exploit this effect. To the best of our knowledge, this is the first work to show this scaling effect, connected to the COT process. The DCoT method significantly enhances the performance of our model, achieving an improvement exceeding 10.5% in MASE on the GIFT-Eval benchmark. as shown in figure 4.
2. We present YINGLONG, a flexible encoder-only architecture for time series forecasting utilizing a masking token during training. Our approach stems from the premise that time series forecasting shares greater similarity with natural language understanding (NLU) rather than natural language generation (NLG). YINGLONG achieves exceptional performance across a variety of zero-shot forecasting tasks on both traditional ETT and weather datasets. In our evaluations using the GIFT-Eval benchmarks, which include 23 datasets, we outperformed all baseline methods in CRPS and ranking metrics by a large margin.
3. We propose an innovative ensemble method that leverages mirror symmetry with inputs of varying lengths. This approach serves as a "free lunch," boosting performance with n-time inference compared to traditional single inference.

2 Related Works

Time Series Forecasting. In recent years, deep learning models have significantly advanced time series forecasting, primarily categorized into: (1) *univariate models*, like TFT (Lim et al., 2021), DeepAR (Salinas et al., 2020), and N-BEATS (Oreshkin et al., 2020b), focused on single time series, and (2) *multivariate models* that include transformer-based techniques (Wu et al., 2021; Zhou et al., 2022c; Nie et al., 2023b; Liu et al., 2024a) and others (Sen et al., 2019; Zhou et al., 2022b). Although these models excel within their training areas, the quest continues for pretrained models like LLMs with powerful zero-shot generalization capabilities.

Time Series Foundation Models. Universal forecasting approaches with foundational time series models can be divided into: (1) *Encoder-only models*, such as Moirai (Woo et al., 2024), Moment (Goswami et al., 2024) and VisionTS Chen et al. (2024) utilizing masked reconstruction, and TabPFN (Hoo et al., 2025) encoding to tabular data; (2) *Encoder-decoder models*, like Chronos (Ansari et al., 2024) and TTMS Ekambaram et al. (2024) with T5 and MLP architectures respectively; and (3) *Decoder-only models*, including TimesFM (Das et al., 2024) Lag-Llama (Rasul et al., 2023) and Timer (Liu et al., 2024c). Our work follows the encoder-only approach, highlighting the efficiency of the masked reconstruction learning framework and its capability for inference scaling, leading to superior zero-shot forecasting accuracy through self-consistency ensembles. We argue that time series forecasting aligns more with tasks requiring bidirectional consistency, unlike decoder-based generative tasks handled by LLMs (Ni et al., 2021).

Chain-of-Thought (CoT) Reasoning. Chain-of-thought refers to methods generating intermediate reasoning before deriving a final answer, including LLM prompting (Yang et al., 2024; Wei et al., 2022) and reasoning chain training (Zhou et al., 2022a). CoT enhances model expressivity (Feng et al., 2024) by feeding generated outputs back as inputs, yet its autoregressive nature limits complex reasoning tasks requiring planning (Xie et al., 2024).In our work, we utilize a bidirectional attention encoder-only framework for time series forecasting, effectively leveraging latent COT reasoning to overcome the autoregressive constraints and discrete language space limitations commonly found in large language models (LLMs).

3 Methods

Preliminary We consider the univariate forecasting setting. Let $x_1, x_2, \ldots, x_T$ be the historical observations from time $t = 1$ to $t = T$. Our goal is to predict the next P time points. We denote by $\hat{\boldsymbol{x}}_{T+i}$ the predicted statistics and by x_{T+i} the real observation at time $T + i$, for $i = 1, 2, \ldots, P$. Without loss of generality, we refer to the predicted statistics as α-quantiles, for some $\alpha \in (0, 1)$. For probabilistic forecasting models, the negative log-likelihood loss functions for direct and recursive paradigms can be compactly expressed as:

$$\text{Direct: } \min_{\theta} -\mathbb{E}\left[\sum_{k=j+1}^{T-1} \log f_\theta(x_k \mid x_{1:j})\right], \quad \text{Recursive: } \min_{\theta} -\mathbb{E}\left[\sum_{j=1}^{T-1} \log f_\theta(x_{j+1} \mid x_{1:j})\right] \quad (1)$$

where $x_{1:j} = \{x_1, \ldots, x_j\}$. This formulation highlights the structural similarity between paradigms: direct forecasting optimizes multi-step predictions, while recursive forecasting optimizes single-step predictions with autoregressive conditioning. The negative log-likelihood ensures probabilistic calibration by minimizing divergence between predictions and observations.

Although prevalent in time series modeling, both paradigms have limitations. Recursive forecasting focuses on transitions between observations but can accumulate errors over long horizons. Direct forecasting uses independent models for each horizon, enabling parallel predictions but sacrificing scalability and inter-output correlation, potentially compromising temporal coherence.

A toy example for illustration: Consider a random walk starting from the origin: $x_0 = 0$, with $x_{t+1} = x_t + \epsilon$, where $\epsilon \sim (-1, 1)$ for $t = 1, 2, \ldots$. Given observations up to time m, we aim to forecast the next n points using mean square error (MSE) as the metric. The optimal forecasts here are $\hat{x}_{m+j} = x_m$ for $j = 1, 2, \ldots, n$.

Assume an ideal model f_θ perfectly learns this one-step process, i.e., $f_\theta(x_{t+1} \mid x_t) = x_t \pm 1$ with equal probability. Generating N sample paths with f_θ and averaging them provides the future point forecast, illustrating the anti-concentration phenomenon as follows:

Lemma 3.1. *Let $\{x^i_{m+1}, \ldots, x^i_{m+n}\}$ for $i = 1, \ldots, N$ be N sample paths drawn from f_θ starting at x_m. For $\epsilon \in (0, \sqrt{j/N})$, there exists a positive constant C such that*

$$\mathbb{P}\Big(\Big|\tfrac{1}{N}\sum_{i=1}^{N} x^i_{m+j} \;-\; x_m\Big| \;>\; \epsilon\Big) \;\geq\; C\Big(1 - \tfrac{\epsilon^2 N}{j}\Big)^2.$$

Lemma 3.1 shows that for any fixed number of sample paths, there is always a non-negligible probability that the average estimator will deviate significantly from its target. Moreover, as the time horizon j grows, the chance of such a large deviation also increases. This suggests that relying on an average of sample paths can be problematic over relatively long forecasting windows.

Joint forecasting enhances performance beyond time series forecasting, benefiting NLP and CV domains as well. Diffusion methods generate all image tokens concurrently, and recent autoregressive image models adopt joint token generation approaches Chang et al. (2023); Teng et al. (2024). Although language models are typically recursive, techniques like beam search Freitag & Al-Onaizan (2017) capture token correlations and compute joint probabilities, unlike the greedy methods used in time series forecasting. Recent studies show that simultaneous token generation improves performance Gloeckle et al. (2024). We propose that this paradigm may shape the future of time series forecasting.

4 Model Architecture

To overcome these limitations, we propose a joint forecasting framework using a non-autoregressive architecture that generates probabilistic predictions for all temporal horizons in parallel while preserving inter-horizon dependencies. One architectural solution is a bidirectional transformer with masked tokens (see Figure 2). We describe our model architecture in detail below. Notably, **a vanilla transformer still achieves top performance on the GIFT-Eval benchmark** (see structural ablation section), demonstrating that our Unet transformer design is not essential. **The key factor is the joint forecasting framework leveraging DCoT**.

4.1 Tokenization

We adopt the patching technique (Nie et al., 2023a) to convert the input time series $[x_1, \ldots, x_T]$ into tokens, where the patch length is P. The series is transformed into a matrix $\tilde{X} \in \mathbb{R}^{N \times P}$, where $N = \lfloor T/P + 1 \rfloor$. During training, we randomly exclude a fraction $\rho \in (0, 1)$ of the patches as forecast targets. In testing, we append these masked patches as placeholders, extending the sequence to X_{m} with corresponding indices $\mathcal{M}$.

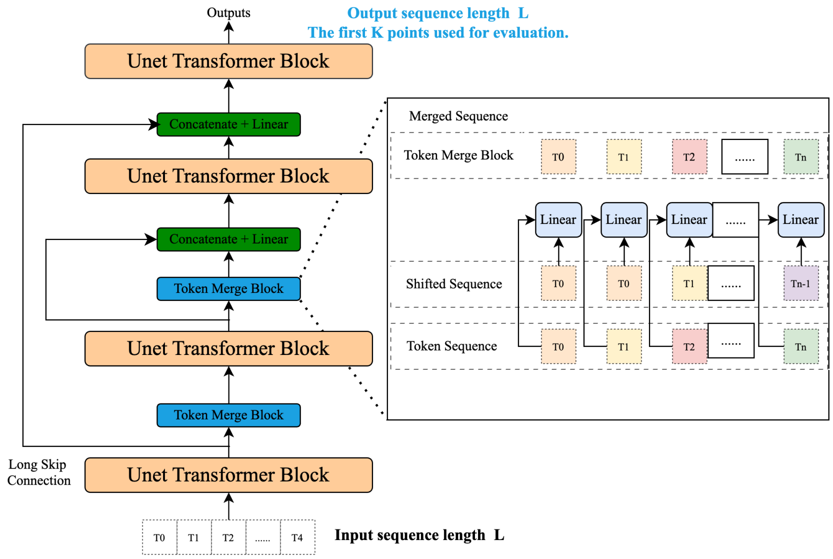


Figure 2: Illustration of the architecture of YINGLONG.

4.2 Unet-Transformer

For our transformer block, we adopt the standard architecture (Vaswani, 2017) with bidirectional attention. We use RMSNorm (Zhang & Sennrich, 2019) for pre-normalization, SwiGLU (Shazeer, 2020) as the activation function, and rotary positional embeddings (Su et al., 2024).

U-shaped design Inspired by U-shaped generative models (e.g., Bao et al. 2023; Zhang & Yan 2023), we incorporate a token merging module in the shallow layers and introduce long skip connections from shallow to deep layers. This design allows the network to process varying granularities of information (from coarse to fine), which is beneficial for point-level forecasting. The skip connections help propagate fine-grained information through the network, improving learning and performance.

Let $x^{i,j,\text{in}}, x^{i,j,\text{out}} \in \mathbb{R}^{1\times d}$ represent the input and output of the i-th token in the j-th transformer block, with $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, D$. Assuming D is even, after each shallow layer we generate coarse-grid tokens using a fully connected layer $F_c : 2d \to d$:

$$x^{i+1,j+1,\text{in}} \;=\; F_c\Big(\big[x^{i,j,\text{out}},\, x^{i+1,j,\text{out}}\big]\Big). \tag{2}$$

Moreover, before each deep layer, we apply another merging layer $F_m : 2d \to d$ to combine tokens from the corresponding shallow layer:

$$x^{i,j+1,\text{in}} \;=\; F_m\Big(\big[x^{i,j,\text{out}},\, x^{i,D-j+1,\text{out}}\big]\Big). \tag{3}$$

4.3 Output Layer and Loss Function

In this work, we directly predict R quantiles using a set of fully connected layers $F_{\alpha_k} : d \to P$ for each $\alpha_k \in (0, 1)$, $k = 1, 2, \ldots, R$:

$$q^i_{\alpha_k} \;=\; F_{\alpha_k}\Big(x^{i,D,\text{out}}\Big) \;\cdot\; \big(\sigma_{X_\text{m}} + \epsilon\big) \;+\; \mu_{X_\text{m}}, \tag{4}$$

where $q^i_{\alpha_k}$ is the predicted α_k-quantile for token i.

Weighted quantiles loss We adopt a weighted quantiles loss (WQL), leading to the following optimization objective:

$$\min_\theta \;\mathbb{E}\left[\sum_{i\in\mathcal{M}}\sum_{k=1}^{R}\sum_{s=1}^{P} w_{\alpha_k,x^i}\,\ell_{\alpha_k}\big(q^{i,s}_{\alpha_k},\, x^{i,s}\big)\right], \tag{5}$$

where $q^{i,s}_{\alpha_k}$ and $x^{i,s}$ denote the s-th elements of $q^i_{\alpha_k}$ and the ground-truth patch x^i, respectively. Here, ℓ_{α_k} is the standard α_k-quantile loss, and $w_{\alpha_k,x^i} > 0$ is a weight parameter that balances the contribution of each individual quantile loss term.

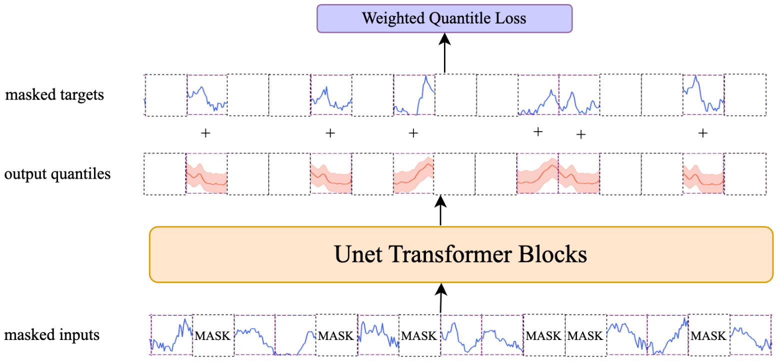


Figure 3: Training with Masked Token Prediction.

5 Output Scaling

5.1 Delayed Chain of Thoughts

Beyond the efficient direct prediction scheme described earlier, our approach also facilitates a new form of *chain of thoughts* (CoT) for time series forecasting. Chain of thoughts has been extensively studied in the field of natural language processing (NLP), where generating additional tokens (i.e., “thoughts”) can significantly improve model performance. In a classic NLP setting, CoT is obtained by training on sequences of the form [*prompt, CoT, target*].

In time series forecasting, future data is typically unavailable, rendering conventional Chains of Thought (CoT), which position intermediate tokens before the target, impractical. Historical observations function as the "prompt," directly preceding the forecasting targets. We propose Delayed CoT (DCoT), wherein time points beyond the forecasting horizon form the chain of thoughts. This approach exploits the periodic behavior and low-frequency structures inherent in time series data, allowing certain future points to be more predictable and serve as conditional anchors for target forecasting.

Probabilistic Interpretation of CoT Consider the probabilistic viewpoint of a classical CoT sequence:

$$\text{prompt},\ \text{CoT}_1,\ \ldots,\ \text{CoT}_n,\ \text{target}.$$

With n intermediate CoT steps, improved accuracy is expressed as:

$$\mathbb{P}\big(\text{target} = \text{truth} \mid \text{prompt},\ \text{CoT}_1, \ldots, \text{CoT}_n\big) \geq \mathbb{P}\big(\text{target} = \text{truth} \mid \text{prompt}\big).$$

This suggests that CoT tokens provide additional information, thereby increasing the likelihood of correctly predicting the target. In unidirectional language models, such auxiliary information appears before the target. However, when using a bidirectional model:

$$\text{prompt},\ \text{target},\ \text{CoT}_1,\ \ldots,\ \text{CoT}_n,$$

can still maintain the same conditional probability structure. Hence, positioning CoT after the target is a delayed CoT (DCoT).

5.2 Multi-Input Ensemble

A challenge in multi-horizon forecasting lies in optimizing the input window length—our preliminary experiments identify an accuracy tradeoff where short inputs enhance immediate predictions but degrade long-horizon performance, while extended windows exhibit the inverse behavior. To mitigate these horizon-specific limitations, we introduce an input-length ensemble that adaptively combines forecasts from varying temporal contexts. Furthermore, we implement temporal mirroring: reversing input sequences while correspondingly flipping prediction targets, then aggregating outputs through ensemble averaging to exploit bidirectional temporal patterns and enhance forecast stability.

Concretely, let $x_{1:n}$ denote the time series over the total lookback window of length n. We select k indices $1 = n_1 < n_2 < \cdots < n_k < n$ for shorter lookback windows. The final prediction y is then given by:

$$y \;=\; \frac{1}{2k}\,\sum_{j=1}^{k}\Big[\mathrm{Model}\big(x_{n_j:n}\big)\;-\;\mathrm{Model}\big(-\,x_{n_j:n}\big)\Big]. \tag{6}$$

While existing ensemble approaches combine multiple models trained under varied configurations Oreshkin et al. (2020a), our method achieves enhanced forecasting robustness through input-space bootstrapping—applying sign inversion and variable-length sampling to a single pre-trained model. Crucially, our architecture generates all horizon predictions via parallel computation in one forward pass (eliminating sample-path averaging), maintaining computational efficiency even when implementing the ensemble strategy in (6).

6 Experiments

6.1 Datasets and Training Details

In the training phase, we leverage a subset of datasets from the Monash Forecasting Repository Godahewa et al. (2021), the 5.625° WeatherBench Rasp et al. (2020), and a subset of data in Ansari et al. (2024), which altogether comprise approximately 78 billion time points. Recent studies (e.g., Lin et al. 2024; Woo et al. 2024; Liu et al. 2024b) note the critical impact of pretraining dataset size on overall model performance. However, in this work we do not specifically explore that aspect; rather, we aim to keep training costs low for efficiency.

We evaluate models on various tasks. For point forecasting, we use the ETT datasets Zhou et al. (2021) (two hourly series and two 15-minute series) and a 10-minute weather forecasting dataset Wetterstation. For broader generalization and probabilistic forecasting, we employ the GIFT-Eval benchmarks Aksu et al. (2024), which cover 23 different datasets. To avoid data leakage, we carefully exclude any data from our training sets that appears in the GIFT-Eval benchmarks. For baseline results, we utilize published data from the GIFT-Eval benchmark.

6.2 Training Details

We train four models with parameter sizes ranging from 7M to 300M, each for 100,000 steps. The batch size is 512, and the maximum sequence length is 8192. We use a patch size of 32, resulting in 256 tokens per sequence, and set the random masking ratio ρ to 0.2. Our optimizer is AdamW with a learning rate of 1×10^{-4}, weight decay 0.1, $\beta_1 = 0.9$, and $\beta_2 = 0.95$. The learning rate schedule includes a linear warmup over 2,000 steps followed by cosine annealing. All training is performed on eight NVIDIA A100 at BF16 precision. Refer to Section B in the Appendix for further details.

6.3 Zero-shot Forecasting

We compare our models against popular end-to-end approaches (FEDformer, TimesNet, DLinear, PatchTST) and recently proposed foundation models (Moirai, TimesFM, Chronos, TimeMoE, VisionTS, and Moment). For end-to-end baselines, we report the best performance from existing literature, while for foundation models we provide the best zero-shot results across all scales. Our model employs a forecasting window of 4096 and ensembles multiple input lengths (512, 1024, 2048, and 4096) for robust predictions. We assess performance using mean squared error (MSE) and mean absolute error (MAE).

Table 1 summarizes our performance across four model sizes. YINGLONG$_{\mathrm{110m}}$and YINGLONG$_{\mathrm{300m}}$ rank highest, with YINGLONG$_{\mathrm{110m}}$ achieving top 2 results in 70% of MSE and 75% of MAE cases.YINGLONG$_{\mathrm{300m}}$ also performs well with 60% and 90% in MSE and MAE respectively. In complex weather tasks, we observe a scaling law in model size, while in simpler tasks, YINGLONG$_{\mathrm{110m}}$ and YINGLONG$_{\mathrm{300m}}$ perform similarly, likely due to a low signal-to-noise ratio. Our smallest model, YINGLONG$_{\mathrm{6m}}$, efficiently achieves average ranks of 5.45 (MSE) and 5.00 (MAE), outperforming foundation models 30 times its size. For a more detailed result, please refer to Appendix Table 6.

6.4 Generalization Across Diverse Datasets

We recognize that ETT series and weather datasets may be insufficient for a truly comprehensive evaluation, which is a subject of ongoing debate in the time series community. Consequently, we adopt GIFT-Eval Aksu et al. (2024) as a benchmark. It includes 23 datasets from domains such

Table 1: Zero-shot forecasting experiments for another set of benchmarks. A lower MSE or MAE indicates a better prediction. TimesFM, due to its use of Weather datasets in pretraining, is not evaluated on this dataset and is denoted by a dash (−). **Red**: the best, Blue: the 2nd best.

	Ours								Foundation Models												End-to-end Models			
Models	YINGLONG$_{\text{6m}}$		YINGLONG$_{\text{50m}}$		YINGLONG$_{\text{110m}}$		YINGLONG$_{\text{300m}}$		Moirai		TimesFM		Moment		visionTS		Chronos		TimeMoE		Dlinear		PatchTST	
Metrics	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
ETTh1	0.408	0.412	0.405	0.408	0.399	0.408	0.398	**0.407**	0.415	0.418	0.473	0.444	0.684	0.566	**0.390**	0.414	0.524	0.455	0.393	0.417	0.423	0.437	0.413	0.431
ETTh2	0.344	0.382	0.339	0.370	**0.330**	**0.366**	0.337	0.370	0.359	0.377	0.392	0.406	0.362	0.410	0.333	0.375	0.398	0.411	0.362	0.399	0.456	0.445	**0.330**	0.379
ETTm1	0.370	0.368	0.367	0.366	**0.347**	**0.356**	0.351	0.358	0.407	0.385	0.433	0.418	0.670	0.536	0.374	0.372	0.555	0.465	0.356	0.392	0.362	0.379	0.351	0.381
ETTm2	0.250	0.302	0.250	0.298	**0.246**	**0.296**	0.249	**0.296**	0.303	0.337	0.328	0.346	0.316	0.365	0.282	0.321	0.295	0.338	0.288	0.344	0.356	0.331	0.255	0.315
Weather	0.239	0.268	0.231	0.257	0.227	0.255	**0.217**	**0.245**	0.264	0.273	-	-	0.294	0.326	0.269	0.292	0.279	0.306	0.256	0.289	0.240	0.300	0.226	0.264
Rank	5.45	5.00	4.75	2.90	**2.30**	**1.60**	2.40	1.65	9.25	6.90	11.50	10.50	11.80	12.35	5.45	6.25	11.20	10.65	5.85	7,9	7.45	9.15	4.20	6.25

Table 2: Results on GIFT-Eval (Best: **Red**, Second: Blue)

Model	MASE	CRPS	Rank	Model	MASE	CRPS	Rank	Model	MASE	CRPS	Rank
Foundation Models				**Foundation Models**				**End-to-end Models**			
YINGLONG$_{\text{300m}}$	0.716	**0.463**	**7.38**	Chronos$_{\text{base}}$	0.786	0.550	14.84	Crossformer	2.310	1.383	21.47
YINGLONG$_{\text{110m}}$	0.726	0.471	8.04	TimesFM	0.967	0.575	15.09	N-BEATS	0.842	0.689	21.92
Chronos-bolt$_{\text{base}}$	0.725	0.485	8.61	Chronos$_{\text{small}}$	0.800	0.560	15.85	Lag-Llama	1.101	0.744	22.34
YINGLONG$_{\text{50m}}$	0.738	0.479	8.74	VisionTS	0.775	0.638	21.02	DLinear	0.952	0.714	23.56
TabPFN-TS	0.748	0.480	8.88	TTMs	0.969	0.753	24.04	**Statistical Models**			
TimesFM2$_{500m}$	**0.680**	0.465	8.90	Timer	1.019	0.820	25.67	Auto Arima	0.964	0.770	21.91
Chronos-bolt$_{\text{small}}$	0.738	0.487	9.36	**End-to-end Models**				Auto Theta	0.978	1.051	24.34
Moirai$_{\text{large}}$	0.785	0.506	10.71	PatchTST	0.762	0.496	10.61	Auto ETS	1.088	6.327	25.21
Moirai$_{\text{base}}$	0.809	0.515	10.83	iTransformer	0.802	0.524	11.90	Seasonal Naive	1.00	1.00	26.36
YINGLONG$_{\text{6m}}$	0.790	0.515	12.55	TFT	0.822	0.511	12.03	Naive	1.260	1.383	28.24
Moirai$_{\text{small}}$	0.849	0.549	14.17	TIDE	0.980	0.652	19.22				
Chronos$_{\text{large}}$	0.781	0.547	14.74	DeepAR	1.206	0.721	19.43				

as economics, energy, healthcare, nature, sales, transportation, and cloud operations, and evaluates all major foundation time series models, end-to-end methods, and statistical techniques, mitigating concerns about generalization.

Our testing further incorporates five statistical approaches: Naive and Seasonal Naive Hyndman (2018), Auto ARIMA, Auto ETS, and Auto Theta Garza et al. (2022). For foundation models, we consider Chronos, Moirai, TimesFM, TTMS, VisionTS, Timer, Lag-Llama, and TabPFN-TS. We also benchmark end-to-end methods including PatchTST, TFT, Tide Das et al. (2023), DeepAR Salinas et al. (2020), Crossformer Zhang & Yan (2023), N-BEATS, and DLinear Zeng et al. (2023)

GIFT-Eval performance is reported in terms of mean absolute scaled error (MASE) and continuous ranked probability score (CRPS). Table 2 presents the geometric averages of these metrics. Notably, two of our models achieved top rankings, with average ranks of 7.38 and 8.04, respectively, highlighting their robustness and generalizability across diverse datasets. Furthermore, with respect to the MASE and CRPS metrics, our 300M model clearly outperforms the recent TabPFN-TS model, showing significant improvements of 4.3% and 3.5%, respectively. It's worth mentioning TabPFN-TS ranks at the top among all baselines, excluding our models, while our participation propels Chronos-Bolts ahead. Moreover, we observe scaling-law improvements from 50M to 300M parameters. Our largest 300M model is similar in size to other models like Morirai (311M), Moment (385M), and Chronos (710M), and is much smaller than Time-MoE as shown in table 2. Our performance gains are thus not solely driven by parameter scaling. Detailed results can be found in the appendix, with data aggregated by domain, prediction length, frequency, and multivariate settings.

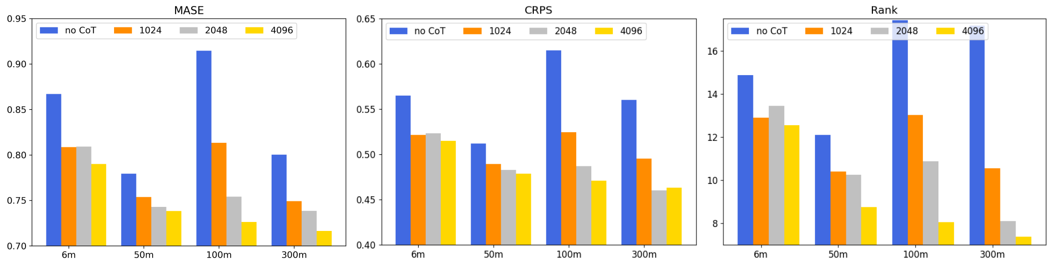


Figure 4: Ablation on DCoT. Four different output lengths up to 4096 are considered. In no CoT setting, we only output the target length. The MASE, CRPS,RANK of Gift-Eval are reported.

6.5 Influence of DCoT

We perform ablation studies to demonstrate the effectiveness of DCoT by varying its length, as illustrated in Figure 4. DCoT clearly demonstrates the phenomenon of inference scaling, resulting in substantial performance improvements, including reductions of up to 10.1% in MASE and 11.9% in CRPS. These advancements are likely the primary contributors to YinLong's success.

General scaling observations We observe substantial improvements across all pretrained YINGLONG models (6M–300M). Interestingly, the no-COT baseline model does not strictly follow the usual scaling trends; for example, the 50M model outperforms its 6M, 100M, and 300M counterparts in some cases. This behavior is not uncommon in time series foundation models, where scaling laws are less consistent than in large language models (LLMs). However, by integrating a sufficiently long DCoT and leveraging the output scaling effect, we achieve a more robust scaling relationship. For instance, using $\text{YINGLONG}_{\mathrm{300m}}$ with a 4096-length DCoT reduces MASE by 10.5% (from 0.800 to 0.716),and improves its average rank on GIFT-Eval from 17.2 to 7.38. By contrast, $\text{YINGLONG}_{\mathrm{100m}}$ with the same DCoT length yields only a 5.3% MASE drop (from 0.779 to 0.738). These findings suggest that DCoT has a stronger impact on larger, more robust baseline models.

Output Scaling for Other Model: We investigated the output scaling effect in vanilla transformer models ranging from 6M to 300M parameters using our joint forecasting paradigm (see Appendix Figure 7). Applying DCoT with a token length of 4096 significantly improved the 300M model's MASE and CRPS by 24.9% and 30.0%, respectively, compared to a non-DCoT setup. This scaling effect persisted across DCOT lengths and various model sizes. validated on the GIFT-Eval benchmark with 23 datasets, these results demonstrate that the scaling effect is robust across different model architectures and datasets.

Effect of DCoT length. As DCoT length grows from 1k to 4k, the performance gains become more pronounced and consistent. For instance, $\text{YINGLONG}_{\mathrm{300m}}$ achieves MASE reductions of 6.4% and 7.7% with DCoT lengths of 1024 and 2048, respectively, compared to a 10.5% reduction at 4096 length. This trend is evident across various model sizes, indicating that for a fixed test horizon, extending the output sequence enhances accuracy. This improvement is attributed to delayed CoT interactions affecting earlier outputs, a phenomenon specific to our joint forecasting paradigm. In contrast, direct/recursive forecasting do not exhibit this effect, as extending the output sequence does not impact prior outputs.

Structural Ablation We conducted a structure ablation on the uTransformer block and the token merge design, as shown in the Appendix Figure 6, resulting in moderate performance improvements over the standard transformer block. We compared the YINGLONG-300M model to vanilla transformer models ranging from 6M to 300M parameters, all trained using our joint forecasting paradigm (see Appendix Figure 7). YINGLONG-300M reduced the MASE from 0.726 to 0.716 and the CRPS from 0.473 to 0.463 compared to the 300M transformer baseline. Even vanilla transformer with joint forecasting performs comparably to or better than previous SOTA foundation models, including TabPFN-TS (MASE=0.748, CRPS=0.480) and Chronos-bolt (MASE=0.725, CRPS=0.471).

Error Reduction Pattern Our analysis reveals that the primary error reduction achieved by DCoT is due to a decrease in trend error. As shown in Appendix Table 12 and Figure 5, both absolute and relative reductions in MSE are attributable to trend rather than changes in the seasonal component.

Influence of Input Ensemble We compare the average of multiple input-length forecasts with the single-best component in Table 13 in Appendix. Our input-ensemble strategy improves accuracy by 1%–4% without training additional models. This represents a scalable "free lunch" approach.

7 Conclusion and Future Work

Our study presents a novel joint forecasting paradigm for time series forecasting, distinguishing itself from traditional direct and recursive approaches by incorporating masked patch placeholders and non-causal, bidirectional attention mechanisms. This approach uncovers a previously unreported scaling phenomenon: longer outputs appear to enhance model accuracy due to a delayed chain-of-thought reasoning process. Comprehensive evaluations using the GIFT-Eval benchmark confirm that YingLong consistently delivers superior results across various metrics, outperforming SOTA methods. Compared to LLMs, we face interpretability challenges: LLMs make reasoning explicit through textual tokens, while our delayed thought process remains hidden in latent space. Future work will focus on better interpreting and understanding COT within time series models.

References

Aksu, T., Woo, G., Liu, J., Liu, X., Liu, C., Savarese, S., Xiong, C., and Sahoo, D. Gift-eval: A benchmark for general time series forecasting model evaluation. *arxiv preprint arxiv:2410.10393*, 2024.

Alexandrov, A., Benidis, K., Bohlke-Schneider, M., Flunkert, V., Gasthaus, J., Januschowski, T., Maddix, D. C., Rangapuram, S., Salinas, D., Schulz, J., Stella, L., TÃ¼rkmen, A. C., and Wang, Y. Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020. URL `http://jmlr.org/papers/v21/19-820.html`.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22669–22679, 2023.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Chen, M., Shen, L., Li, Z., Wang, X. J., Sun, J., and Liu, C. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*, 2024.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., and Yu, R. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Ekambaram, V., Jati, A., Dayama, P., Mukherjee, S., Nguyen, N. H., Gifford, W. M., Reddy, C., and Kalagnanam, J. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *CoRR*, 2024.

Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., and Wang, L. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

Freitag, M. and Al-Onaizan, Y. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*, 2017.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Garza, F., Canseco, M. M., Challú, C., and Olivares, K. G. Statsforecast: Lightning fast forecasting with statistical and econometric models. *PyCon: Salt Lake City, UT, USA*, 2022.

Gloeckle, F., Idrissi, B. Y., Rozière, B., Lopez-Paz, D., and Synnaeve, G. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024.

Godahewa, R. W., Bergmeir, C., Webb, G. I., Hyndman, R., and Montero-Manso, P. Monash time series forecasting archive. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL `https://openreview.net/forum?id=wEc1mgAjU-`.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models. In *Forty-first International Conference on Machine Learning*, 2024.

Hoo, S. B., Müller, S., Salinas, D., and Hutter, F. The tabular foundation model tabpfn outperforms specialized time series forecasting models based on simple features. *arXiv preprint arXiv:2501.02945*, 2025.

Howard, A., Yui, H., McDonald, M., and Cukierski, W. Recruit restaurant visitor forecasting. `https://kaggle.com/competitions/recruit-restaurant-visitor-forecasting`, 2017. Kaggle.

Howard, A., inversion, Makridakis, S., and vangelis. M5 forecasting - accuracy. `https://kaggle.com/competitions/m5-forecasting-accuracy`, 2020. Kaggle.

Hyndman, R. *Forecasting: principles and practice*. OTexts, 2018.

Lai, G., Chang, W.-C., Yang, Y., and Liu, H. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., and Ping, W. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., Pan, S., and Wen, Q. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6555–6565, 2024.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

Lin, B., Tang, Z., Ye, Y., Cui, J., Zhu, B., Jin, P., Zhang, J., Ning, M., and Yuan, L. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a.

Liu, Y., Qin, G., Huang, X., Wang, J., and Long, M. Autotimes: Autoregressive time series forecasters via large language models. *arXiv preprint arXiv:2402.02370*, 2024b.

Liu, Y., Zhang, H., Li, C., Huang, X., Wang, J., and Long, M. Timer: Generative pre-trained transformers are large time series models. In *Forty-first International Conference on Machine Learning*, 2024c.

Mancuso, P., Piccialli, V., and Sudoso, A. M. A machine learning approach for forecasting hierarchical time series. *Expert Systems with Applications*, 182:115102, 2021.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

Ni, J., Abrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., and Yang, Y. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023a. URL `https://openreview.net/forum?id=Jbdc0vTOcol`.

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

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020a. URL `https://openreview.net/forum?id=r1ecqn4YwB`.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020b.

Palaskar, S., Ekambaram, V., Jati, A., Gantayat, N., Saha, A., Nagar, S., Nguyen, N. H., Dayama, P., Sindhgatta, R., Mohapatra, P., et al. Automixer for improved multivariate time-series forecasting on bizitops data. *arXiv preprint arXiv:2310.20280*, 2023.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

Rasul, K., Ashok, A., Williams, A. R., Khorasani, A., Adamopoulos, G., Bhagwatkar, R., Biloš, M., Ghonia, H., Hassen, N. V., Schneider, A., Garg, S., Drouin, A., Chapados, N., Nevmyvaka, Y., and Rish, I. Lag-llama: Towards foundation models for time series forecasting, 2023.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

Sen, R., Yu, H.-F., and Dhillon, I. S. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. *Advances in neural information processing systems*, 32, 2019.

Shazeer, N. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Shen, S., Van Beek, V., and Iosup, A. Statistical characterization of business-critical workloads hosted in cloud datacenters. In *2015 15th IEEE/ACM international symposium on cluster, cloud and grid computing*, pp. 465–474. IEEE, 2015.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Teng, Y., Shi, H., Liu, X., Ning, X., Dai, G., Wang, Y., Li, Z., and Liu, X. Accelerating autoregressive text-to-image generation with training-free speculative jacobi decoding. *arXiv preprint arXiv:2410.01699*, 2024.

Vaswani, A. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Wang, J., Jiang, J., Jiang, W., Li, C., and Zhao, W. X. Libcity: An open library for traffic prediction. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '21, pp. 145–148, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386647. doi: 10.1145/3474717.3483923. URL `https://doi.org/10.1145/3474717.3483923`.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wetterstation. Weather. URL `https://www.bgc-jena.mpg.de/wetter/`.

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

Wu, H., Xu, J., Wang, J., and Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

Xie, Y., Kawaguchi, K., Zhao, Y., Zhao, J. X., Kan, M.-Y., He, J., and Xie, M. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36, 2024.

Yang, S., Gribovskaya, E., Kassner, N., Geva, M., and Riedel, S. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*, 2024.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Zhang, B. and Sennrich, R. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, K., Wen, Q., Zhang, C., Cai, R., Jin, M., Liu, Y., Zhang, J. Y., Liang, Y., Pang, G., Song, D., et al. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Zhang, Y. and Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022a.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Zhou, T., Ma, Z., Wen, Q., Sun, L., Yao, T., Yin, W., Jin, R., et al. Film: Frequency improved legendre memory model for long-term time series forecasting. *Advances in neural information processing systems*, 35:12677–12690, 2022b.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. 39th International Conference on Machine Learning (ICML 2022)*, 2022c.

A Proof of Lemma 3.1

Proof. Since the sample paths of the random walk are martingale and generated independently, we have

$$\mathbb{E}[x^i_{m+j}] = x_m, \quad \mathbb{E}[(x^i_{m+j} - x_m)^2] = j.$$

Denote

$$Z_j = \left(\frac{1}{N} \sum_{i=1}^{N} x^i_{m+j} - x_m \right)^2.$$

It follows that

$$Z_j \geq 0, \quad \mathbb{E}[Z_j] = \frac{j}{N}, \quad \mathbb{E}[Z_j^2] \leq \mathcal{O}(j^2 N^{-2}).$$

Via the Paley–Zygmund inequality, for $\epsilon \in (0, 1)$ we have

$$\mathbb{P}(Z_j \geq \epsilon \mathbb{E}[Z_j]) \geq (1-\epsilon)^2 \cdot \frac{\mathbb{E}[Z_j]^2}{\mathbb{E}[Z_j^2]}.$$

Thus,

$$\mathbb{P}\left(\left| \frac{1}{N} \sum_{i=1}^{N} x^i_{m+j} - x_m \right| \geq \tilde{\epsilon} \right) \geq C \left(1 - \frac{\tilde{\epsilon}^2 N}{j} \right)^2,$$

where C is a positive constant and $\tilde{\epsilon} = \sqrt{\epsilon j / N}$. □

B Training Details

We summarize the details for pretraining datasets in Table 3. The weights in (5) is set as $w_{\alpha_k, x^i} = 1/(\sqrt{\alpha_k(1-\alpha_k)} \sum |x^{i,s}|)$, where we control the influence on the target scale and quantile estimation variance simultaneously. In Table 4 we present the model configurations. In order minimizing the influence of hyperparameter tuning, we keep all training settings the same except model size related ones.

C Supplement for zero-shot forecasting experiment

Table 5 summarize the statistics of datasets. We use Mean Absolute Error (MAE) and Mean Squared Error (MSE) as metrics. The full results are provided in Table 6. In particular,YINGLONG$_{100\mathrm{m}}$ and YINGLONG$_{300\mathrm{m}}$ reach 20 and 16 best results. Those performances beat other foundation models.

D Supplement for the GIFT-Eval Benchmark

For each GIFT-Eval experiment, we measure performance using mean absolute scaled error (MASE) and weighted quantile loss (WQL). Table 2 presents the geometric average of these metrics. Notably, three of our models achieve top scores in MASE, WQL, and Rank, demonstrating their robustness and generalizability across the diverse datasets.Furthermore, we observe scaling law improvements from 50M to 300M. Our largest 300M model is comparable in size to Morirai 311M, Moment 385M, and Chronos 710M, yet significantly smaller than Time-MoE. Our improvements do not rely on increasing parameter size through scaling laws. A more detailed performance analysis, aggregated by domain, is presented in Table 8. YinLong demonstrates robust performance across various domains, exhibiting leading results in Energy, Nature, and Transportation, with rank improvements from 9.2 to 4.3, 4.6 to 3.8, and 6.9 to 5.2, respectively. Additionally, YinLong achieves top performance in CloudOps and Sales. Interestingly, in business-heavy domains such as CloudOps and Sales, the end-to-end model outperforms the time series foundation model. This may be attributed to their extensive within-channel interactions.

We provide a summary on the datasets of GIFT-Eval in Table 7. We consider mean absolute scaled error (MASE), continuous ranked probability score (CRPS) and Rank as metrics.

$$\mathrm{MASE} = \frac{m-s}{n} \cdot \frac{\sum_{t=m+1}^{m+n} |\hat{x}_t - x_t|}{\sum_{t}^{m-s} |x_t - x_{t+s}|},$$

where m in the lookback length, n is the forecasting length, and s is the seasonality parameter.

In this work, we follow the setting in Aksu et al. (2024) and use weighted quantile loss (WQL) to approximate CRPS as follows

$$\mathrm{CRPS} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{WQL}[\alpha_k]$$

$$\mathrm{WQL}[\alpha] = 2\frac{\sum_{t=m+1}^{m+n} l_\alpha(q_t(\alpha), x_t)}{\sum_{t=m+1}^{m+n} |x_t|},$$

where we set $K = 9$ and $alpha_k = k/10$ for $k = 1, 2, ..., K$.

Readers may refer Aksu et al. 2024 for more details. The addition results aggregated in prediction lengths, frequency and number of variates are given in Table 9, 10 and 11.

Table 3: Pretraining Datasets

Dataset	Source	domain	Frequency	# Length
Australian Electricity Demand	Godahewa et al. 2021	Energy	30T	1153584
Bitcoin	Godahewa et al. 2021	Econ/Fin	D	68927
Cif 2016	Godahewa et al. 2021	Econ/Fin	M	7108
Cif 2016	Godahewa et al. 2021	Econ/Fin	M	7108
Fred MD	Godahewa et al. 2021	Econ/Fin	M	71624
Kaggle Web Traffic Daily	Godahewa et al. 2021	Web/CloudOps	W	15206232
Kaggle Web Traffic Weekly	Godahewa et al. 2021	Web/CloudOps	D	332586145
London smart meters	Godahewa et al. 2021	Energy	30T	160041727
M1 Monthly	Godahewa et al. 2021	Econ/Fin	M	1047
M1 Quarterly	Godahewa et al. 2021	Econ/Fin	3M	9628
M1 Yearly	Godahewa et al. 2021	Econ/Fin	Y	3136
M3 Monthly	Godahewa et al. 2021	Econ/Fin	M	538
M3 Quarterly	Godahewa et al. 2021	Econ/Fin	3M	36960
M3 Yearly	Godahewa et al. 2021	Econ/Fin	Y	18319
NN5 Daily	Godahewa et al. 2021	Econ/Fin	D	35303
Pedestrian Counts	Godahewa et al. 2021	Transport	H	3125914
Sunspot	Godahewa et al. 2021	Nature	D	45312
Tourism Monthly	Godahewa et al. 2021	Econ/Fin	M	98867
Tourism Quarterly	Godahewa et al. 2021	Econ/Fin	Q	39128
Tourism Yearly	Godahewa et al. 2021	Econ/Fin	Y	11198
Traffic Hourly	Godahewa et al. 2021	Transport	H	14858016
Traffic Weekly	Godahewa et al. 2021	Transport	W	78816
Oikolab Weather	Godahewa et al. 2021	Nature	H	615574
Wind Power	Godahewa et al. 2021	Energy	4s	7397147
Wind Farm	Godahewa et al. 2021	Energy	T	172165370
M5	Howard et al. 2020	Sales	D	58327370
Wiki Daily	Ansari et al. 2024	Web/CloudOps	D	247099892
USHCN	Ansari et al. 2024	Nature	D	235396770
Uber TLC Daily	Alexandrov et al. 2020	Transport	D	42533
Uber TLC Hourly	Alexandrov et al. 2020	Transport	H	510284
Weatherbench Hourly	Rasp et al. 2020	Nature	H	74630250518
Weatherbench Daily	Rasp et al. 2020	Nature	D	3223513345
Weatherbench Weekly	Rasp et al. 2020	Nature	W	462956049
KernelSynth	Ansari et al. 2024	Synthetic	-	3221225472

Table 4: Model Configurations

Config	YINGLONG$_{6m}$	YINGLONG$_{50m}$	YINGLONG$_{100m}$	YINGLONG$_{300m}$
Max Length	8192	8192	8192	8192
Layers	6	8	12	18
Heads	16	16	12	32
Embed Dim	256	512	768	1024
Intermediate Dim	1024	2048	3072	4096
Query per Groups	4	4	4	4
Pos Embed	Rope	Rope	Rope	Rope
Norm	RMSNorm	RMSNorm	RMSNorm	RMSNorm
MLP	SwiGLU	SwiGLU	SwiGLU	SwiGLU
Patch Size	32	32	32	32
Batch Size	512	512	512	512
Max Step	100000	100000	100000	100000
Warm-up Steps	2000	2000	2000	2000
Learning Rate	1e-4	1e-4	1e-4	1e-4
Scheduler	Cosine	Cosine	Cosine	Cosine
Weight Decay	0.1	0.1	0.1	0.1
β_1, β_2	0.9, 0.95	0.9, 0.95	0.9, 0.95	0.9, 0.95
Min LR	1e-5	1e-5	1e-5	1e-5
Mix Up Aug	0.2	0.2	0.2	0.2
Rescaling Aug	(1,4)	(1,4)	(1,4)	(1,4)
mask ratio ρ	0.2	0.2	0.2	0.2

Table 5: Dataset details for zero-shot forecasting experiment

Dataset	Source	Length	Dimension	Frequency
ETTm1	Zhou et al. 2021	69680	7	15 T
ETTm2	Zhou et al. 2021	69680	7	15 T
ETTh1	Zhou et al. 2021	17420	7	1H
ETTh2	Zhou et al. 2021	17420	7	1H
Weather	Wu et al. 2021	52696	21	10T

Table 6: Full results of zero-shot forecasting experiments. A lower MSE or MAE indicates a better prediction. TimesFM, due to its use of Weather datasets in pretraining, is not evaluated on this dataset and is denoted by a dash (−). **Red**: the best, Blue: the 2nd best.

Models		Ours								Foundation Models												End-to-end Models							
		YINGLONG$_{6m}$		YINGLONG$_{50m}$		YINGLONG$_{110m}$		YINGLONG$_{300m}$		Moirai		TimesFM		Moment		visionTS		Chronos		TimeMoE		FEDformer		TimesNet		Dlinear		PatchTST	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MAE	MAE
ETTh1	96	0.366	0.380	0.359	0.375	0.354	**0.372**	0.355	0.374	0.376	0.388	0.414	0.404	0.688	0.557	0.353	0.383	0.440	0.390	**0.349**	0.379	0.376	0.419	0.384	0.402	0.375	0.399	0.370	0.399
	192	0.406	0.405	0.404	0.403	0.397	**0.401**	0.396	**0.401**	0.412	0.413	0.465	0.434	0.688	0.560	**0.392**	0.410	0.492	0.426	0.384	0.404	0.420	0.448	0.436	0.429	0.405	0.416	0.413	0.421
	336	0.433	0.421	0.429	0.416	0.419	0.416	0.418	**0.415**	0.433	0.428	0.503	0.456	0.675	0.563	**0.407**	0.423	0.550	0.462	0.411	0.430	0.459	0.465	0.491	0.469	0.439	0.443	0.422	0.466
	720	0.428	0.444	0.427	0.438	0.427	0.444	0.422	**0.437**	0.439	0.444	0.511	0.481	0.683	0.585	**0.406**	0.441	0.615	0.543	0.427	0.455	0.506	0.507	0.521	0.500	0.472	0.490	0.447	0.466
	avg	0.408	0.412	0.405	0.408	0.399	0.408	0.398	**0.407**	0.415	0.418	0.473	0.444	0.684	0.566	**0.390**	0.414	0.524	0.455	0.393	0.417	0.440	0.460	0.458	0.450	0.423	0.437	0.413	0.431
ETTh2	96	0.273	0.326	0.275	0.319	**0.267**	**0.316**	0.271	0.319	0.294	0.330	0.315	0.349	0.342	0.396	0.271	0.328	0.307	0.343	0.292	0.352	0.358	0.397	0.340	0.374	0.289	0.353	0.274	0.336
	192	0.339	0.372	0.340	0.364	0.330	**0.360**	0.336	0.364	0.361	0.371	0.388	0.395	0.354	0.402	**0.328**	0.367	0.376	0.392	0.347	0.379	0.429	0.439	0.402	0.414	0.383	0.418	0.339	0.379
	336	0.370	0.399	0.366	0.388	0.354	0.382	0.364	0.388	0.370	0.390	0.422	0.427	0.356	0.407	0.345	0.381	0.408	0.430	0.391	0.418	0.497	0.487	0.452	0.452	0.448	0.465	**0.329**	**0.380**
	720	0.394	0.429	0.376	0.409	**0.369**	**0.407**	0.378	0.411	0.411	0.418	0.443	0.454	0.395	0.434	0.388	0.422	0.501	0.477	0.419	0.447	0.463	0.474	0.462	0.468	0.605	0.551	0.379	0.422
	avg	0.344	0.382	0.339	0.370	**0.330**	**0.366**	0.337	0.370	0.359	0.377	0.392	0.406	0.362	0.410	0.333	0.375	0.398	0.411	0.362	0.399	0.437	0.449	0.414	0.427	0.456	0.445	**0.330**	0.379
ETTm1	96	0.297	0.324	0.294	0.320	**0.281**	**0.313**	0.284	0.315	0.363	0.356	0.361	0.370	0.654	0.527	0.341	0.347	0.454	0.403	**0.281**	0.341	0.464	0.416	0.338	0.375	0.299	0.343	0.290	0.342
	192	0.346	0.355	0.344	0.352	**0.328**	**0.343**	0.332	0.346	0.388	0.375	0.414	0.405	0.662	0.532	0.360	0.360	0.530	0.450	0.305	0.358	0.426	0.441	0.371	0.387	0.355	0.365	0.332	0.369
	336	0.384	0.378	0.383	0.376	**0.362**	**0.366**	0.365	0.369	0.416	0.392	0.445	0.429	0.672	0.537	0.377	0.374	0.577	0.481	0.369	0.395	0.445	0.459	0.410	0.411	0.369	0.386	0.366	0.392
	720	0.441	0.414	0.449	0.416	0.418	**0.401**	0.423	0.404	0.460	0.418	0.512	0.471	0.692	0.551	0.416	0.405	0.660	0.526	0.469	0.472	0.543	0.490	0.478	0.450	0.425	0.421	**0.416**	0.420
	avg	0.370	0.368	0.367	0.366	**0.347**	**0.356**	0.351	0.358	0.407	0.385	0.433	0.418	0.670	0.536	0.374	0.372	0.555	0.465	0.356	0.392	0.448	0.452	0.400	0.406	0.362	0.379	0.351	0.381
ETTm2	96	0.163	0.240	0.162	0.235	0.161	**0.234**	**0.160**	**0.234**	0.205	0.273	0.202	0.270	0.260	0.335	0.228	0.282	0.197	0.271	0.197	0.286	0.203	0.287	0.187	0.267	0.167	0.260	0.165	0.255
	192	**0.218**	0.281	0.219	0.277	**0.218**	**0.276**	**0.218**	**0.276**	0.275	0.316	0.289	0.321	0.289	0.350	0.262	0305	0.254	0.314	0.235	0.312	0.269	0.328	0.249	0.309	0.224	0.303	0.220	0.292
	336	0.269	0.317	0.270	0.313	**0.266**	0.312	0.268	**0.311**	0.329	0.350	0.360	0.366	0.324	0.369	0.293	0.328	0.313	0.353	0.293	0.348	0.325	0.366	0.321	0.351	0.281	0.342	0.274	0.329
	720	0.350	0.371	0.351	0.367	**0.340**	**0.363**	0.348	0.366	0.402	0.408	0.462	0.430	0.394	0.409	0.343	0.370	0.416	0.415	0.427	0.428	0.421	0.415	0.497	0.403	0.397	0.421	0.362	0.385
	avg	0.250	0.302	0.250	0.298	**0.246**	**0.296**	0.249	**0.296**	0.303	0.337	0.328	0.346	0.316	0.365	0.282	0.321	0.295	0.338	0.288	0.344	0.305	0.349	0.391	0.333	0.356	0.331	0.255	0.315
Weather	96	0.169	0.240	0.159	0.195	0.153	0.190	**0.144**	**0.180**	0.198	0.211	-	-	0.243	0.255	0.220	0.257	0.194	0.235	0.157	0.211	0.217	0.396	0.172	0.220	0.152	0.237	0.149	0.198
	192	0.212	0.249	0.202	0.237	0.196	0.234	**0.186**	**0.223**	0.246	0.251	-	-	0.278	0.329	0.244	0.275	0.249	0.285	0.208	0.256	0.276	0.336	0.219	0.261	0.220	0.282	0.194	0.241
	336	0.258	0.285	0.250	0.275	0.246	0.273	**0.235**	**0.263**	0.274	0.291	-	-	0.306	0.375	0.280	0.299	0.302	0.327	0.255	0.290	0.339	0.380	0.280	0.306	0.265	0.319	0.245	0.282
	720	0.316	0.328	0.313	0.320	0.314	0.325	**0.302**	**0.313**	0.337	0.340	-	-	0.350	0.374	0.330	0.337	0.372	0.378	0.405	0.397	0.403	0.428	0.365	0.359	0.323	0.362	0.314	0.334
	avg	0.239	0.268	0.231	0.257	0.227	0.255	**0.217**	**0.245**	0.264	0.273	-	-	0.294	0.326	0.269	0.292	0.279	0.306	0.256	0.289	0.309	0.360	0.259	0.287	0.240	0.300	0.226	0.264
Average Rank		5.45	5.00	4.75	2.90	**2.30**	**1.60**	2.40	1.65	9.25	6.90	11.50	10.50	11.80	12.35	5.45	6.25	11.20	10.65	5.85	7,9	11.70	12.55	10.15	9.80	7.45	9.15	4.20	6.25

Table 7: GIFT-Eval datasets

Dataset	Source	domain	Frequency	# Series	# Tasks
Weather	Wu et al. 2021	Nature	10T,H,D	1	7
BizITObs	Palaskar et al. 2023	Web/CloudOps	10S,5T,H	24	12
Bitbrains	Shen et al. 2015	Web/CloudOps	5T,H	1750	8
Restaurant	Howard et al. 2017	Sales	D	807	1
ETT	Zhou et al. 2021	Energy	15T,H,D,W-THU	2	18
libcity	Wang et al. 2021	Transport	5T,H,D,15T,	518	14
Solar	Lai et al. 2018	Energy	10T,H,D,W-FRI	137	8
Hierarchical Sales	Mancuso et al. 2021	Salses	D,W-WED	118	2
M4	Godahewa et al. 2021	Econ/Fin	A-DEC,Q-DEC,M,W-SUN,D,H	100741	6
Hospital	Godahewa et al. 2021	Healthcare	M	767	1
COVID Death	Godahewa et al. 2021	Healthcare	D	226	1
US Births	Godahewa et al. 2021	Healthcare	D,W-TUE,M	1	3
Saugeen	Godahewa et al. 2021	Nature	D,W-THU,M	1	3
Temperature Rain	Godahewa et al. 2021	Nature	D	32072	1
KDD CUP 2018	Godahewa et al. 2021	Nature	H,D	270	4
Car Parts	Godahewa et al. 2021	Sales	M	2674	1
Electricity	Godahewa et al. 2021	Energy	15T,H,D,W-FRI	370	8

Table 8: Results on GIFT-Eval aggregated by domain. The table shows MASE,CRPS and Rank for each method.

Model	Econ/Fin			Energy			Healthcare			Nature			Sales			Transport			Web/CloudOps		
	MASE	CRPS	Rank	MASE	CRPS	Rank	MASE	CRPS	Rank	MASE	CRPS	Rank	MASE	CRPS	Rank	MASE	CRPS	Rank	MASE	CRPS	Rank
YINGLONG$_{300m}$	0.899	0.786	14	0.870	0.539	5.19	0.704	0.574	13.20	0.746	0.287	5.27	0.742	0.368	13.00	0.666	0.435	6.40	0.501	0.461	8.65
YINGLONG$_{100m}$	0.939	0.790	15.5	0.906	0.560	6.25	0.710	0.579	14	0.742	0.288	4.8	0.737	0.365	12.75	0.692	0.452	8.33	0.483	0.455	8.45
YINGLONG$_{50m}$	0.918	0.793	15.33	0.909	0.570	7.16	0.706	0.546	12.8	0.748	0.296	5.93	0.788	0.390	17.25	0.698	0.458	9.47	0.512	0.465	8.15
YINGLONG$_{6m}$	1.080	0.899	20	0.978	0.617	11.06	0.727	0.589	16.6	0.779	0.318	11.2	0.767	0.385	15.5	0.746	0.488	12.93	0.554	0.499	11.85
TabPFN-TS	0.802	0.717	7.33	0.900	0.564	8.53	0.554	0.411	3	0.780	0.297	6.2	0.713	0.351	9.25	0.682	0.447	9	0.618	0.552	13.2
chronos-bolt$_b$	0.799	0.717	6.67	0.845	0.550	7	0.691	0.532	12	0.667	0.265	3.07	0.693	0.344	4	0.686	0.474	9.93	0.624	0.591	15.00
chronos-bolt$_s$	0.816	0.700	6.00	0.864	0.564	8.19	0.671	0.511	11.80	0.704	0.284	4.67	0.696	0.347	6.25	0.692	0.474	11.67	0.627	0.559	14.05
TimesFM-V2$_{500m}$	0.641	0.546	2.67	0.890	0.574	8.81	0.597	0.455	5.80	0.624	0.284	10.67	0.699	0.342	4.75	0.645	0.432	5.73	0.515	0.518	13.55
Moirai$_{\mathrm{large}}$	0.845	0.732	8.33	1.025	0.63	11.53	0.700	0.534	11.40	0.750	0.31	8.87	0.711	0.364	10.75	0.601	0.39	8	0.666	0.591	13.35
Moirai$_{\mathrm{base}}$	0.906	0.786	9.83	0.998	0.623	11.38	0.683	0.518	9.8	0.771	0.314	8.87	0.694	0.346	5	0.637	0.413	9.07	0.745	0.617	14.5
Moirai$_{\mathrm{small}}$	0.986	0.785	13.17	1.069	0.656	13.78	0.849	0.703	20.8	0.807	0.335	12.47	0.731	0.361	11.75	0.732	0.476	14.47	0.673	0.614	15
Chronos$_{\mathrm{large}}$	0.783	0.758	11.17	0.919	0.628	13.47	0.599	0.446	6.20	0.813	0.364	16.53	0.724	0.362	13.25	0.714	0.512	15.33	0.675	0.647	18.5
Chronos$_{\mathrm{base}}$	0.783	0.751	10.17	0.924	0.631	13.63	0.645	0.485	8.8	0.823	0.366	17.27	0.726	0.363	13.5	0.712	0.512	15.13	0.676	0.651	17.95
Chronos$_{\mathrm{small}}$	0.797	0.763	12	0.947	0.648	15.34	0.607	0.496	9	0.851	0.383	18.4	0.733	0.366	15	0.737	0.530	17.93	0.678	0.629	16.2
TimesFM	0.824	0.716	8.67	1.016	0.673	15.06	0.698	0.652	12.4	0.880	0.333	13.07	0.700	0.344	5	0.741	0.510	14.2	1.419	0.739	21.95
Lag-Llama	2.910	1.839	30.33	1.392	0.923	23	1.621	1.599	29.4	0.932	0.385	18.2	0.841	0.442	20.5	0.842	0.572	19.73	0.753	0.734	22.55
TTMs	1.297	1.142	26.17	1.059	0.847	23.03	1.150	1.226	28.20	0.845	0.404	20.53	0.878	0.540	26.50	0.891	0.730	25.47	0.887	0.851	25.05
Timer	1.809	1.475	28.83	1.287	1.019	26.38	1.390	1.700	29.6	0.881	0.444	23.6	0.775	0.468	23	0.894	0.743	26.07	0.711	0.771	24.40
VisionTS	0.931	1.046	24.83	0.993	0.782	21.72	0.749	0.681	21	0.860	0.406	20.8	0.817	0.492	25.50	0.739	0.601	22.20	0.472	0.603	17.15
PatchTST	0.908	0.803	11.5	0.983	0.612	11.03	0.686	0.576	15	0.916	0.347	14.87	0.690	0.348	6.75	0.709	0.461	11.53	0.462	0.437	5.45
iTransformer	0.989	0.848	15.17	1.110	0.695	13.44	0.774	0.628	15.80	0.851	0.342	13.73	0.699	0.351	9.25	0.707	0.460	11.53	0.488	0.454	6.9
TFT	1.034	0.841	14.67	1.009	0.630	13.28	0.660	0.512	13.2	0.871	0.348	14.87	0.716	0.352	13.25	0.679	0.443	9.47	0.662	0.503	8.5
TIDE	1.507	1.084	25.5	1.167	0.751	17.91	0.803	0.912	23.00	1.372	0.561	23.07	0.981	0.484	24.25	0.790	0.531	17.80	0.623	0.568	16.65
DeepAR	1.541	1.221	22	1.782	1.072	23.56	0.765	0.723	17.8	1.644	0.535	21.47	0.707	0.352	11	0.745	0.484	12.4	0.850	0.633	17.9
Crossformer	29.323	109.221	32.00	2.193	1.224	22.72	8.529	5.177	24.40	2.979	0.690	21	1.593	5.768	31.75	1.769	0.811	15.93	0.92	0.615	18.05
N-BEATS	0.861	0.967	20.67	1.184	0.935	24.59	0.691	0.713	12.4	0.933	0.532	25.13	0.704	0.414	18.25	0.731	0.593	21.67	0.543	0.570	16.4
DLinear	1.133	1.124	26.00	1.151	0.879	23.97	0.792	0.806	24.20	1.117	0.496	25.20	0.808	0.481	24.25	0.808	0.654	24.13	0.724	0.660	20.30
Auto Arima	0.866	0.821	13.50	1.011	0.833	21	0.784	0.570	13.00	1.018	0.658	24.47	0.813	0.458	23.75	0.974	0.763	25.67	0.957	0.904	23.15
Auto Theta	0.983	0.841	14.17	1.358	1.702	26.72	0.951	0.803	20.4	1.060	0.910	27.07	0.873	0.48	24.50	1.082	1.326	29.13	0.521	0.608	18.9
Auto ETS	0.899	0.940	16	1.479	10.102	25.41	0.797	0.586	11.8	1.121	7.021	27.13	0.887	2.195	28.25	1.205	42.407	28.73	0.718	2.637	26.3
Seasonal Naive	1.000	1.000	22.67	1.000	1.000	25.44	1.000	1.000	24.60	1.000	1.000	29.67	1.000	1.000	30.25	1.000	1.000	27.80	1.000	1.000	25.10
Naive	1.433	1.170	23.17	1.555	1.533	28.44	1.157	1.194	26.80	0.962	1.326	30.13	1.002	0.896	30	1.260	2.069	31.13	1.133	1.066	25.9

Model	Long			Medium			Short		
	MASE	CRPS	Rank	MASE	CRPS	Rank	MASE	CRPS	Rank
YINGLONG$_{300\mathrm{m}}$	0.529	0.368	5.62	0.837	0.450	6.14	0.758	0.511	8.53
YINGLONG$_{100\mathrm{m}}$	0.540	0.379	6.71	0.853	0.465	7.71	0.765	0.514	8.67
YINGLONG$_{50\mathrm{m}}$	0.550	0.384	7.1	0.872	0.472	7.71	0.776	0.524	9.76
YINGLONG$_{6\mathrm{m}}$	0.599	0.418	11.43	0.948	0.515	12.19	0.819	0.557	13.13
TabPFN-TS	0.587	0.405	10.48	0.936	0.498	10.38	0.754	0.506	7.69
Chronos-bolt$_b$	0.568	0.424	12.62	0.894	0.515	11.10	0.735	0.500	6.13
Chronos-bolt$_s$	0.587	0.431	13.38	0.916	0.531	13.19	0.741	0.494	6.36
TimesFM-V2$_{500m}$	0.525	0.412	11.38	0.807	0.486	10.81	0.704	0.479	7.22
Moirai$_{\mathrm{large}}$	0.601	0.418	11.10	0.957	0.510	10.90	0.806	0.543	10.49
Moirai$_{\mathrm{base}}$	0.628	0.426	11.57	1.013	0.532	11.86	0.817	0.547	10.16
Moirai$_{\mathrm{small}}$	0.631	0.437	12.38	1.018	0.535	12.67	0.888	0.606	15.44
Chronos$_{\mathrm{large}}$	0.632	0.502	18.71	1.030	0.622	18.14	0.761	0.538	11.93
Chronos$_{\mathrm{base}}$	0.634	0.504	18.33	1.037	0.630	19.00	0.768	0.542	11.93
Chronos$_{\mathrm{small}}$	0.658	0.522	19.76	1.044	0.625	18.76	0.779	0.552	13.24
TimesFM	0.990	0.518	18.90	1.441	0.630	18.00	0.823	0.577	12.53
Lag-Llama	0.724	0.536	20.10	1.175	0.672	21.19	1.262	0.876	23.64
TTMs	0.731	0.596	22.48	1.202	0.756	23.76	0.993	0.822	24.75
Timer	0.753	0.650	24.71	1.223	0.830	25.67	1.067	0.891	26.04
VisionTS	0.522	0.456	16.71	0.847	0.583	18.38	0.871	0.751	23.67
PatchTST	0.537	0.368	7.52	0.856	0.461	7.33	0.832	0.571	13.04
iTransformer	0.566	0.391	9.29	0.867	0.470	8.62	0.889	0.611	14.15
TFT	0.589	0.379	8.38	0.949	0.468	8.24	0.883	0.592	14.87
TIDE	0.655	0.448	15.29	0.986	0.563	16.05	1.140	0.795	21.93
DeepAR	1.105	0.628	20.48	1.334	0.640	17.14	1.200	0.795	19.91
Crossformer	0.921	0.255	13.62	1.849	0.461	13.95	3.574	4.008	27.35
N-BEATS	0.644	0.565	20.29	1.032	0.678	21.24	0.862	0.748	22.80
DLinear	0.700	0.566	22.05	1.094	0.684	22.95	1.015	0.794	24.40
Auto Arima	0.985	0.805	24.57	1.022	0.833	23.90	0.935	0.735	20.13
Auto Theta	0.869	1.397	27.05	1.175	1.531	27.10	0.955	0.816	22.25
Auto ETS	0.956	369.207	29.24	1.610	6.230	28.71	0.984	1.347	22.33
Seasional Naive	1.000	1.000	26.48	1.000	1.000	25.62	1.000	1.000	26.60
Naive	1.403	1.892	30.29	1.462	1.873	29.57	1.143	1.093	26.96

Table 9: Results on GIFT-Eval aggregated by prediction length. The table shows MASE,CRPS and Rank for each method.

Table 10: Results on GIFT-Eval aggregated by frequency. The table shows MASE, CRPS and Rank for each method.

Daily	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	0.712	0.724	0.729	0.758	0.706	0.685	0.689	0.700	0.716	0.714	0.737	0.727	0.738	0.752	0.943	0.982
CRPS	0.367	0.372	0.370	0.386	0.347	0.346	0.350	0.357	0.377	0.378	0.397	0.375	0.379	0.384	0.586	0.637
Rank	7.67	8.60	9.53	12.73	7.53	5.00	6.13	7.33	12.33	12.73	15.00	11.60	10.53	12.20	24.93	26.00
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	0.746	0.822	1.188	4.832	0.887	0.906	0.775	0.749	0.725	1.146	0.831	0.882	0.901	0.936	1.000	1.000
CRPS	0.413	0.504	0.644	3.596	0.543	0.491	0.524	0.392	0.370	0.610	0.438	0.469	0.907	0.543	0.794	1.000
Rank	10.73	23.33	24.80	28.73	24.20	19.07	23.33	13.20	12.60	22.00	15.33	17.93	22.60	22.93	27.80	29.53
Hourly	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	0.725	0.738	0.737	0.796	0.770	0.662	0.690	0.734	0.763	0.763	0.773	0.763	0.775	0.867	0.853	0.941
CRPS	0.390	0.396	0.397	0.426	0.414	0.374	0.382	0.409	0.464	0.462	0.468	0.409	0.413	0.454	0.569	0.633
Rank	7.03	8.16	7.77	12.03	10.10	6.65	7.55	9.71	14.87	14.55	15.45	8.39	8.61	14.58	22.74	24.55
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	0.824	0.770	0.895	1.733	0.943	1.309	0.872	0.774	0.825	0.959	0.805	1.022	1.269	1.276	1.460	1.000
CRPS	0.469	0.525	0.489	0.534	0.606	0.623	0.600	0.407	0.428	0.511	0.424	0.743	50.063	1.573	1.667	1.000
Rank	15.29	20.68	18.55	17.94	24.06	18.39	23.23	10.00	12.52	18.55	11.81	25.84	29.68	29.77	30.58	28.39
Minutely	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	0.786	0.801	0.821	0.872	0.825	0.796	0.816	0.727	0.914	0.914	0.950	0.868	0.867	0.924	1.081	1.155
CRPS	0.497	0.512	0.527	0.572	0.534	0.553	0.574	0.534	0.659	0.659	0.686	0.560	0.568	0.597	0.784	0.955
Rank	4.70	5.43	6.67	10.63	9.13	9.50	11.53	10.57	17.00	16.87	18.30	10.93	11.50	12.43	22.33	26.67
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	1.458	0.871	1.199	2.044	1.139	1.564	0.996	0.892	0.907	1.042	0.872	0.991	1.306	1.106	1.211	1.000
CRPS	0.657	0.693	0.840	0.868	0.793	0.862	0.792	0.551	0.554	0.677	0.552	0.981	7.131	1.388	1.701	1.000
Rank	16.93	19.30	22.67	19.50	22.73	21.40	22.07	9.07	9.53	17.53	9.67	24.70	26.50	27.43	28.77	26.00
Monthly	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	0.867	0.846	0.868	0.895	0.744	0.842	0.810	0.700	0.812	0.857	0.827	0.825	0.815	1.016	1.182	1.210
CRPS	0.806	0.780	0.809	0.846	0.680	0.787	0.758	0.660	0.807	0.849	0.818	0.781	0.753	0.979	1.409	1.450
Rank	14.60	12.80	14.20	18.00	2.80	9.40	9.40	6.60	14.20	15.80	14.40	10.80	8.00	21.60	27.40	26.60
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	0.800	0.915	1.430	2.950	0.997	1.215	0.851	0.859	0.901	1.099	0.907	0.759	0.821	0.932	1.204	1.000
CRPS	0.733	1.030	1.352	5.640	1.141	1.030	0.962	0.832	0.840	1.157	0.803	0.759	0.770	0.873	1.524	1.000
Rank	8.40	25.20	22.40	19.60	26.20	19.20	20.40	14.80	15.20	25.40	12.00	12.60	10.80	16.80	28.80	23.60
Quarterly	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	0.869	0.921	0.876	0.918	0.732	0.765	0.780	0.603	0.769	0.769	0.775	0.713	0.713	0.775	1.240	1.838
CRPS	0.865	0.889	0.882	0.914	0.757	0.777	0.787	0.623	0.840	0.840	0.846	0.740	0.740	0.793	1.287	1.787
Rank	19.00	21.00	20.00	22.00	4.00	5.00	6.00	1.00	15.00	14.00	17.00	3.00	2.00	7.00	29.00	30.00
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	0.875	0.850	3.730	15.875	0.913	0.900	0.756	0.825	0.813	1.050	0.769	0.800	0.725	0.744	0.925	1.000
CRPS	0.853	1.048	2.517	119.960	1.109	0.841	0.972	0.835	0.837	1.018	0.797	0.823	0.798	0.797	0.951	1.000
Rank	18.00	27.00	31.00	32.00	28.00	16.00	24.00	12.00	13.00	26.00	9.00	11.00	10.00	8.00	23.00	25.00
Secondly	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	0.264	0.241	0.276	0.321	0.467	0.682	0.632	0.227	0.506	0.523	0.523	0.524	0.786	0.510	0.728	0.517
CRPS	0.497	0.482	0.513	0.573	0.785	1.138	0.874	0.429	0.818	0.859	0.793	0.873	1.025	0.977	1.529	0.957
Rank	7.83	6.50	8.83	12.17	16.50	27.00	20.33	3.83	18.33	19.50	17.17	21.17	24.33	22.17	28.83	23.17
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	0.787	0.216	0.598	0.615	0.368	0.376	0.271	0.224	0.537	0.323	0.235	1.000	0.551	0.159	1.983	1.000
CRPS	1.298	0.691	1.116	0.729	0.782	0.754	0.598	0.536	0.672	0.705	0.510	1.000	1.934	0.315	1.441	1.000
Rank	29.17	14.83	24.33	16.67	17.67	16.83	9.67	7.00	10.67	15.00	4.67	13.83	29.17	1.00	25.00	14.83
Weekly	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	0.851	0.902	0.899	0.963	0.739	0.747	0.763	0.863	0.737	0.762	0.745	0.931	0.911	0.967	1.133	1.136
CRPS	0.595	0.621	0.642	0.692	0.520	0.523	0.532	0.589	0.529	0.542	0.536	0.634	0.640	0.688	1.058	0.979
Rank	10.00	10.75 12.25	15.25	4.50	5.25	5.88	9.63	7.50	8.13	8.88	11.38	11.25	14.38	26.88	25.75	
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	0.847	1.038	1.629	4.364	1.137	1.460	1.082	0.929	0.921	1.294	1.248	0.946	0.932	1.028	1.000	1.000
CRPS	0.602	0.943	1.203	11.775	0.948	0.994	0.971	0.666	0.726	0.956	0.956	0.731	0.774	0.787	0.874	1.000
Rank	9.38	25.13	27.63	31.13	25.50	20.88	24.25	13.88	18.25	20.75	19.88	17.75	18.25	20.50	22.00	25.2
Yearly	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	1.086	1.188	1.097	1.328	0.796	0.883	0.929	0.639	0.917	0.917	0.942	0.748	0.758	0.751	1.288	2.899
CRPS	1.099	1.151	1.090	1.286	0.822	0.880	0.926	0.657	0.978	0.978	1.007	0.754	0.761	0.761	1.390	2.923
Rank	23.00	25.00	22.00	28.00	8.00	13.00	14.00	1.00	18.00	17.00	21.00	2.00	3.00	4.00	29.00	31.00
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	0.844	0.965	2.409	7.909	1.048	0.856	0.793	0.829	0.778	1.264	0.849	0.935	0.776	0.783	1.000	1.000
CRPS	0.848	1.152	2.039	102.899	1.217	0.819	0.971	0.848	0.797	1.123	0.848	0.942	0.804	0.833	0.993	1.000
Rank	11.00	26.00	30.00	32.00	27.00	7.00	16.00	10.00	5.00	24.00	12.00	15.00	6.00	9.00	19.00	20.00

Multivariate	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	0.658	0.65	0.668	0.709	0.756	0.725	0.736	0.629	0.788	0.794	0.804	0.802	0.826	0.800	0.930	0.895
CRPS	0.421	0.422	0.43	0.459	0.482	0.490	0.487	0.448	0.552	0.555	0.555	0.515	0.516	0.519	0.694	0.716
Rank	5.95	5.63	6.19	9.95	10.60	10.70	11.21	10.72	16.67	16.47	16.40	11.98	12.09	12.72	23.30	24.81
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	1.173	0.695	0.957	1.160	0.963	1.495	0.782	0.711	0.840	1.007	0.737	1.033	1.055	0.801	1.147	1.000
CRPS	0.582	0.585	0.647	0.669	0.663	0.802	0.641	0.451	0.495	0.659	0.478	0.837	4.053	0.926	1.259	1.000
Rank	17.67	19.35	21.60	21.28	22.37	22.30	20.77	9.33	11.95	18.98	9.98	23.02	26.74	23.63	27.63	26.00
Unvariate	YINGLong$_{\mathrm{300m}}$	YINGLong$_{\mathrm{100m}}$	YINGLong$_{\mathrm{50m}}$	YINGLong$_{\mathrm{6m}}$	TabPFN-TS	Chronos-bolt$_b$	Chronos-bolt$_s$	TimesFM-V2$_{500m}$	Chronos$_{\mathrm{large}}$	Chronos$_{\mathrm{base}}$	Chronos$_{\mathrm{small}}$	Moirai$_{\mathrm{large}}$	Moirai$_{\mathrm{base}}$	Moirai$_{\mathrm{small}}$	TTMs	Timer
MASE	0.766	0.793	0.799	0.861	0.742	0.725	0.739	0.724	0.775	0.780	0.797	0.773	0.795	0.890	1.001	1.131
CRPS	0.500	0.514	0.522	0.564	0.479	0.482	0.488	0.479	0.543	0.547	0.564	0.499	0.514	0.575	0.803	0.913
Rank	8.52	9.96	10.78	14.63	7.50	6.94	7.89	7.44	13.20	13.56	15.41	9.70	9.83	15.33	24.63	26.35
	TimesFM	VisionTS	Lag-Llama	Crossformer	DLinear	DeepAR	N-BEATS	PatchTST	TFT	TIDE	iTransforme	Auto Arima	Auto ETS	Auto Theta	Naive	Seasonal Naive
MASE	0.829	0.845	1.233	4.000	0.944	1.016	0.892	0.805	0.808	0.959	0.857	0.912	1.115	1.147	1.358	1.000
CRPS	0.569	0.683	0.831	2.464	0.758	0.662	0.730	0.535	0.524	0.646	0.564	0.721	9.018	1.162	1.491	1.000
Rank	13.04	22.35	22.93	21.63	24.54	17.15	22.83	11.63	12.09	19.41	13.43	21.02	23.98	24.91	28.74	26.65

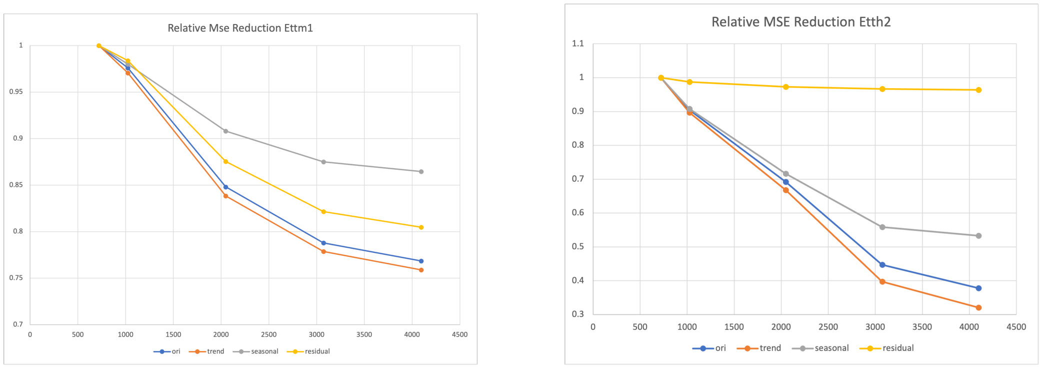
Table 11: Results on GIFT-Eval aggregated by number of variates. The table shows MASE, CRPS and Rank for each method.

E Error Reduction Pattern Analysis

We performed STL decomposition on the target and predicted sequences generated by our model with various DCoT lengths to determine the primary source of error reduction. It is hypothesized that error reduction primarily arises from DCoT tokens, which likely encapsulate information related to general signal patterns or low-frequency trends. Consequently, the majority of error reduction may be attributed to improvements in trend prediction. Our initial experiments corroborate this hypothesis. Furthermore, we analyzed relative improvements with extending DCoT lengths, as absolute MSE reductions may yield dataset-dependent results, given that some datasets are more amenable to trend-based enhancements rather than seasonal improvements. As illustrated in Figure 5, relative improvements highlight that error reductions in the trend are more pronounced compared to seasonal reductions, while the residual component remains stable due to its generally unlearnable nature.

DCOT(Output)	MSE_all	MSE_trend	MSE_seasonal	MSE_residual
720	0.550 93	0.374 31	0.017 02	0.062 45
1024	0.537 69	0.363 28	0.016 68	0.061 43
2048	0.467 23	0.313 81	0.015 45	0.054 66
3072	0.434 10	0.291 48	0.014 89	0.051 30
4096	0.423 35	0.284 06	0.014 71	0.050 25

Table 12: Performance metrics across different DCOT outputs for analysis in ETTm1.



(a) Relative MSE Reduction ETTm1

(b) Relative MSE Reduction ETTh2

Figure 5: Comparison of relative MSE reduction in ETTm1 models.

F Influence of Input Ensemble

We compare the average performance of forecasts generated using multiple input lengths against the single-best component, as detailed in Appendix Table 13. Our input-ensemble strategy consistently enhances accuracy by 1% to 4% without necessitating additional model training. This presents a scalable 'free lunch' approach, yielding post-training improvements through simple ensemble averaging. When compared to the single-worst component in the setup, the accuracy improvement is even more pronounced. Since comparative accuracy assessments cannot be made prior to determining the target value, this ensemble technique proves to be both practical and robust for real-world applications.

G Structure ablations

We performed a structural ablation study on our U-transformer and token merge designs. As illustrated in Figure 6, both designs contribute to 1% to 5% improvements in MSE and MAE, depending on

Table 13: Influence of Multi-Input Ensemble

Model	ETTh1								Weather							
	1024		2048		4096		Ensemble		1024		2048		4096		Ensemble	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
YINGLONG$_{6m}$	0.411	0.413	0.414	0.423	0.437	0.45	0.408	0.412	0.248	0.27	0.247	0.272	0.243	0.274	0.239	0.268
YINGLONG$_{50m}$	0.409	0.412	0.396	0.409	0.388	0.41	0.405	0.408	0.246	0.266	0.235	0.261	0.226	0.257	0.231	0.257
YINGLONG$_{100m}$	0.406	0.410	0.412	0.414	0.389	0.407	0.399	0.408	0.242	0.262	0.233	0.256	0.23	0.256	0.227	0.255
YINGLONG$_{300m}$	0.425	0.417	0.412	0.414	0.393	0.412	0.398	0.407	0.233	0.255	0.22	0.245	0.215	0.245	0.217	0.245

the dataset tested. Generally, the U-transformer architecture employed in $YINGLONG$ yields the best results, which is why it serves as the backbone architecture. However, it is important to note that the architectural design alone is not the primary driver of the state-of-the-art performance seen in $YINGLONG$. The key contributors are the joint forecasting paradigm and the DCoT design.

Dataset	**Transformer**	**Token-Merge+**	**uTransformer+**
etth1	0.407	0.408	0.408
etth2	0.361	0.348	0.344
ettm1	0.381	0.375	0.370
ettm2	0.255	0.252	0.250
weather	0.242	0.241	0.239

(a) MSE Comparison

Dataset	**Transformer**	**Token-Merge+**	**uTransformer+**
etth1	0.423	0.420	0.412
etth2	0.402	0.390	0.382
ettm1	0.374	0.367	0.368
ettm2	0.317	0.305	0.302
weather	0.277	0.273	0.268

(b) MAE Comparison

Figure 6: Structure ablation:6M size Transformer, add Token-Merge, and uTransformer($Yinglong$) model variates across different datasets for MSE and MAE metrics.

H Output Scaling for vanilla transformer model

We investigated the output scaling effect across a range of vanilla transformer models with sizes from 6 million to 300 million parameters within our joint forecasting paradigm. For all models of varying sizes, the output scaling effect was significant, and the scaling effect with model size persisted. Specifically, for the largest 300M parameter model, our joint forecasting approach resulted in substantial improvements in Mean Absolute Scaled Error (MASE) and Continuous Ranked Probability Score (CRPS) by 24.9% and 30.0%, respectively, when comparing the DCoT setting with a token length of 4096 to a non-DCoT configuration. This effect clearly demonstrates robust output scaling within the extensive GIFT-Eval benchmark, which encompasses 23 datasets. These findings indicate that the observed scaling effect is neither specific to a particular architectural model design nor limited to a single dataset.

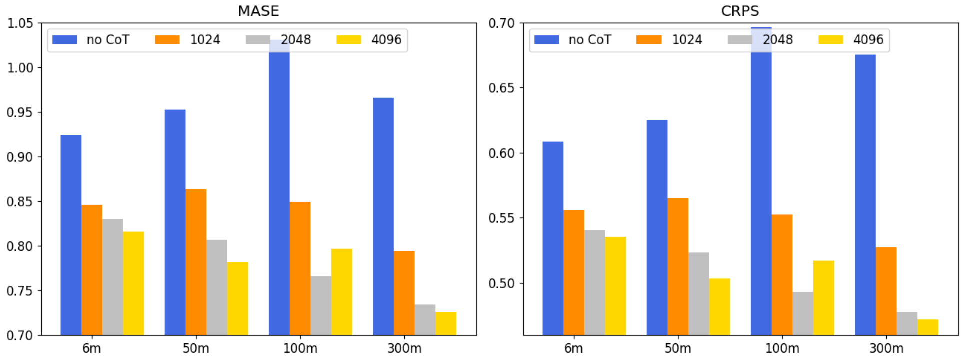


Figure 7: Output Scaling for 6M to 300M vanilla transformer model following joint forecasting paradigm under different DCOT setting