DIVERSIFIED SCALING INFERENCE IN TIME SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The advancement of Time Series Foundation Models (TSFMs) has been driven primarily by large-scale pre-training, but inference-time compute potential remains largely untapped. This work systematically investigates two questions: how do TSFMs behave under standard sampling-based inference scaling, and can controlled sampling diversity enhance performance? We first examine the properties of TSFMs under standard sampling, revealing that they often fail to adhere to scaling laws due to insufficient exploration of the solution space. Building on this, we then delve into diversified inference scaling via tailored time series perturbations to expand the generative distribution's support. We theoretically analyze the diversity-fidelity trade-off and derive a critical sample threshold for diversified sampling to outperform standard sampling. Extensive experiments across various TSFMs and datasets show that proper diversified inference scaling yields substantial performance gains without model parameter updates, establishing inference scaling as a critical, highly parallelizable paradigm for TSFM optimization. As an application, we propose RobustMSE, a rigorous metric to quantify TSFM headroom under a fixed budget. Overall, our findings clarify these factor interactions, enabling reliable performance via diverse large-scale time series inference in parallel environments without re-training TSFMs.

1 INTRODUCTION

The emergence of Time Series Foundation Models (TSFMs) has fundamentally altered the landscape of forecasting (Das et al., 2024; Li et al., 2025). By leveraging large-scale pre-training on heterogeneous corpora, models such as Chronos, TimesFM, and Moirai have demonstrated remarkable zero-shot generalization and the ability to learn transferable temporal representations (Goswami et al., 2024; Ansari et al., 2024; Das et al., 2024; Woo et al., 2024). To date, the primary driver of this progress has been the scaling of training compute, increasing model parameters and dataset size to refine representation quality and generalization (Kaplan et al., 2020; Hoffmann et al., 2022). However, the potential of inference compute, i.e., the computational resources expended at test time to refine predictions, remains largely untapped in the time series domain. This leaves a critical open question: *To what extent can the performance of a fixed TSFM be improved solely by scaling inference strategies, without additional training or parameter updates?*

In adjacent generative domains, particularly Natural Language Processing (NLP) and reasoning, extending computation at inference time has proven to be a potent lever for performance gains (Chen et al., 2024; Wang et al., 2025; Kadavath et al., 2022). Strategies such as chain-of-thought, best-of-N sampling, and verifiers allow models to explore a broader solution space, effectively trading test-time compute for accuracy (Wei et al., 2022; Wang et al., 2023; Sun et al., 2024), especially when combined with multi-sample aggregation (Wang et al., 2023). Critically, recent findings suggest that diversity is a prerequisite for effective scaling. Simply sampling more from a biased distribution yields diminishing returns, whereas introducing controlled stochasticity or input diversity expands the effective support of the generation distribution, enabling the model to uncover higher-quality predictions (Wang et al., 2025; Vijayakumar et al., 2017). These findings suggest that inference exploration when properly structured can serve as a powerful complement to large foundation models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Motivated by these developments, we present a systematic study of diversified scaling inference in TSFMs, as shown in Figure 1. We posit that the stochastic nature of time series forecasting makes it well-suited for inference-time search, whereas standard greedy or low-temperature sampling methods may often converge to sub-optimal modes due to limited exploration. To address this, we investigate two complementary axes of scaling: (i) **Inference Scaling**, where we analyze the effects of expanding the candidate pool size under varied model sizes, context lengths, and temperatures, and (ii) **Diversified Sampling**, where we introduce controlled diversity at the input level via task-agnostic and task-specific perturbations. Unlike standard sampling, which is constrained by the fixed conditioning of the time-series context, diversified sampling approximates a marginalization over the input neighborhood. As depicted by the arrows pointing towards the ground truth in the bottom distribution plot, this theoretically expands the search space to cover regions closer to the target. We unify these strategies under a rigorous evaluation framework, utilizing aggregation mechanisms such as majority voting and exact match to synthesize candidate predictions. Furthermore, to fairly quantify the attainable utility of a model given a compute budget, we propose RobustMSE, a metric designed to capture the performance envelope of TSFMs under optimal sampling strategies.

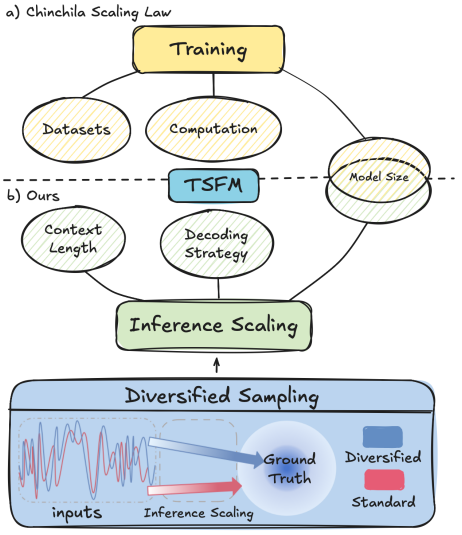


Figure 1: Relationship between scaling training (Hoffmann et al., 2022; Yao et al., 2025) and diversified scaling inference in time series foundation models.

Our empirical analysis across multiple architectures and datasets reveals recurring empirical patterns for TSFM inference. We observe that while scaling the number of samples alone improves stability, there is still an upper limit for TSFMs that does not follow the laws similar to NLP (Kaplan et al., 2020; Hoffmann et al., 2022). For example, smaller models occasionally surpass larger models, and the performance is neither monotonic in context length nor log-linear in the number of samples. In addition, we demonstrated that when appropriate diversified sampling takes effect, demonstrating that significant gains in accuracy can be unlocked with different aggregations. These findings establish inference design not merely as a post-processing step, but as a critical dimension of TSFM scaling. By clarifying how these factors interact, our study helps users easily achieve reliable performance by designing diverse large-scale time series in parallel environments, rather than pre-training a TSFM.

Our contributions are summarized as follows:

- We provide the first comprehensive study of inference-time computation in TSFMs, characterizing the interaction between model capacity, context length, and temperature, establishing that TSFMs exhibit significant performance gains when scaling test-time compute.
- We theoretically motivate and empirically validate the necessity of exploration diversity. We show that diversifying the input via controlled perturbations strictly expands the support of the sampling distribution.
- We systematically perform an empirical evaluation of diversified sampling in various TSFMs and datasets, showing that appropriate perturbations can consistently improve performance under scaling inference, along with an actionable insight to guide practical deployment.
- As an application, we introduce RobustMSE, a new metric allowing for a fair comparison of the headroom of TSFMs under a fixed inference budget.

2 PRELIMINARIES

Time Series Foundation Model A forecasting TSFM is regarded as a generative model $f : \mathbb{R}^{L\times D} \to \mathbb{R}^{H\times D_{\text{out}}}$ with parameters θ pre-trained on a massive, heterogeneous collection of time series, where D and D_{out} denote the input and output dimensions respectively (typically $1 \le D_{\text{out}} \le D$). Given a historical sequence $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_L\} \in \mathbb{R}^{L\times D}$ with a context length L, the model

produces a predicted sequence $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_{L+1}, \ldots, \hat{\mathbf{y}}_{L+H}\} \in \mathbb{R}^{H \times D_{\text{out}}}$ for a forecast horizon H. The model is optimized by minimizing a global loss $\mathcal{L}$ over the time series as

$$\theta^* = \arg\min_{\theta} \sum_{\mathbf{X}} \mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}),$$

where $\mathbf{Y} = \{\mathbf{y}_{L+1}, \ldots, \mathbf{y}_{L+H}\} \in \mathbb{R}^{H \times D_{\text{out}}}$ represents the ground truth of target variables. This formula allows the model to leverage multi-dimensional context (e.g., exogenous variables) to predict target dimensions, while maintaining generalization across diverse temporal patterns and scales.

Sampling Strategy Given an input $\mathbf{X}$ and a generative model $f(\cdot \mid \mathbf{X})$, a sampling strategy S defines a probabilistic procedure for generating a finite set of candidate outputs

$$\mathcal{Y} = \{\hat{\mathbf{Y}}_1, \hat{\mathbf{Y}}_2, \ldots, \hat{\mathbf{Y}}_N\}, \quad \hat{\mathbf{Y}}_i \sim f(\cdot \mid \mathbf{X}, S),$$

where N is the number of sampled candidates. The sampling strategy may correspond to a fixed decoding distribution such as temperature τ and top-p nucleus sampling, or a diversified strategy π_S from perturbed samples $\mathbf{X}'$, formally $\mathbf{X}' \sim \pi_S(\cdot \mid \mathbf{X})$. The candidate set $\mathcal{Y}$ is subsequently combined by an aggregation function A, e.g., exact match or majority voting (Wu et al., 2025), to produce the final prediction. Our TSFM inference applies parallel sampling, and we analyze the latency in Appendix I. These sampling strategies have prepared us for the scaling inference of TSFMs.

Scenarios and Settings To better investigate the impact of inference-time scaling on time series forecasting, we select 4 datasets under long-term settings, including *ETTh1*, *ETTm1*, *Electricity*, and *Traffic* from TSLib[1], covering diverse application domains, data frequencies, and noise levels. To verify the universal applicability of the diversified sampling approach we propose, we choose 4 representative TSFMs based on these datasets: *Chronos* (Ansari et al., 2024) for the encoder-decoder, *TimesFM* (Das et al., 2024) and *Time-MoE* (Shi et al., 2025) for the decoder-only, and *Moirai* (Woo et al., 2024) for the encoder-only. All evaluations are assessed based on the Mean Squared Error (MSE). We then systematically evaluate the general performance gains of scaling-based inference strategies across these TSFMs. Detailed experimental settings are provided in Appendix D.

3 Scaling Inference in TSFM

While prior work (Yao et al., 2025) has investigated neural scaling laws in the training of TSFMs, the behavior during the inference phase remains largely underexplored. Evidence from LLMs (Wang et al., 2025) suggests that allocating additional compute at inference time can yield greater efficiency gains than simply increasing model parameters. Motivated by this paradigm shift, we explore a fundamental question in this section: *Do TSFMs exhibit inference scaling behaviors analogous to those observed in LLMs?* To answer this question, we proceed from the aggregators and examine three dimensions of TSFMs: model size, context length, and decoding strategy.

First, we consider two widely adopted aggregators for candidates, Exact Match (EM) and Majority Voting (MV). Notably, EM employs a selection strategy by exploiting the optimal individual candidate, typically used for validation; whereas MV leverages distributed aggregation to enhance robustness, typically deployed for inference. Note that EM is not necessarily always less than MV.

- **Exact Match** also termed best-of-N in reasoning (Wang et al., 2023) and pass@N in code (Chen et al., 2024) tasks, we consider it as EM@N in our time series forecasting task. This aggregator prepares a set of time series candidates $\mathcal{Y}$ and picks the one with the lowest MSE loss compared with the ground truth:

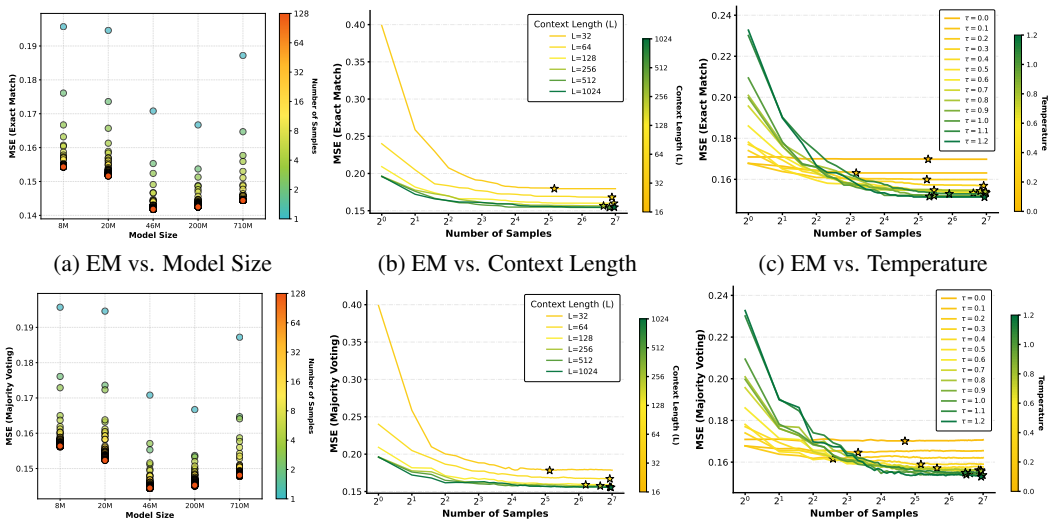
$$\hat{\mathbf{Y}}_{\texttt{EM}} = \arg\min_{\hat{\mathbf{Y}}_i \in \mathcal{Y}} \frac{1}{H} ||\hat{\mathbf{Y}}_i - \mathbf{Y}||_2^2, \ \texttt{EM} = \mathcal{L}(\hat{\mathbf{Y}}_{\texttt{EM}}, \mathbf{Y}). \tag{1}$$

- **Majority Voting** is referred to as self-consistency in reasoning tasks (Brown et al., 2024), generating multiple candidates and determining the final answer as the most frequently occurring one among all outputs. In our scenarios, MV computes the median across candidates:

$$\hat{\mathbf{Y}}_{\texttt{MV}} = \text{quantile}_{0.5}(\mathcal{Y}), \ \texttt{MV} = \mathcal{L}(\hat{\mathbf{Y}}_{\texttt{MV}}, \mathbf{Y}). \tag{2}$$

Now, we empirically evaluate inference scaling in model size, context length, and decoding strategy.

[1]Time-Series-Library



(a) EM vs. Model Size (b) EM vs. Context Length (c) EM vs. Temperature

(d) MV vs. Model Size (e) MV vs. Context Length (f) MV vs. Temperature

Figure 2: **Performance of *Chronos* on the data *ETTh1* under different scaling factors and aggregation functions**. We systematically evaluate the impact of three critical factors: model size, context length, temperature. The number of sampling candidates is set to 128, plotted on a logarithmic axis. The stars indicate the initial convergence number to the minimum MSE for each configuration.

Effect of Model Size Inspired by LLM scaling laws (Kaplan et al., 2020), we systematically investigate how TSFM size affects scaling inference. We evaluate different sizes of the same model under both aggregation functions, examining the joint effect of model size and sample count. In Figure 2 *(a)* and *(d)*, and consistently across additional TSFMs in Appendix E, MSE loss generally decreases as model size increases. However, this improvement is not strictly monotonic on log scales. For example, *Chronos* on *ETTh1* achieves its best performance at an intermediate size before slightly degrading with further scaling. We attribute this to the highly domain-dependent nature of time series tasks, where larger model capacities do not universally guarantee better performance (Zeng et al., 2023; Kim et al., 2025). Overall, across all datasets and TSFMs, increasing model sizes yields substantial performance gains up to approximately 80% optimally, yet the absence of consistent results indicates that performance in TSFMs does not strictly follow scaling inference. This observation suggests simply enlarging model capacity is insufficient to guarantee stable improvements, highlighting the need for complementary mechanisms beyond parameter scaling.

Effect of Context Length We next analyze the impact of context length (L) on inference scaling, given its importance in temporal modeling (Wu et al., 2023). In all experiments (Figure 2 *(b)* and *(e)* and Appendix E), an obvious pattern emerges when jointly considering context length and the number of samples. With a small sampling budget, increasing the context length generally reduces MSE, whereas as the number of samples increases, prediction errors decrease rapidly and often converge across different lengths, yielding performance gains of up to approximately 90%. However, such convergence is not universal. Across different models and datasets, increasing the number of samples often reduces MSE but does not necessarily eliminate performance differences across context lengths, suggesting that the effectiveness of longer context depends on both model architecture and data characteristics. Moreover, under limited sampling, MV is more likely than EM to converge earlier. Overall, these results show that extending context length alone does not produce a strictly monotonic or universal inference scaling effect.

Varying Sampling Temperatures As sampling temperature (τ) can directly control the generation randomness and higher temperatures can increase output diversity (Holtzman et al., 2020), we explore the performance of scaling inference by adjusting temperatures. As shown in Figure 2 *(c)* and *(f)*, and consistently observed across Appendix E, we evaluate temperature swept in 0.1 increments from 0.0 to 1.2. Across all datasets, temperature exhibits heterogeneous effects. Under small budgets, higher temperatures often increase MSE due to excessive stochasticity, while low temperatures provide limited diversity and marginal gains from additional samples. As N increases, higher temperatures

increasingly benefit from aggregation, with MSE rapidly decreasing and often surpassing low-temperature performance. Overall, temperature does not independently induce inference scaling; rather, its effect is tightly coupled with sampling budget and decoding capability.

Remark. *The results do not strictly adhere to logarithmic scaling laws.* The bottleneck of standard scaling lies in its distributional limitations. While scaling inference improves performance, standard sampling saturates rapidly due to distributional bias in the fixed predictive distribution $P_\theta(\hat{\mathbf{Y}}|\mathbf{X})$. Simply drawing more samples traps the model in the dominant but sub-optimal peak, preventing it from discovering the ground truth hidden in low-probability regions. To break this limitation, we introduce input diversity, which constructs a marginal mixture distribution to expand the support set and calibrate the inference toward the true solution.

4 Encouraging Diversified Sampling

In this section, we first derive a theoretical insight into how sample diversification can improve TSFM performance. Guided by this insight, we instantiate a set of perturbations that are likely to be effective for diversified scaling inference, including task-agnostic and task-specific perturbations.

4.1 Theoretical Analysis of Diversified Sampling

To understand why diversified sampling improves the TSFM performance in best-of-N, we provide a theoretical perspective. We examine the behavior of the minimum loss as the sample size N approaches infinity, as well as the expected performance under finite N. The proofs are presented in Appendix B. First, we formalize the settings.

Definition 4.1 (**Sampling Strategies**). Let $P(\hat{\mathbf{Y}}|\mathbf{X})$ be the probability density function of the TSFM output given $\mathbf{X}$. Let $\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}})$ be the forecasting loss. We first consider two sampling strategies to generate a candidate set $\mathcal{Y}$:

Standard Sampling: The candidate set is generated by sampling N times, i.i.d. with a fixed configuration ζ (e.g., model size, context length, decoding seed, temperature, etc.), from a fixed distribution:

$$\hat{\mathbf{Y}}_i \sim P(\cdot|\mathbf{X}, \zeta), \quad \forall i \in \{1, \dots, N\}. \tag{3}$$

We denote the minimum EM for the standard sampling as $\mathrm{EM}_N^{\mathrm{std}} = \min_{i=1}^N \mathcal{L}(\hat{\mathbf{Y}}_i, \mathbf{Y})$.

Diversified Sampling: Let $Q(\mathbf{X}'|\mathbf{X}, \pi_S)$ be a distribution of π_S over inputs, i.e., Q is a test-time augmentation. The candidate set is generated by the marginal mixture distribution:

$$\hat{\mathbf{Y}}_i \sim P_{\text{mix}} = \int P(\cdot|\mathbf{X}', \zeta) Q(\mathbf{X}'|\mathbf{X}, \pi_S) d\mathbf{X}'. \tag{4}$$

Equivalently, for each i, we first sample $\mathbf{X}' \sim Q(\cdot|\mathbf{X}, \pi_S)$ and then $\hat{\mathbf{Y}}_i \sim P(\cdot|\mathbf{X}', \zeta)$. We denote the EM for diversified sampling as $\mathrm{EM}_N^{\mathrm{div}} = \min_{i=1}^N \mathcal{L}(\hat{\mathbf{Y}}_i, \mathbf{Y})$.

Asymptotic Analysis ($N \to \infty$) Our first result concerns the potential of diversified sampling when the sample budget is unlimited. We demonstrate that introducing input perturbations strictly expands the search space without discarding the original solution capabilities.

Proposition 4.2 (**Asymptotic Lower Bound**). *Let $\mathcal{S}_{std}$ and $\mathcal{S}_{div}$ be the support sets of the standard decoding and diversified sampling distributions, respectively. Under the mild assumption that the perturbation distribution covers the original input, we have $\mathcal{S}_{std} \subseteq \mathcal{S}_{div}$. Consequently, as $N \to \infty$, the minimum loss of diversified sampling is bounded by:*

$$\lim_{N\to\infty} EM_N^{div} \leq \lim_{N\to\infty} EM_N^{std}. \tag{5}$$

The inequality becomes strict (e.g., $<$) if the perturbation allows the model to explore regions of the manifold closer to the ground truth $\mathbf{Y}$ that are inaccessible under ζ.

Remark. This proposition suggests that diversified sampling effectively mitigates the inductive bias of the fixed model configuration. By expanding the support set, the method can access blind spots of the TSFM. As N increases, the probability of hitting these superior regions approaches certainty, guaranteeing a lower bound for EM.

Table 1: Effect of diversified sampling across TSFMs and datasets with 128 trials. Cells highlighted in red represent failed perturbations that result in MSE greater than 20% compared to standard sampling.

Diversified Strategy	Chronos-T5-Tiny			Moirai-1.1-R-Small			TimesFM-2.5-200M			TimeMoE-50M			Fail Cout.
	ETTh1	Traffic	Electricity	ETTh1	Traffic	Electricity	ETTh1	Traffic	Electricity	ETTh1	Traffic	Electricity	
None	0.1969±0.0056	0.2970±0.0067	0.6318±0.0532	0.7412±0.7458	0.1856±0.1151	0.6457±0.1441	0.1411±0.0000	0.0440±0.0000	0.1887±0.0000	0.2412±0.0000	0.3205±0.0000	0.9099±0.0000	-
Prefix	0.1967±0.0057	0.2970±0.0094	0.6483±0.0833	0.6283±0.2240	0.1940±0.1380	0.6808±0.0705	0.1508±0.0047	0.0437±0.0001	0.1857±0.0012	0.2410±0.0001	0.3210±0.0003	0.9104±0.0005	0
Suffix	0.8747±0.3240	1.2307±0.7274	1.0046±0.2605	1.2239±0.5135	1.3229±0.8577	1.2391±0.4795	0.3370±0.1154	1.0543±0.8516	0.7706±0.3820	0.2829±0.0296	1.6084±0.8924	1.3404±0.2703	**11**
Insertion	0.1940±0.0059	0.3174±0.0373	0.6727±0.0953	0.6422±0.2749	0.2581±0.1142	0.7368±0.0591	0.1416±0.0009	0.0577±0.0127	0.2055±0.0156	0.2410±0.0002	0.3227±0.0035	0.9039±0.0060	2
Gaussian	0.1701±0.0068	0.2218±0.0290	0.6315±0.0721	3.3840±3.8760	0.2417±0.0768	0.7909±0.1558	0.1418±0.0013	0.0458±0.0019	0.1931±0.0041	0.2406±0.0018	0.3168±0.0022	0.9071±0.0017	3
Random	0.1937±0.0104	0.2614±0.0177	0.6901±0.0934	3.8157±3.4359	0.3185±0.0844	0.9280±0.4061	0.1398±0.0017	0.0474±0.0020	0.1966±0.0047	0.2389±0.0018	0.3190±0.0017	0.9040±0.0030	3
Missing	0.4079±0.1268	0.3758±0.1251	0.7877±0.1283	6.3197±24.2453	0.4693±0.1497	1.1945±0.4356	0.1402±0.0026	0.0641±0.0157	0.2897±0.0236	0.2328±0.0022	0.3193±0.0050	0.8525±0.0318	**8**
Sensitivity	0.1958±0.0051	0.3109±0.0097	0.6612±0.0736	0.6760±0.3648	0.1810±0.0168	0.6582±0.0233	0.1413±0.0002	0.0443±0.0002	0.1887±0.0001	0.2454±0.0016	0.3243±0.0016	0.9251±0.0064	0
Dependency	0.1944±0.0058	0.3048±0.0069	0.6827±0.1045	0.7344±0.5802	0.1863±0.0446	0.6505±0.0159	0.1420±0.0006	0.0451±0.0007	0.1893±0.0004	0.2463±0.0022	0.3261±0.0025	0.9280±0.0076	0
Reconstruction	0.1991±0.0062	0.3409±0.0014	0.6926±0.0719	0.6505±0.3703	0.2346±0.1427	0.7380±0.0439	0.1471±0.0016	0.0433±0.0016	0.1920±0.0022	0.2424±0.0010	0.3064±0.0069	0.9118±0.0050	1

Finite Sample Analysis ($N < \infty$) While asymptotic analysis guarantees superiority in the limit, practical TSFM applications operate with a finite budget N. We model this as a trade-off between the *quality of the best subspace* and the *probability of sampling from it*. We model the expected minimum EM to derive a critical sample size N^*.

Proposition 4.3 (Critical Sample Threshold). *Let $\mathcal{L}_0$ be the baseline loss. For a diversified strategy that introduces a variance in performance (yielding lower loss $\mathcal{L}_{good}$ with probability ρ and higher loss $\mathcal{L}_{bad}$ otherwise), there exists a critical threshold N^* given by:*

$$N^* = \frac{\ln\left(\frac{\mathcal{L}_{bad}-\mathcal{L}_{good}}{\mathcal{L}_0-\mathcal{L}_{good}}\right)}{\ln\left(\frac{1}{1-\rho}\right)}. \tag{6}$$

Diversified sampling outperforms standard sampling in expectation (i.e., $\mathbb{E}[EM_N^{div}] < \mathcal{L}_0$) if and only if the sample budget satisfies $N > N^*$.

Remark. This result highlights a "risk-reward" dynamic. When N is small ($N \leq N^*$), standard sampling is more conservative, as the potential benefit of diversified sampling may not outweigh the risk of harmful perturbations ($\mathcal{L}_{\text{bad}}$). As N exceeds the critical threshold N^*, the increased likelihood of sampling high-quality outputs ($\mathcal{L}_{\text{good}}$) can increasingly offset this risk, making diversified sampling more likely to provide a performance benefit. However, EM generally exhibits more gains as N increases, while MV may provide smaller gains or even degrade performance under unfavorable distributions (Wang et al., 2025). These theoretical findings align with our empirical observations.

4.2 Perturbation Design in Sampling

According to the theoretical insight, we delineate two distinct categories of the perturbation function π_S in diversified sampling: (1) Task-agnostic perturbations, including Prefix Padding, Suffix Padding, Middle Insertion, Gaussian Noise, Random Offset, and Missing Data; and (2) Task-specific perturbations, including Task Sensitivity, Task Dependency, and Task Reconstruction. Task-agnostic perturbations are independent of the target task or dataset, relying on commonly used methods to achieve sample diversification. By contrast, task-specific ones are tightly linked to samples in accordance with their inherent temporal structure or task-defined patterns. Detailed definitions and visualizations of all test-time augmentation variants can be found in Appendix C.

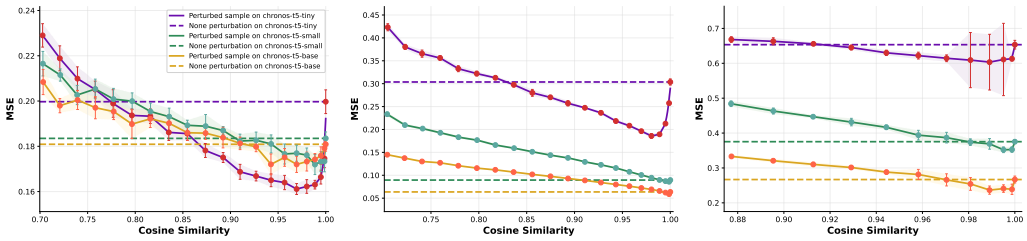
5 Efforts of Diversified Sampling in Scaling Inference

Building on the theory in Section 4, we investigate conditions and mechanisms under which strategies improve performance. We first identify which techniques yield significant gains and then quantify their effect on forecasting accuracy across TSFMs and datasets. Additionally, we conduct ablation studies on inference scaling under different perturbations, with details reported in the Appendix G.

5.1 When Does It Become Effective?

Diversified sampling fundamentally changes the effective support of the forecasting distribution, its benefit depends on whether the induced diversity expands fidelity or noise. Hence, we explicitly study when it is effective, and only effective perturbations are worth preserving.

Some Diversified Strategies Often Fail Diversified sampling does not always improve inference performance, its effectiveness depends on how randomness interacts with model structure and task characteristics (Holtzman et al., 2020; Ahmed et al., 2025; Wang et al., 2025). We therefore evaluate



(a) Performance on *ETTh1* (b) Performance on *Traffic* (c) Performance on *Electricity*

Figure 3: **Relationship between cosine similarity of time series samples and forecasting performance.** Results are obtained on three datasets using *Chronos-T5* models of different sizes. Solid lines represent performance under diversified sampling, with shaded regions indicating standard deviation across multiple runs. Horizontal lines correspond to the standard sampling for each model.

the perturbations in Section 4.2 across multiple TSFMs by measuring their performance changes relative to non-diversified baselines. We define failures as cases where performance drops by more than 20%. The results in Table 1 show that unlike language models in NLP tasks, *Suffix* and *Missing* perturbations incur the highest failure rates and often provide little benefit or even degrade performance, suggesting that they introduce noise rather than useful diversity (further verified across additional datasets in Appendix F.4). Accordingly, we discard these strategies and restrict subsequent analysis to the validated set $\pi_S \in \Pi_{\text{valid}}$.

The Fidelity-Diversity Trade-off We further examine when diversified sampling improves inference performance. Introducing diversification is beneficial only if the induced predictive distribution increases probability mass near $\mathbf{Y}$, while preserving sufficient fidelity to the original input. This reveals a fundamental trade-off: overly conservative perturbations fail to expand the predictive distribution, whereas overly aggressive ones increase diversity at the cost of degraded sample quality, leading to a worse N^* limit. Consequently, diversified sampling is effective only within a limited regime where perturbations provide sufficient exploration without compromising fidelity. Empirically, this trade-off is illustrated in Figure 3. Across all three datasets, diversified sampling performance exhibits a clear non-monotonic relationship with perturbation relevance: MSE decreases as relevance increases, reaches an optimal minimum, and then slightly rises. Notably, the best performance generally occurs within a narrow relevance band, between cosine similarity 0.95 and 1.00, confirming the fidelity–diversity trade-off. Beyond this global trend, diversified sampling also induces qualitative changes in relative model performance. On ETTh1, a representative pattern emerges in which smaller models exhibit higher sensitivity but greater upside under diversified sampling, illustrating that properly balanced diversification can reverse relative performance ordering across models.

> 📌 ***Takeaways:*** *Diversified sampling works only if perturbations expand the predictive support while preserving fidelity. Excessive perturbations will elevate harmful sample probability, raising the critical budget. So it is necessary to design a good perturbation.*

5.2 How Effective Is It?

We then investigate how diversified sampling affects forecasting performance across datasets in practice. Experiments are conducted on multiple TSFMs with different perturbations. The results of *Chronos* are in Figure 4, and other models' results are reported in Appendix F.

Performance Scaling under Baseline and Diversified Inference We first examine how forecasting performance evolves with increasing inference budget. As shown in Figure 4, the sampling temperature is fixed to 0.7 for all models, and the number of generated samples is varied from 2^0 to 2^7 to control the inference budget. The standard sampling serves as a fixed reference, since direct inference produces only a single prediction and therefore does not benefit from inference scaling. In contrast, introducing input perturbations enables diversified sampling and yields clear inference scaling effects. For perturbed inputs, the forecasting performance decreases as the sampling budget grows and converges to levels superior to the standard sampling, with peak improvements of up to approximately 50%. However, the relative effectiveness of different perturbations varies

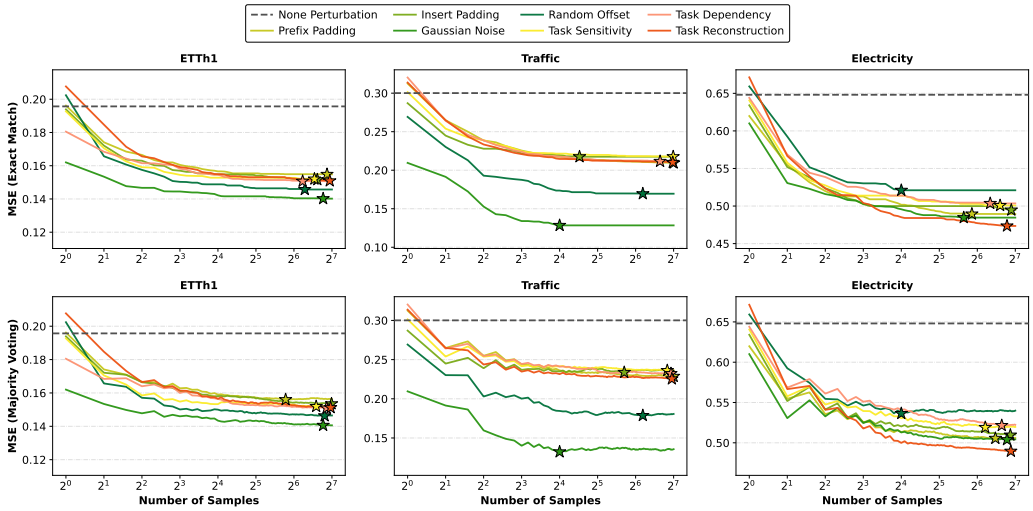


Figure 4: **Performance of *Chronos* on different datasets with effective perturbations.** The dashed line denotes single-sample error; results are reported with EM and MV aggregation.

across models, indicating that the choice of perturbation family can depend on the characteristics of the underlying model, rather than requiring any adaptation of the model itself. In conclusion, diversified sampling is a general inference mechanism to improve performance; however, customized perturbations according to the underlying model and inference settings might achieve optimal results.

Effect of Aggregation and Model-Datasets Interaction We further analyze how aggregation functions and TSFM characteristics influence the effectiveness of diversified inference. All perturbation strategies are evaluated under identical inference budgets using both EM and MV. While both aggregators preserve the overall scaling trend induced by diversified sampling, EM generally yields smoother scaling behavior and lower optimal errors in most cases. Beyond aggregation, the gains from diversified inference vary systematically across models and datasets. Models with stronger standard sampling performance generally achieve lower absolute errors. However, compared to models equipped with decoding strategies, models without decoding strategies (e.g., TimesFM and TimeMoE) exhibit relatively weaker inference scaling effects, although diversified sampling under most perturbations still improves upon standard sampling.

> 📌 ***Takeaways:*** *Inputting perturbations will expand the support, while incorporating scaling increases the hit probability, enabling diversified sampling to surpass the finite sample and achieve improvements.*

Actionable Insights Combining "When" and "How", we derive that diversified sampling should be configured through an explicit coupling between inference computation and diversity, as an example in Figure 5. Operationally, MV typically decreases with increasing N and converges to a stable level (performance typically converges around $N = 2^5$ for MV), beyond which additional sampling yields diminishing returns and can be safely terminated. And given an inference budget N, introduce perturbations while restricting them to high-relevance variants that preserve strong similarity to the original input (e.g., cosine similarity ≥ 0.92), ensuring that diversification expands predictive support without overwhelming sample fidelity. When users choose its type, we recommend using noise-based perturbations as the default, but for specific datasets, we further provide an *optional* oracle selector in the Appendix J to identify the optimal configuration if the user has computing resources. Aggregation provides an effective validity check: if aggregated performance fails to improve as N increases, the current perturbation is misaligned with the inference budget and should be adjusted.

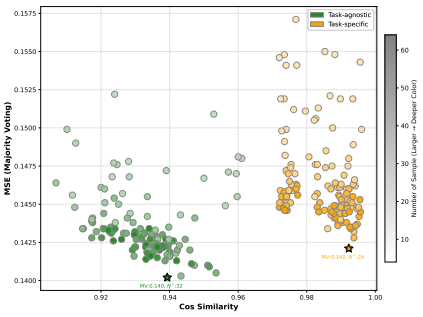


Figure 5: Computation-Diversity.

Table 2: RobustMSE evaluations of TSFMs under diversified scaling inference (the budget $N = 64$) with EM or MV. The MSE is a reference.

Strategy	Dataset	Chronos-T5-Base			Moirai-1.1-R-Base			TimesFM-2.5-200M			TimeMoE-200M		
		EM	MV	MSE	EM	MV	MSE	EM	MV	MSE	EM	MV	MSE
Task-Agnostic	ETTh1	0.1394±0.0003	0.1424±0.0005	0.1805±0.0068	0.1383±0.0004	0.1407±0.0020	0.1458±0.0075	**0.1334**±0.0003	**0.1354**±0.0006	**0.1435**±0.0053	0.2353±0.0003	0.2396±0.0002	0.2401±0.0015
	Traffic	0.0468±0.0004	0.0467±0.0004	0.0669±0.0449	0.0438±0.0015	**0.0431**±0.0004	0.0670±0.0699	**0.0437**±0.0000	0.0439±0.0002	**0.0481**±0.0066	0.3172±0.0017	0.3227±0.0008	0.3232±0.0131
	Electricity	0.2072±0.0013	0.2088±0.0026	0.2645±0.0375	0.1945±0.0002	0.1953±0.0002	0.2141±0.0325	**0.1845**±0.0007	**0.1852**±0.0003	**0.1947**±0.0101	0.8968±0.0031	0.9057±0.0012	0.9068±0.0145
	Average	0.1311±0.0007	0.1326±0.0012	0.1706±0.0297	0.1255±0.0007	0.1264±0.0009	0.1423±0.0366	**0.1205**±0.0003	**0.1215**±0.0004	**0.1288**±0.0073	0.4831±0.0017	0.4893±0.0007	0.4900±0.0097
Task-Specific	ETTh1	0.1416±0.0004	0.1445±0.0004	0.1843±0.0055	0.1413±0.0003	0.1417±0.0002	0.1478±0.0021	**0.1401**±0.0003	**0.1411**±0.0000	**0.1435**±0.0028	0.2390±0.0008	0.2435±0.0002	0.2448±0.0023
	Traffic	0.0497±0.0002	0.0501±0.0002	0.0725±0.0160	**0.0420**±0.0001	**0.0422**±0.0001	**0.0490**±0.0153	0.0441±0.0002	0.0444±0.0001	0.0500±0.0087	0.3112±0.0019	0.3271±0.0001	0.3244±0.0100
	Electricity	0.2084±0.0009	0.2139±0.0015	0.3157±0.0750	0.1919±0.0002	0.1920±0.0000	**0.2027**±0.0160	**0.1881**±0.0003	**0.1887**±0.0000	0.2054±0.0216	0.9028±0.0053	0.9248±0.0009	0.9275±0.0107
	Average	0.1332±0.0005	0.1362±0.0007	0.1908±0.0322	0.1251±0.0002	0.1253±0.0001	0.1332±0.0111	**0.1241**±0.0002	**0.1247**±0.0000	**0.1330**±0.0110	0.4843±0.0027	0.4985±0.0004	0.4989±0.0077
1st Count		0	0	0	1	2	2	5	4	4	0	0	0

6 ROBUSTMSE: A NEW METRIC

Why RobustMSE? Standard time series forecasting typically relies on deterministic metrics (e.g., MSE, MAE), assessed via a single inference pass (Hyndman & Koehler, 2006). While sufficient for discriminative models, this paradigm is inadequate for generative TSFMs operating under stochastic settings. At inference time, prediction quality is not defined by a single realization, but by the distribution of candidate sequences and decoding strategies (Holtzman et al., 2020; Chen et al., 2021). Consequently, single-run metrics exhibit high variance due to stochasticity and fail to capture the model's potential when afforded a computational budget for exploration. Based on our actionable insights, we introduce RobustMSE, a metric that reflects a TSFM's repeatable best achievable performance under a fixed inference budget (N) and an aggregator (EM or MV). Unlike MSE, which often reflects a lower-bound performance tied to unoptimized decoding, RobustMSE aggregates the headroom achievable under validated sampling strategies. This provides a fair comparison by measuring how effectively a TSFM can leverage diversified scaling inference to minimize error.

Computation of RobustMSE Following the empirical findings in Section 5.1, we exclude ineffective perturbations and restrict evaluation to a critical sample size $N = 64$. We perform diversified scaling inference for each model on varying datasets. In each trial, we record the EM and MV across all valid diversified configurations, and RobustMSE is computed as the average of $T = 5$ trials:

$$\textbf{RobustMSE}_{\texttt{EM/MV}} = \frac{1}{T}\sum_{i=1}^{T} \text{MSE}(\hat{\mathbf{Y}}^{i}_{\texttt{EM/MV}}, \mathbf{Y}). \tag{7}$$

Horizontal Comparison Table 2 reports the RobustMSE performance of mainstream TSFMs on approximately $200M$ active parameters across various datasets. These results confirm that TSFMs effectively utilize the headroom provided by diversified sampling to minimize prediction error. Overall, TimesFM demonstrates superior performance, achieving the lowest RobustMSE in 5 out of 6 comparison scenarios, as it was pre-trained on some datasets. Moirai serves as a highly competitive runner-up, frequently ranking second and exhibiting exceptional stability. In contrast, TimeMoE exhibits significantly higher variance and error margins compared to other transformer-based baselines in this evaluation, indicating potential limitations in leveraging diversified inference budgets. Furthermore, we note that the standard deviation of RobustMSE has always been small compared to MSE, indicating the stability of the metric.

7 CONCLUSION

In this work, we systematically explore the underexplored inference scaling and diversified sampling for TSFMs. By dissecting the effect of standard sampling on performance, we uncover counterintuitive findings: traditional scaling levers such as increasing TSFM parameters or extending length yield inconsistent, dataset-dependent gains, with their effects often overshadowed by the benefits of aggregating multiple inferences. To address the inherent limitations, we further formalize a theoretical framework for diversified sampling, proving that input perturbations expand the model's solution space. Guided by this theory, we design perturbation schemes and conduct extensive evaluations across TSFM architectures and real-world datasets, quantifying consistent gains in multi-step forecasting accuracy and robustness. As a practical instantiation, we introduce RobustMSE, a novel evaluation metric quantifying the headroom of TSFMs, enabling fair comparisons of models under fixed compute budgets. Collectively, our findings highlight inference design as a pivotal yet underutilized lever for optimizing TSFM performance, offering a cost-effective alternative to resource-heavy pre-training or finetuning pipelines.

8 AI USE STATEMENT

In this work, we only used generative AI tools to refine parts of the paper. We have reviewed these AI-assisted parts. All outputs were reviewed and verified by the authors, including mathematical derivations, experimental designs, code, and reported results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Eltayeb Ahmed, Uljad Berdica, Martha Elliott, Danijela Horak, and Jakob N. Foerster. Intent factored generation: Unleashing the diversity in your language model. In *The Exploration in AI Today Workshop at ICML*, pp. 1–79, 2025.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, pp. 1–42, 2024.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, pp. 235–256, 2002.

Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Alphamath almost zero: process supervision without process. In *NeurIPS*, pp. 27689–27724, 2024.

Mark Chen et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Hao Cheng, Qingsong Wen, Yang Liu, and Liang Sun. RobustTSF: Towards theory and design of robust time series forecasting with anomalies. In *ICLR*, 2024.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *ICML*, pp. 10148–10167, 2024.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, pp. 1–37, 2014.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. TimeGPT-1. *arXiv preprint arXiv:2310.03589*, 2023.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *ICML*, pp. 16115–16152, 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *NeurIPS*, pp. 30016–30030, 2022.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, pp. 1–16, 2020.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, pp. 679–688, 2006.

Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *PLOS one*, pp. e0254841, 2021.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *ACL*, pp. 2021–2031, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jongseon Kim, Hyungjoon Kim, HyunGi Kim, Dongjun Lee, and Sungroh Yoon. A comprehensive survey of deep learning for time series forecasting: architectural diversity and open challenges. *Artificial Intelligence Review*, pp. 216, 2025.

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.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *SIGIR*, pp. 95–104, 2018.

Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S Jensen, et al. Tsfm-bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *SIGKDD*, pp. 5595–5606, 2025.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *SIGKDD*, pp. 6555–6565, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. In *ACL*, pp. 157–173, 2024a.

Zichuan Liu, Zefan Wang, Linjie Xu, Jinyu Wang, Lei Song, Tianchun Wang, Chunlin Chen, Wei Cheng, and Jiang Bian. Protecting your llms with information bottleneck. In *NeurIPS*, pp. 29723–29753, 2024b.

Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, and Qingsong Wen. Explaining time series via contrastive and locally sparse perturbations. In *ICLR*, pp. 1–21, 2024c.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In *NeurIPS*, pp. 11054–11070, 2021.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *ICML*, pp. 8857–8868, 2021.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *ACL*, 2020.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *ICLR*, pp. 1–33, 2025.

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Fast best-of-N decoding via speculative rejection. In *NeurIPS*, pp. 32630–32652, 2024.

Liaoyuan Tang, Zheng Wang, Guanxiong He, Rong Wang, and Feiping Nie. Perturbation guiding contrastive representation learning for time series anomaly detection. In *IJCAI*, pp. 4955–4963, 2024.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. In *ICLR*, 2017.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. In *ACL*, pp. 2153–2162, 2019.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *ICLR*, pp. 1–27, 2024a.

Tianchun Wang, Yuanzhou Chen, Zichuan Liu, Jonathan Light, Weiyang Liu, Haifeng Chen, Xiang Zhang, and Wei Cheng. On the effect of sampling diversity in scaling llm inference. In *UAI*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. In *NeurIPS*, pp. 469–498, 2024b.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP*, pp. 6382–6388, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pp. 24824–24837, 2022.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *ICML*, pp. 53140–53164, 2024.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *ICLR*, 2023.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *ICLR*, 2025.

Yiyuan Yang, Zichuan Liu, Lei Song, Kai Ying, Zhiguang Wang, Tom Bamford, Svitlana Vyetrenko, Jiang Bian, and Qingsong Wen. Time-RA: Towards time series reasoning for anomaly with llm feedback. In *Findings of ACL*, 2026.

Qingren Yao, Chao-Han Huck Yang, Renhe Jiang, Yuxuan Liang, Ming Jin, and Shirui Pan. Towards neural scaling laws for time series foundation models. In *ICLR*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, pp. 11809–11822, 2023.

Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. In *NeurIPS*, 2019.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI*, pp. 11121–11128, 2023.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, pp. 12697–12706, 2021.

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 *AAAI*, pp. 11106–11115, 2021.

A Related Work

Time Series Foundation Model (TSFM) has established a pre-training paradigm that learns domain-agnostic temporal dynamics from large-scale and heterogeneous time series, and recent progress has demonstrated that such large-scale pre-training enables models to acquire cross-domain temporal representations across diverse data sources (Li et al., 2025; Ansari et al., 2024; Liang et al., 2024; Woo et al., 2024; Shi et al., 2025). Existing TSFM architectures can be grouped into decoder-based and encoder-based formulations. Decoder-based models, such as Chronos (Ansari et al., 2024), TimesFM (Das et al., 2024), and TimeMoE (Shi et al., 2025), treat forecasting as conditional language modeling, using normalization, symbolization, and sparse expert mixtures to support multi-step forecasting and cross-domain generalization. Encoder-based models, such as Moirai (Woo et al., 2024), learn unified representations across variables, frequencies, and tasks through flexible attention mechanisms. Additional developments (Garza et al., 2023; Goswami et al., 2024; Wang et al., 2024b) further expand the capabilities of TSFMs. Across these architectures, the common strategy is to leverage large-scale pre-training to acquire transferable temporal patterns and improve robustness and generalization in downstream tasks.

Scaling Inference improves language models' performance by increasing test-time computation via deeper decoding, broader search, and lightweight aggregation without changing model parameters (Wu et al., 2025; Wang et al., 2025). Prior studies have shown that multi-pass reasoning (Wei et al., 2022; Yao et al., 2023) and test-time ensembling (Vijayakumar et al., 2017; Wang et al., 2023) can substantially enhance model performance without additional training. For example, these approaches implement deeper decoding or stochastic constraint relaxation by extending chain- or tree-structured searches (Wei et al., 2022; Yao et al., 2023), promoting sampling to generate different candidate sequences through randomness and diversity (Holtzman et al., 2020; Vijayakumar et al., 2017) to expand alternative inference paths. However, in time series, although existing work has studied training scaling laws (Liang et al., 2024; Kottapalli et al., 2025; Yao et al., 2025), the inference phase remains unchanged and lacks expanding computational depth or breadth. This gap motivates our systematic investigation of scaling inference, where we examine temporal expansion strategies and evaluate their effects on prediction accuracy, robustness, and generalization.

Diversified Sampling provides a mechanism to enhance model robustness and uncertainty handling by generating multiple heterogeneous outputs during inference (Vijayakumar et al., 2017). Such strategies have been extensively explored in generative modeling (Wang et al., 2025; Sun et al., 2024; Wang et al., 2023) where multi-path generation and voting-based aggregation can substantially improve reliability in reasoning and math tasks. Similarly, existing time series robustness studies primarily focus on assessing model sensitivity through input perturbations (Cheng et al., 2024; Tang et al., 2024). These methods capture signal level variability but lack systematic perturbation studies, and leave unexplored whether diversifying samples under fixed model parameters can enhance forecasting accuracy or robustness. Consequently, current time series inference frameworks remain unable to fully exploit the potential benefits of diversified sampling (Yao et al., 2025; Kottapalli et al., 2025). In this work, we incorporate diversified sampling mechanisms in time series inference to expand the feasible solution space and enhance prediction performance under both task-specific and task-agnostic perturbations.

B Omitted Proofs

B.1 Asymptotic Analysis ($N \to \infty$)

We first characterize the theoretical lower bound of the loss when the sample budget approaches infinity. This analysis relies on the support sets of the respective distributions.

Assumption B.1 (**Identity Inclusion**). The perturbation distribution $Q(\mathbf{X}'|\mathbf{X}, \pi_S)$ assigns non-zero probability mass (or density) to the original input $\mathbf{X}$, or equivalently, to an ϵ-neighborhood of $\mathbf{X}$.

Let $\mathcal{S}_{\text{std}} = \text{supp}(P(\cdot|\mathbf{X}, \zeta))$ and $\mathcal{S}_{\text{div}} = \text{supp}(P_{\text{mix}})$ denote the support sets of the standard and diversified distributions, respectively. Under Assumption B.1, it logically follows that $\mathcal{S}_{\text{std}} \subseteq \mathcal{S}_{\text{div}}$.

Proposition B.2 (**Asymptotic Lower Bound**). *As the sample size $N \to \infty$, the expected minimum EM of diversified sampling is strictly bounded by that of standard sampling:*

$$\lim_{N\to\infty} EM_N^{div} \leq \lim_{N\to\infty} EM_N^{std}. \tag{8}$$

The inequality holds strictly ($<$) if there exists $\mathbf{Y}^* \in \mathcal{S}_{div} \setminus \mathcal{S}_{std}$ such that $\mathcal{L}(\mathbf{Y}^*, \mathbf{Y}) < \inf_{\hat{\mathbf{Y}} \in \mathcal{S}_{std}} \mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y})$.

Proof. According to extreme value theory, the minimum statistic of i.i.d. samples converges to the essential infimum of the distribution's support. Thus:

$$\lim_{N\to\infty} \mathrm{EM}_N^{\text{std}} = \inf_{\hat{\mathbf{Y}} \in \mathcal{S}_{\text{std}}} \mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y}), \tag{9}$$

$$\lim_{N\to\infty} \mathrm{EM}_N^{\text{div}} = \inf_{\hat{\mathbf{Y}} \in \mathcal{S}_{\text{div}}} \mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y}). \tag{10}$$

Since $\mathcal{S}_{\text{std}} \subseteq \mathcal{S}_{\text{div}}$, and the infimum operator is monotonic (i.e., $\inf_{\mathbf{X}\in A} f(\mathbf{X}) \geq \inf_{\mathbf{X}'\in B} f(\mathbf{X}')$ if $A \subseteq B$), the proposition holds. Geometrically, this implies diversified sampling explores an expanded manifold that may contain solutions closer to the ground truth $\mathbf{Y}$, which are inaccessible under the fixed configuration ζ. □

B.2 Finite Sample Analysis ($N < \infty$)

In the finite regime, simply expanding the support is insufficient due to the variance introduced by perturbations. We must balance the *potential for lower loss* against the *probability of sampling it*. We derive a critical sample size N^* using a simplified probabilistic model.

Proposition B.3 (Critical Sample Threshold). *Let $\mathcal{L}_0$ be the baseline EM of standard sampling at one time. Assume the diversified distribution yields a reduced loss $\mathcal{L}_{good} < \mathcal{L}_0$ with probability ρ (valid perturbation), and an increased loss $\mathcal{L}_{bad} > \mathcal{L}_0$ with probability $1 - \rho$. Diversified sampling outperforms standard sampling in expectation (i.e., $\mathbb{E}[EM_N^{div}] < \mathcal{L}_0$) if and only if the sample size N exceeds a critical threshold N^*:*

$$N > N^* = \frac{\ln\left(\frac{\mathcal{L}_{bad} - \mathcal{L}_{good}}{\mathcal{L}_0 - \mathcal{L}_{good}}\right)}{\ln\left(\frac{1}{1-\rho}\right)}. \tag{11}$$

Proof. The expected minimum EM for diversified sampling is the weighted sum of the loss when at least one "good" sample is found, and the loss when all samples are "bad":

$$\mathbb{E}[\mathrm{EM}_N^{\text{div}}] = \underbrace{(1-\rho)^N}_{\text{All failure}} \mathcal{L}_{\text{bad}} + \underbrace{\left[1 - (1-\rho)^N\right]}_{\text{At least one success}} \mathcal{L}_{\text{good}}. \tag{12}$$

We seek the condition where $\mathbb{E}[\mathrm{EM}_N^{\text{div}}] < \mathcal{L}_0$. Substituting the expectation:

$$(1-\rho)^N \mathcal{L}_{\text{bad}} + \mathcal{L}_{\text{good}} - (1-\rho)^N \mathcal{L}_{\text{good}} < \mathcal{L}_0, \tag{13}$$

grouping terms by $(1-\rho)^N$:

$$(1-\rho)^N (\mathcal{L}_{\text{bad}} - \mathcal{L}_{\text{good}}) < \mathcal{L}_0 - \mathcal{L}_{\text{good}}. \tag{14}$$

Then, dividing by $(\mathcal{L}_{\text{bad}} - \mathcal{L}_{\text{good}})$ (which is positive) and taking the natural logarithm:

$$N \ln(1-\rho) < \ln\left(\frac{\mathcal{L}_0 - \mathcal{L}_{\text{good}}}{\mathcal{L}_{\text{bad}} - \mathcal{L}_{\text{good}}}\right). \tag{15}$$

Since $\ln(1-\rho) < 0$, dividing by it reverses the inequality:

$$N > \frac{\ln\left(\frac{\mathcal{L}_{\text{bad}} - \mathcal{L}_{\text{good}}}{\mathcal{L}_0 - \mathcal{L}_{\text{good}}}\right)}{-\ln(1-\rho)} = \frac{\ln\left(\frac{\mathcal{L}_{\text{bad}} - \mathcal{L}_{\text{good}}}{\mathcal{L}_0 - \mathcal{L}_{\text{good}}}\right)}{\ln\left(\frac{1}{1-\rho}\right)} = N^*. \tag{16}$$

This concludes the proof. □

Table 3: Description of time series perturbation strategies with analogous methods in LMs.

Taxonomy	Perturbation Type	Method	Definition	Default Setting	Analogous Method in LLMs	Ref.
Task-Agnostic	Structural Perturbations	Prefix Padding	$\mathbf{X}' = [\mathbf{c}, \mathbf{X}],\ i = 0$	$\ell \sim \mathcal{U}\{16, 32\}$; $c \sim \mathcal{U}(-1, 1)$	**Original:** "The result is significant." **Perturbed:** "According to the professor, the result is significant."	(Perez et al., 2021; Zhao et al., 2021)
		Suffix Padding	$\mathbf{X}' = [\mathbf{X}, \mathbf{c}],\ i = L$	$\ell \sim \mathcal{U}\{16, 32\}$; $c \sim \mathcal{U}(-1, 1)$	**Original:** "The experiment failed." **Perturbed:** "The experiment failed. [END OF REPORT]"	(Wallace et al., 2019)
		Middle Insertion	$\mathbf{X}' = [\mathbf{X}_{1:i}, \mathbf{c}, \mathbf{X}_{i+1:L}]$, $0 < i < L$	$\ell \sim \mathcal{U}\{16, 32\}$; $i \sim \mathcal{U}(0.3L, 0.7L)$, $c = \mathbf{x}_i$	**Original:** "The policy is harmful." **Perturbed:** "The policy is, this is a neutral statement, harmful."	(Jia & Liang, 2017; Liu et al., 2024a)
	Noise Perturbations	Gaussian Noise	$\mathbf{X}' = \mathbf{X} + \mathcal{N}(0, (\eta\sigma)^2\mathbf{I})$	$\eta \in \{0.05, 0.1, 0.2\}$	$e_i' = e_i + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$	(Liu et al., 2024c)
		Random Offset	$\mathbf{x}_i' = \mathbf{x}_i + c,\ 0 < i < L$	$c \in \mathcal{U}\{-\sigma, \sigma\}$	**Original:** "The result is important." **Perturbed:** "The result is crucial."	(Wei & Zou, 2019)
		Missing Data	$\mathbf{X}' = \mathbf{M} \cdot \mathbf{X}$	Mask $\mathbf{M} \in \{0, 1\}^{L \times D}$	**Original:** "The market reacts strongly." **Perturbed:** "The market [MASK] strongly."	(Liu et al., 2024b)
Task-Specific	Task-specific Perturbations	Task Sensitivity	$\mathbf{X}' = \mathbf{X} + \eta\pi_S^s(\mathbf{X}, t)$	$\eta \sim \mathcal{U}(0.01, 0.05)$	-	Refer to Algorithm1
		Task Dependency	$\mathbf{X}' = \mathbf{X} + \eta\pi_S^d(\mathbf{X}, t)$	$\eta \sim \mathcal{U}(0.01, 0.05)$	-	Refer to Algorithm2
		Task Reconstruction	$\mathbf{X}' \sim \pi_S^r(\cdot \mid \mathbf{X})$	$\tau \in \{0.3, 0.4, \cdots, 0.9\}$	**Original:** "The experiment failed due to noise." **Paraphrase:** "Noise caused the experiment to be unsuccessful."	(Ribeiro et al., 2020; Wang et al., 2025)

C The Illustration for Diversified Strategies

Task-Agnostic Perturbations. This category introduces controlled perturbations to raw data sequences without accounting for the model's designated forecasting objective. The primary goals of these methods are to assess model robustness and the impact of input diversity. Task-agnostic perturbations are further partitioned into two orthogonal subclasses: (i) *Structural Perturbations*, encompassing Prefix Padding, Suffix Padding, and Middle Insertion, which tamper with the boundary constraints or internal topological structure of the sequence. Notably, the degree of diversity induced by such perturbations exhibits a positive correlation with the perturbation length (Iwana & Uchida, 2021). (ii) *Noise Perturbations* which cover Gaussian Noise, Random Offset, and Missing Data implantation, where the intensity of perturbation η serves as the direct control variable for tuning the level of sample diversity. This suite is widely recognized as a standard benchmark for evaluating model robustness in time series analysis (Cheng et al., 2024).

Task-Specific Perturbations. To probe the model's prediction mechanism and simulate realistic contextual distribution shifts, we design three dedicated perturbation injection paradigms tailored to the inherent characteristics of time series tasks: (i) *Task Sensitivity* Perturbation mimics latent time distribution drifts by systematically modulating the global trend of the input time series. This perturbation strategy aligns with real-world scenarios where temporal data often undergoes gradual trend shifts due to external environmental factors (Gama et al., 2014). (ii) *Task Dependency* Perturbation embeds a predefined, task-indicative pattern into the original sequence. The core objective of this perturbation is to expose potential patterns associated with the exploitation of spurious task-specific features (Yang et al., 2026). (iii) *Task Reconstruction* leverages a pre-specified TSFM to reconstruct time series $\mathbf{X}' \sim f(\cdot|\mathbf{X}, \zeta)$ under a given configuration ζ. By sampling from the reconstructed time series, this paradigm generates inputs that are semantically congruent with the original sequence yet structurally divergent, thereby augmenting the diversity of the sample pool (Yoon et al., 2019; Rasul et al., 2021).

While the definitions above provide a structural overview, the core motivation for this taxonomy lies in the fundamental trade-off between universality and fidelity in inference scaling. Task-Agnostic strategies serve as a cost-effective, universal baseline, allowing us to quantify the model's inherent robustness against pure stochasticity without assuming any domain priors. In contrast, Task-Specific strategies are designed to ensure semantic consistency. By leveraging intrinsic temporal properties (such as trend and seasonality), they constrain the exploration within a plausible manifold. This distinction is essential for our analysis, as it enables us to isolate whether performance gains stem merely from tolerance to random noise or from structurally valid exploration that aligns with the time

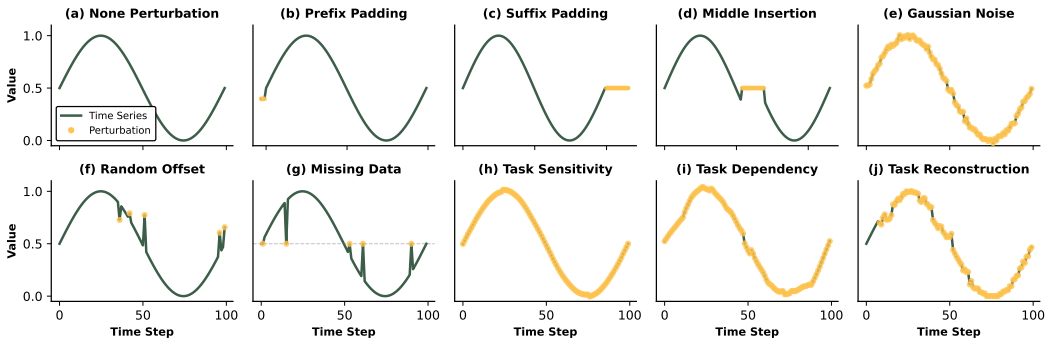


Figure 6: Visualization of time series with different perturbations.

series' physical nature. *Furthermore, we classify failure perturbations as MSE exceeding 20% of the standard sampling.* This threshold is empirically determined based on the fidelity-diversity buffer observed in Figure 3 and the transition behavior in Table 1. Beyond this threshold, the required N^* increases rapidly, rendering diversified inference impractical under realistic computation budgets.

Furthermore, we summarize nine diversified strategies π_S considered in this work in Table 3, providing a unified taxonomy of time series perturbations together with their analogous counterparts in natural language processing. We categorize all perturbation methods into *Task-Agnostic* and *Task-Specific* classes according to whether the injected modifications are agnostic to the downstream forecasting objective. Throughout the Table 3, $\mathbf{X}'$ denotes the perturbed time series derived from the original input $\mathbf{X}$ by π_S. The index i indicates the perturbation position, where $i = 0$ corresponds to prefix insertion, $i = L$ to suffix insertion, and $0 < i < L$ to middle insertion. $\mathbf{c}^{\ell \times D}$ is padding for structural perturbations, where ℓ and c are the perturbation length and padding value, respectively. For noise perturbations, σ denotes the standard deviation of $\mathbf{X}$ and η controls the perturbation intensity. For task-specific perturbations, π_S^s, π_S^d, π_S^r represent parameterized perturbation functions, respectively, where π_S^r uses a TSFM (Chronos in here) for time series reconstruction. In detail, π_S^s denotes structure-sensitive perturbations modulated by task difficulty, where regions that are more difficult to predict or intrinsically unstable receive stronger perturbations, while more predictable and stable structures are perturbed more mildly. π_S^d means structure-aligned, task-dependent perturbations, where perturbation directions are guided by the task-relevant structural components of the input, so that the perturbations induce maximal changes in the task output while preserving the natural data geometry. In order to visually demonstrate different strategies, we present a schematic diagram for toy data in Figure 6.

Algorithm 1: Structure-sensitive Perturbation π_S^s Modulated by Task Difficulty

Input: Original input sequence $\mathbf{X}$, decomposition period p, intensity control η
Output: Perturbed time series $\mathbf{X}'$

// 1. Decompose input into structural components for sensitivity analysis
$(\mathbf{T}, \mathbf{S}, \mathbf{R}) \leftarrow \text{STL}(\mathbf{X}, p)$;
// 2. Modulate perturbation strength by task difficulty and intrinsic instability
$\mathbf{f}_T \leftarrow |\nabla \mathbf{T}|$; // Heightened sensitivity to trend shifts
$\mathbf{S}_{shift} \leftarrow \text{Roll}(\mathbf{S}, p)$;
$\mathbf{f}_S \leftarrow |\mathbf{S} - \mathbf{S}_{shift}|$; // Capturing seasonal non-stationarity
$\mathbf{f}_R \leftarrow |\mathbf{R}|$; // Identifying unpredictable residual regions
// 3. Consolidate perturbations with sign and mean constraints
$\mathbf{f} \leftarrow \mathbf{f}_T + \mathbf{f}_S + \mathbf{f}_R$;
$\mathbf{f} \leftarrow \text{sign}(\mathbf{X} + \epsilon) \odot |\mathbf{f}|$;
$\mathbf{f} \leftarrow \mathbf{f} - \text{mean}(\mathbf{f})$;
// 4. Final intensity normalization for output consistency
$\eta \leftarrow \eta \cdot \frac{\|\mathbf{X}\|_2}{\|\mathbf{f}\|_2 + \epsilon}$;
$\mathbf{X}' \leftarrow \mathbf{X} + \eta \mathbf{f}$;
return $\mathbf{X}'$

Algorithm 2: Structure-aligned Task-dependent Perturbation π_S^d

Input: Original input sequence $\mathbf{X}$, decomposition period p, intensity control η, local window w
Output: Perturbed time series $\mathbf{X}'$

// 1. Extract task-relevant structural components via STL decomposition
$(\mathbf{T}, \mathbf{S}, \mathbf{R}) \leftarrow \text{STL}(\mathbf{X}, p)$;
// 2. Guide perturbation directions to preserve natural data geometry
$\mathbf{f}_T \leftarrow \nabla \mathbf{T}$; // Trend-aligned directional guidance
$\mathbf{f}_S \leftarrow |\mathbf{S}|$; // Seasonal structural alignment
$\mathbf{f}_R \leftarrow \text{LocalStd}(\mathbf{R}, w)$; // Residual local variance capturing geometry
// 3. Enforce sign consistency and zero-mean constraints
$\mathbf{f} \leftarrow \mathbf{f}_T + \mathbf{f}_S + \mathbf{f}_R$;
$\mathbf{f} \leftarrow \text{sign}(\mathbf{X} + \epsilon) \odot |\mathbf{f}|$;
$\mathbf{f} \leftarrow \mathbf{f} - \text{mean}(\mathbf{f})$;
// 4. Scale perturbation intensity according to the input signal-to-noise ratio
$\eta \leftarrow \eta \cdot \frac{\|\mathbf{X}\|_2}{\|\mathbf{f}\|_2 + \epsilon}$;
$\mathbf{X}' \leftarrow \mathbf{X} + \eta \mathbf{f}$;
return $\mathbf{X}'$

D Experimental Settings Details

The code is available in https://anonymous.4open.science/r/Diversified-Scaling-Inference-in-TSFMs-8FDE/.

D.1 DATASETS

We evaluate the universal forecasting and inference capabilities of TSFMs on widely used public datasets, including ETTh1, ETTm1, Electricity, and Traffic from TSLib[2]. Additionally, to further assess the generality of our findings, we extend the evaluation to Exchange Rate and Weather in Appendix F.4. These datasets span diverse sampling frequencies and domain characteristics, and are commonly adopted in prior long-term forecasting literature (Liang et al., 2024; Yao et al., 2025; Wang et al., 2024a). Table 4 summarizes the datasets used in our inference-only evaluation. For all datasets, we extract the univariate Output Target (OT) as the forecasting variable (i.e., $D = D_{\text{out}} = 1$), as some TSFMs only support univariate prediction. Following prior benchmarks (Wu et al., 2023), each dataset is originally evaluated under multiple prediction horizons. In this work, unless otherwise specified, we fix the prediction horizon to $H = 96$ and perform fixed-horizon forecasting for all inference and robustness evaluations, in order to ensure fair and consistent comparisons across different models.

Table 4: Time series forecasting datasets for diversified scaling inference experiments.

Dataset	Dimension	Dataset Size	Frequency	Prediction Horizon
ETTh1	7	(8640, 2880, 2880)	Hourly	96
ETTm1	7	(34560, 11520, 11520)	15 mins	96
Electricity	321	(18412, 2630, 5261)	Hourly	96
Traffic	862	(12280, 1754, 3509)	Hourly	96
Weather	21	(36792, 5271, 10540)	10-min	96
Exchange Rate	8	(5311, 759, 1518)	Daily	96

D.2 TIME SERIES FOUNDATION MODELS (TSFMS) ARCHITECTURES

We evaluates four representative TSFMs, including TimesFM (Liang et al., 2024), Chronos (Ansari et al., 2024), Time-MoE (Shi et al., 2025) and Moirai (Woo et al., 2024), where all used versions are summarized in the Table 5. Except for scaled model size variants, the default TSFM we use is indicated with a gray background. Specifically, each model is invoked through a common inference interface, receiving identical input preprocessing, context length, prediction horizon, sliding-window stride, decoding strategy, and perturbation configurations. All TSFMs are evaluated under a unified, inference-only framework that standardizes the forecasting protocol across architectures (Bommasani et al., 2021).

Within a diversified scaling inference framework, distinguishing between probabilistic and deterministic TSFMs is crucial, as it fundamentally dictates their inference scaling behavior. Chronos and Moirai are natively probabilistic, utilizing multi-sample inference as their standard mode. In contrast, TimesFM and Time-MoE are essentially deterministic. This distinction leads to two different paradigms for diversity generation: for deterministic models, input perturbation is the sole mechanism to introduce diversity. Conversely, for probabilistic models, input perturbation acts as an incremental expansion of an already-diverse sampling base. By further broadening the generative support set, perturbation-induced diversity tends to yield more significant and compounding performance improvements.

D.3 EXPERIMENTAL SETTING

For the scaling inference experiments, all are under a unified inference-only evaluation, *without any model finetuning or architectural modification.* Our experiments are conducted on a single NVIDIA H100 GPU, and the inference process is highly amenable to parallelization. The detailed experimental settings are provided in Table 6. Specifically, we adopt a sliding window **forecasting protocol** over each time series $\{\mathbf{x}_1, \ldots, \mathbf{x}_L\}$ (Lai et al., 2018). Each forecast sample is set to a length $L = 512$ of the context window and a prediction horizon of $H = 96$ as defaults (Zhou et al., 2021). We also vary $L \in \{32, 64, 128, 256, 512, 1024\}$ in the scaling inference experiments. The sliding window

[2] https://github.com/thuml/Time-Series-Library

Table 5: Summary of TSFMs and links used in our diversified scaling inference experiments.

Model	Name	Size	HuggingFace Link
TimesFM	TimesFM-2.5-200M	200M	https://huggingface.co/google/timesfm-2.5-200m-pytorch
	TimesFM-2.0-500M	500M	https://huggingface.co/google/timesfm-2.0-500m-pytorch
Moirai	Moirai-1.1-R-Small	55M	https://huggingface.co/Salesforce/moirai-1.1-R-small
	Moirai-1.1-R-Base	365M	https://huggingface.co/Salesforce/moirai-1.1-R-base
	Moirai-1.1-R-Large	1.24B	https://huggingface.co/Salesforce/moirai-1.1-R-large
Chronos	Chronos-T5-Tiny	8M	https://huggingface.co/amazon/chronos-t5-tiny
	Chronos-T5-Mini	20M	https://huggingface.co/amazon/chronos-t5-mini
	Chronos-T5-Small	46M	https://huggingface.co/amazon/chronos-t5-small
	Chronos-T5-Base	200M	https://huggingface.co/amazon/chronos-t5-base
	Chronos-T5-Large	710M	https://huggingface.co/amazon/chronos-t5-large
TimeMoE	TimeMoE-50M	113M	https://huggingface.co/Maple728/TimeMoE-50M
	TimeMoE-200M	453M	https://huggingface.co/Maple728/TimeMoE-200M

advances along the time axis with a stride of 32, producing overlapping but distinct time series samples. The **decoding strategy** is stochastic decoding with top-$p = 1.0$, and we set a maximum of $N = 128$ inferences per context window to scale inference. The temperature τ in TSFMs is usually set to 0.7, but in the scaling inference experiment, we also explored values between 0.0 to 1.2 for decoding diversity. For **diversified sampling**, each single forecast sample is obtained through a π_S of six task-agnostic or three task-specific perturbations, and the results are aggregated through repeated decoding in the input of a kind of independent perturbation. The detailed configurations of diversified strategies are in Table 3.

Table 6: Configurations of our scaling inference experiments for all TSFMs.

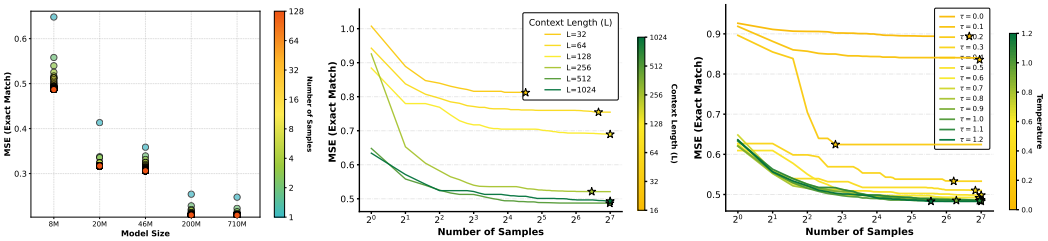
Category	Variable Name	Variable Value
Forecasting Protocol	Prediction Horizon	96
	Context Length	$\{32, 64, 128, 256, 512, 1024\}$
	Sliding Window Stride	32
	Aggregation Function	{Exact Match, Majority Voting}
Decoding Strategy	Decoding Type	Stochastic Decoding
	Temperature	$\{0.0, 0.1, 0.2, \ldots, 1.2\}$
	Top-p	1.0
	Numbers of Sampling	128
Diversified Sampling	Task-agnostic Perturbations	{Prefix, Suffix, Insert, Missing, Gaussian, Random}
	Task-specific Perturbations	{Sensitivity, Dependency, Reconstruction}

E Additional Scaling Inference Experiments

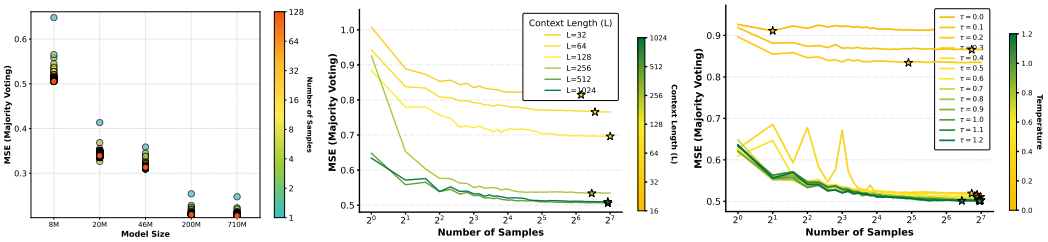
In this section, we provide a detailed relationship for each type of TSFMs between the forecasting task performance and the amount of compute expended during various model families, model sizes, and decoding strategies.

E.1 Chronos

Chronos exhibits dataset-dependent inference scaling behavior that partially aligns with but does not strictly follow standard inference scaling expectations. The results are shown in Figures 2, 7, and 8. We observe that the optimal model size shifts systematically with dataset complexity, specifically, medium or higher-sized models achieve the lowest error on simpler datasets. The results scaling curves are irregular, indicating the absence of a universal scaling law. Increasing context length generally improves performance across all datasets, yet the magnitude of improvement varies substantially with dataset complexity. As sequence complexity increases, the optimal sampling temperature decreases, suggesting that Chronos benefits from reduced output diversity in more volatile settings. Overall, Chronos demonstrates structured but irregular inference scaling, where optimal configurations emerge from dataset-specific interactions rather than predictable scaling trends.

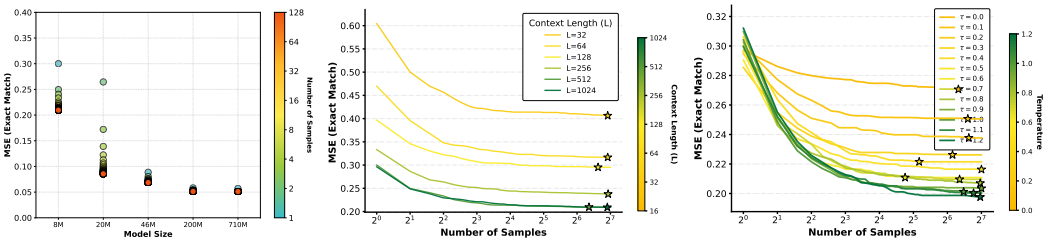


(a) EM vs. Model Size (b) EM vs. Context Length (c) EM vs. Temperature

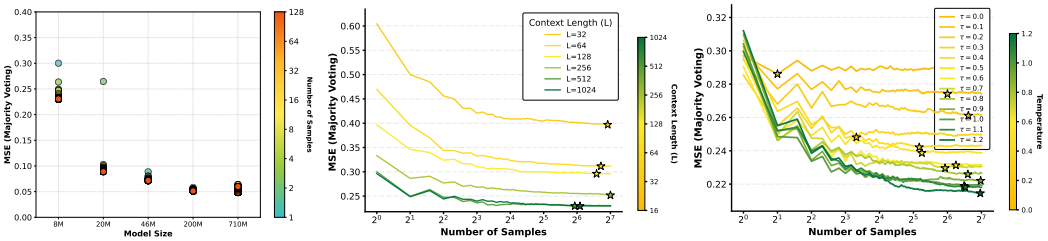


(d) MV vs. Model Size (e) MV vs. Context Length (f) MV vs. Temperature

Figure 7: *Chronos* performance on *Electricity* under varying scaling factors and aggregation methods.



(a) EM vs. Model Size (b) EM vs. Context Length (c) EM vs. Temperature

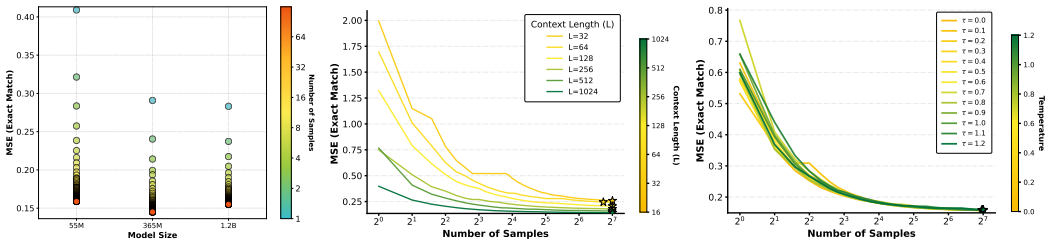


(d) MV vs. Model Size (e) MV vs. Context Length (f) MV vs. Temperature

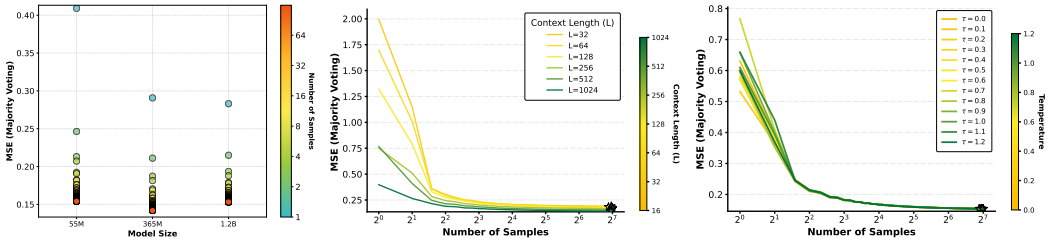
Figure 8: *Chronos* performance on *Traffic* under varying scaling factors and aggregation methods.

E.2 TIMESFM

Due to the lack of temperature and decoding strategies in TimesFM, scaling inference is ineffective. Thus, we only investigate the impact of context length on the model here. The results are shown in Table 7, with the best results in **bold** and second-best underline. Since TimesFM was pre-trained on the Electricity and Traffic datasets (Das et al., 2024), its predictive performance on these two datasets is generally superior, and the MSE shows a significant decreasing trend as L increases, demonstrating relatively stable scaling benefits. On both ETTh1 and ETTm1, the forecasting error generally decreases with increasing context length, suggesting that TimesFM broadly conforms to the inference scaling law. Nevertheless, local non-monotonic variations (e.g., the optimal performance at $L = 256$ on ETTh1) indicate that the benefits of longer contexts may saturate or fluctuate at specific scales.

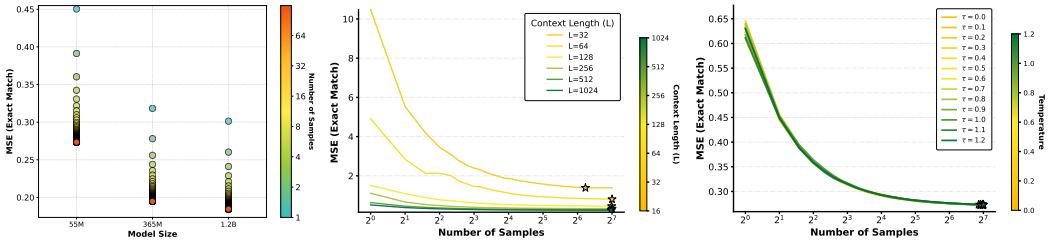


(a) EM vs. Model Size (b) EM vs. Context Length (c) EM vs. Temperature

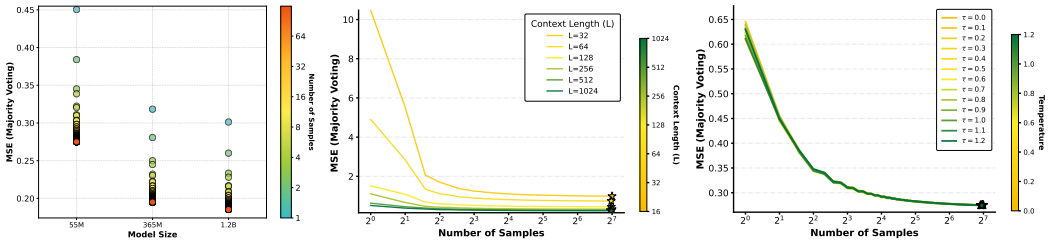


(d) MV vs. Model Size (e) MV vs. Context Length (f) MV vs. Temperature

Figure 9: *Moirai* performance on *ETTh1* under varying scaling factors and aggregation methods.



(a) EM vs. Model Size (b) EM vs. Context Length (c) EM vs. Temperature



(d) MV vs. Model Size (e) MV vs. Context Length (f) MV vs. Temperature

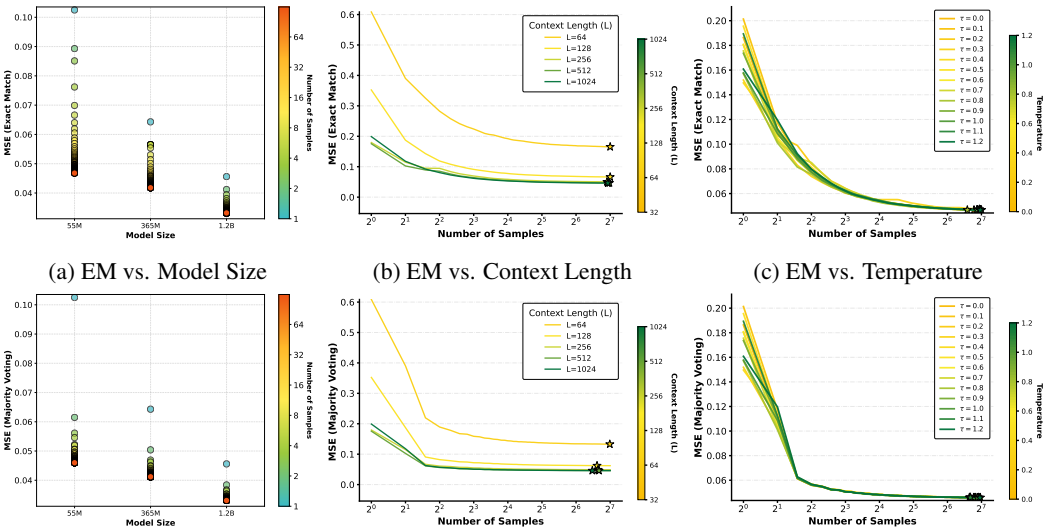
Figure 10: *Moirai* performance on *Electricity* under varying scaling factors and aggregation methods.

Table 7: *TimesFM* performance on all datasets under varying context length (MSE/MAE).

Dataset	$L = 32$		$L = 64$		$L = 128$		$L = 256$		$L = 512$		$L = 1024$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.1693	0.3041	0.1607	0.2936	0.1467	0.2819	**0.1376**	**0.2728**	0.1411	0.2757	0.1404	0.2729
ETTm1	0.1028	0.2308	0.0973	0.2253	0.0941	0.2225	0.0853	0.2080	0.0752	0.1934	**0.0707**	**0.1867**
Electricity	0.8421	0.6881	0.6461	0.5942	0.3119	0.3944	0.2151	0.3240	0.1887	0.2982	**0.1857**	**0.2956**
Traffic	0.2379	0.3373	0.1283	0.2390	0.0622	0.1485	0.0494	0.1208	0.0440	0.1108	**0.0400**	**0.1036**

E.3 MOIRAI

Moirai exhibits a relatively regular inference scaling behavior, which is shown in Figures 9,10, and 11. As the model size increases, the prediction error generally decreases. However, on the ETTh1 dataset, the model achieves optimal performance at a medium scale, and further increasing the model size



(a) EM vs. Model Size (b) EM vs. Context Length (c) EM vs. Temperature

(d) MV vs. Model Size (e) MV vs. Context Length (f) MV vs. Temperature

Figure 11: *Moirai* performance on *Traffic* under varying scaling factors and aggregation methods.

Table 8: *TimeMoE* (50M vs 200M) performance across different context lengths (MSE/MAE).

Dataset	TimeMoE-50M								TimeMoE-200M							
	$L = 128$		$L = 256$		$L = 512$		$L = 1024$		$L = 128$		$L = 256$		$L = 512$		$L = 1024$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.2429	0.3777	0.2423	0.3771	0.2412	0.3761	**0.2403**	**0.3755**	0.2443	0.3787	0.2432	0.3778	0.2412	0.3760	**0.2403**	**0.3755**
ETTm1	0.1192	0.2515	0.1185	0.2508	0.1178	**0.2500**	**0.1177**	**0.2500**	0.1199	0.2518	0.1191	0.2511	0.1183	0.2502	**0.1181**	**0.2500**
Electricity	**0.8757**	**0.7023**	0.9032	0.7151	0.9099	0.7183	0.9144	0.7207	**0.8781**	**0.7032**	0.9004	0.7137	0.9110	0.7183	0.9201	0.7222
Traffic	0.3311	0.4200	0.3274	0.4160	**0.3205**	**0.4118**	0.3222	0.4138	0.3261	0.4167	**0.3255**	**0.4158**	0.3260	0.4163	0.3281	0.4184

leads to an increase in MSE, indicating that scaling the model size does not always yield monotonic improvements. In terms of context length, when the number of samples is small, the MSE decreases significantly with increasing context length. As the number of samples increases, the prediction error decreases rapidly and converges to a similar level, with an overall performance improvement of up to approximately 90%. Notably, with a small number of samples, the MV aggregation function shows significantly faster error reduction and convergence compared to EM. Regarding the temperature setting, because Moirai's output distribution is not explicitly divided by the temperature coefficient τ during inference (refer to Eq.4 in (Woo et al., 2024)), making it difficult to effectively sample the distribution under low conditions. This renders the temperature's regulatory effect on inference scaling largely ineffective in Moirai. Therefore, regardless of the temperature value or the number of samples, the model performance shows a significant convergence trend.

E.4 TIMEMOE

Due to the lack of temperature and decoding strategies in TimeMoE, scaling inference is ineffective. Thus, we only investigate the impact of context length and model size on the model here. The results are shown in Table 8, with the best results in **bold** and second-best underline. We further compared the performance differences of TimeMoE at different model sizes as the context length varied. Taking the Traffic dataset as an example, *TimeMoE-50M* achieved the best results at $L = 512$ and the second-best at $L = 1024$, while *TimeMoE-200M*'s optimal performance occurred at $L = 256$, followed by $L = 512$. This difference indicates that different model sizes have inconsistent sensitivity to context length. Overall, regardless of model size, the performance trend of TimeMoE across different datasets as the context length increases remains largely consistent, but it does not exhibit the standard inference scaling behavior of monotonically improving with L, suggesting that simply increasing the context length is not sufficient to guarantee stable performance improvement.

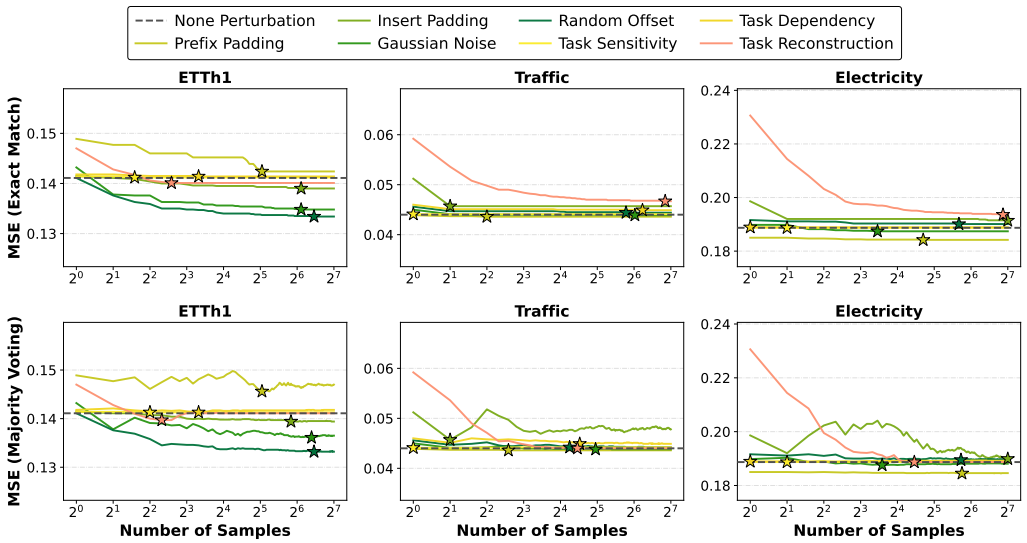


Figure 12: Performance of *TimesFM* on different datasets with effective perturbations.

F Additional Diversified Scaling Inference Experiments

We provide all the results of how diversified scaling influences the inference performance under more TSFMs and varying datasets in this part, and show the characteristics of different models with diversified sampling.

F.1 TimesFM

Figure 12 reports the diversified inference results of TimesFM, a TSFM without decoding strategies. Since TimesFM was pre-trained on the Electricity and Traffic datasets, we primarily focus on its performance on the dataset ETTh1. Experimental results show that on ETTh1, all effective input perturbation methods, except for Prefix Padding, significantly reduce the MSE as the number of samples increases, eventually converging to a level comparable to or even better than standard sampling. In contrast, on the Traffic and Electricity datasets, the inference performance after introducing input perturbations did not improve to the same extent with increasing sampling budget. Although the MSE still showed a decreasing and gradually converging trend, its performance ceiling usually could not surpass or even approach that of standard sampling. This may be related to their use in pre-training. Further comparison of the two aggregation functions reveals that under the EM aggregation method, the MSE changes more smoothly with the number of samples, and in most settings, it achieves optimal performance superior to MV.

F.2 Moirai

Figure 13 shows that Moirai benefits clearly from increased inference budgets. On all three datasets, MSE always decreases as the number of samples grows, with a sharp drop at small budgets followed by gradual convergence, and the performance can be partly attributed to Moirai, which inherently adopts a multi-sample averaging strategy during inference, making additional samples directly contributive to performance improvement. The reduction is observed under both EM and MV aggregation, although EM typically yields smoother curves and lower final errors. Notably, similar to Table 1, *Gaussian Noise* and *Random Offset* perturbations are largely ineffective for the Moirai model. In comparative analyses across perturbation types, these two methods are more likely to result in higher MSEs than other perturbations and even display an atypical scaling behavior in which performance improves initially but fails to converge toward the curves of more effective perturbations. In contrast, *Task Dependency* and *Task Sensitivity*, perturbations achieve low errors across sampling budgets on Moirai. Overall, these results suggest that while increasing the sampling budget universally benefits inference,

the effectiveness of diversified sampling critically depends on the alignment between perturbation design and model-specific inductive biases.

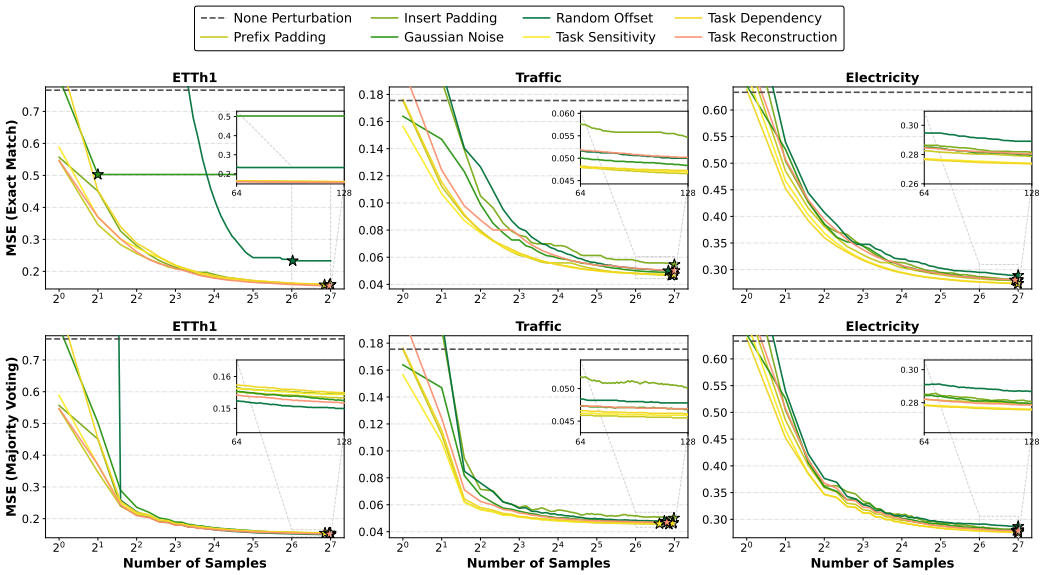


Figure 13: Performance of *Moirai* on different datasets with effective perturbations.

F.3 TIMEMOE

Figure 14 indicates that TimeMoE without decoding strategies exhibits weaker sensitivity to inference scaling. In this setting, *Noise Perturbations* and *Task Reconstruction perturbations* achieve good performance, benefiting from increased sampling budgets. In contrast, *Structural Perturbations* display little variation as the number of samples grows: their MSE remains higher than that of standard sampling and in some cases even degrades with additional samples. Moreover, due to the differences in aggregation methods, the optimal MSE achieved under MV aggregation in TimeMoE is often inferior to that obtained with EM aggregation. Overall, these results indicate that the effect of inference scaling in TimeMoE is highly dependent on the characteristics of the induced perturbations.

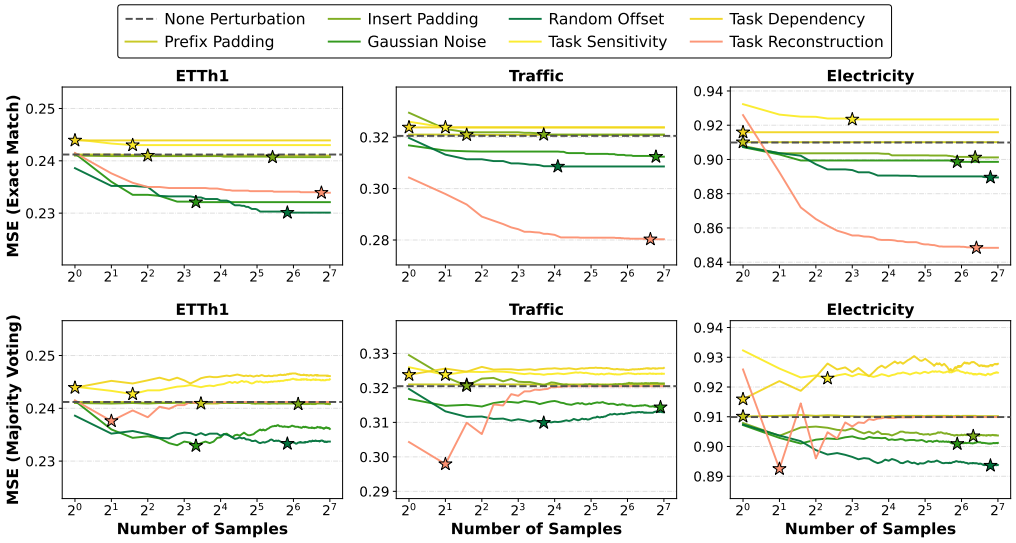


Figure 14: Performance of *TimeMoE* on different datasets with effective perturbations.

F.4 Extended Experiments on Additional Datasets

We further extend our evaluation to additional TSFMs, including Chronos, TimeMoE, and Moirai. Due to the structural differences between language and time series data, perturbations that are effective in LLM prompting do not necessarily transfer to time series forecasting. In particular, suffix perturbations append irrelevant signals beyond the observation window, leading the model to misinterpret noise as future trends and thus violating causal boundaries. Missing perturbations disrupt temporal continuity and reduce input fidelity, impairing the model's ability to capture local dependencies. As a result, these perturbations introduce harmful noise rather than meaningful diversity, and are unlikely to yield trends comparable to effective strategies. As shown in Table 9, *suffix* and *missing* perturbations underperform relative to standard sampling, often failing to produce reliable predictions. This observation further justifies our decision to exclude these perturbations from subsequent analysis, as their inclusion may introduce substantial bias during aggregation.

We additionally evaluate the effect of effective perturbation strategies under increasing sample sizes, as illustrated in Figure 15 and Figure 16. The results show that as the number of samples increases, the performance gains from effective perturbations become more pronounced, with MSE reductions improving by up to approximately 50%. Notably, performance tends to converge before reaching a moderate sample budget, indicating diminishing returns beyond this regime. For the Weather dataset, we adopt a univariate forecasting setting, where predictions are generated solely based on the historical target variable. In this case, the absence of auxiliary variables limits the model's ability to capture complex dependencies, which partially explains the relatively constrained performance compared to multivariate settings. Overall, the results on the Weather and Exchange Rate datasets are largely consistent with the observations derived from the initial four datasets, further supporting the generality of our findings.

Table 9: Performance of Diversified Strategy on exchange rate and weather.

Diversified Strategy	Chronos-T5-base		TimeMoE-200M		Moirai-1.1-R-base	
	Exchange Rate	Weather	Exchange Rate	Weather	Exchange Rate	Weather
Baseline	0.1136±0.0120	1.1212±0.0120	0.1618±0.0000	1.2421±0.0000	0.2017±0.0130	1.6417±0.0696
Suffix	**0.8680**±0.4153	**1.2680**±0.1319	**0.1951**±0.0246	1.2445±0.0124	**1.0654**±0.4142	**INF**
Missing	**0.2377**±0.1235	1.1116±0.0833	0.1616±0.0062	1.2150±0.0182	**0.5583**±0.0967	**INF**

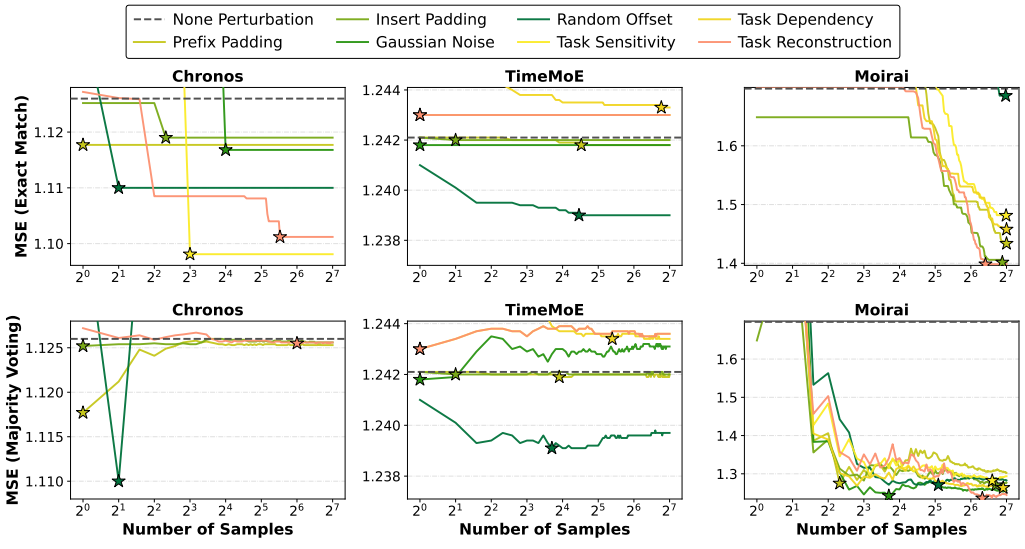


Figure 15: Performance of *Chronos, TimeMoE and Moirai* on *Weather*.

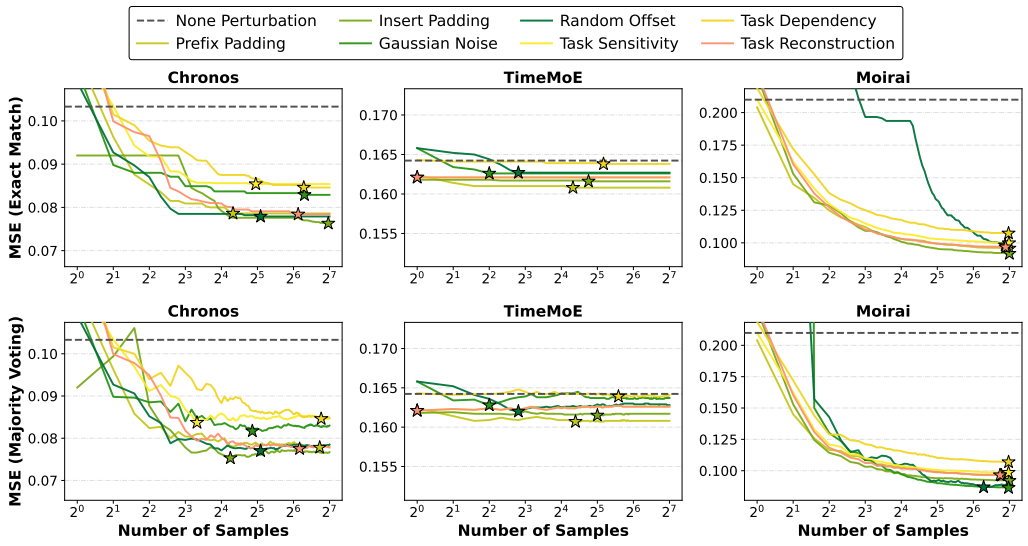


Figure 16: Performance of *Chronos, TimeMoE and Moirai* on *Exchange Rate*.

G Combination of Perturbation and Scaling Inference

We conduct a series of ablation studies on model sizes, context length, and temperature to better understand how inference scaling interacts with stochasticity and input capacity. Input diversity in these experiments is mainly introduced via Gaussian Noise perturbations. The results reveal how these factors jointly influence the diversity and quality of sampled candidates under varying aggregation functions.

G.1 Ablation of Model Sizes

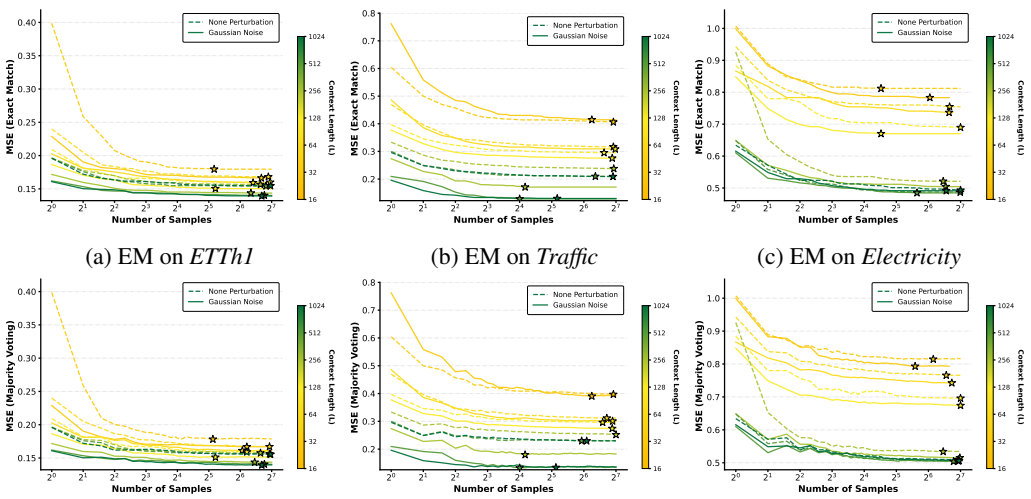
We evaluate diversified inference scaling across *Chronos-T5* models of varying sizes on ETTh1, as reported in Table 10. While larger models achieve lower absolute error, the relative gains from diversified sampling decrease with model size. In particular, strong models already attain low error without perturbations, leaving limited space for additional gains. Diversified sampling therefore benefits smaller and medium-sized models most, and should be applied more selectively to larger models where inference scaling returns saturate.

Table 10: Comparison of *Chronos-T5* across different size and perturbations under 128 samples.

Perturbation	Chronos-T5-Tiny		Chronos-T5-Mini		Chronos-T5-Small		Chronos-T5-Base		Chronos-T5-Large	
	EM	MV	EM	MV	EM	MV	EM	MV	EM	MV
None	0.1543	0.1564	0.1516	0.1523	**0.1418**	**0.1445**	0.1425	0.1453	0.1444	0.1481
Prefix	0.1548 (-0.3%)	0.1563 (+0.1%)	0.1532 (-1.1%)	0.1550 (-1.8%)	**0.1421** (-0.2%)	**0.1456** (-0.8%)	0.1432 (-0.5%)	0.1463 (-0.7%)	0.1439 (+0.3%)	0.1486 (-0.3%)
Gaussian	0.1402 (+9.1%)	0.1406 (+10.1%)	0.1527 (-0.7%)	0.1548 (-1.6%)	0.1414 (+0.3%)	0.1441 (+0.3%)	0.1414 (+0.8%)	0.1431 (+1.5%)	**0.1359** (+5.9%)	**0.1377** (+7.0%)
Dependency	0.1539 (+0.3%)	0.1539 (+1.6%)	0.1592 (-5.0%)	0.1617 (-6.2%)	**0.1402** (+1.1%)	**0.1432** (+0.9%)	0.1434 (-0.6%)	0.1467 (-1.0%)	0.1471 (-1.9%)	0.1502 (-1.4%)

G.2 Ablation of Context Length

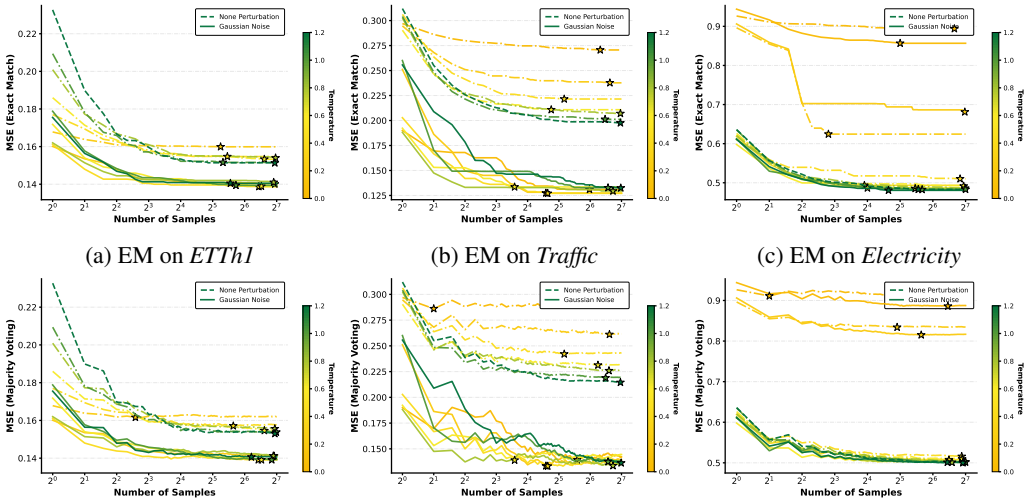
We vary the input context length while increasing the number of sampled predictions on *Chronos-T5*, as shown in Figure 17. Extending the context length basically improves inference performance across datasets and aggregation functions by lowering the achievable error floor. Without input perturbations, these gains quickly saturate. When input perturbations are introduced, longer contexts lead to more stable and sustained scaling as N increases. This effect is more pronounced under EM aggregation, while MV exhibits smoother but weaker improvements, indicating that longer contexts primarily enhance the quality of the best sampled candidates rather than the average prediction.



(a) EM on *ETTh1* (b) EM on *Traffic* (c) EM on *Electricity*

(d) MV on *ETTh1* (e) MV on *Traffic* (f) MV on *Electricity*

Figure 17: Relationship between context length and inference scaling performance on *Chronos-T5*.



(a) EM on *ETTh1* (b) EM on *Traffic* (c) EM on *Electricity*

(d) MV on *ETTh1* (e) MV on *Traffic* (f) MV on *Electricity*

Figure 18: Relationship between temperature and inference scaling performance on *Chronos-T5*.

G.3 Ablation of Temperature

As shown in Figure 18, we vary the sampling temperature on *Chronos-T5* while fixing the model and context length. Excessively high temperatures already improve performance as N increases but quickly saturate, while lower temperatures lead to unstable or degraded scaling. An intermediate temperature range yields stronger and more stable inference gains under aggregations. Overall, the results indicate that while temperature tuning alone can improve early scaling, sustained inference gains require controlled diversity beyond purely stochastic decoding.

H General Horizons

Beyond the most commonly adopted protocol in long-term forecasting, $H = 96$, we conduct some additional experiments on ETTh1, Traffic, and Electricity using Moirai and TimesFM to evaluate the performance of $H = \{36, 60, 96, 128, 192, 336, 720\}$ under the fixed number of samples $N = 64$. As Table 11 shows, diversified sampling provides additional forecasting headroom under both EM

Table 11: Performance of $H = \{36, 60, 96, 128, 192, 336, 720\}$ on Moirai and TimesFM.

Horizon	Dataset	Moirai (Task-Agnostic)	TimesFM (Task-Agnostic)	Moirai (Task-Specific)	TimesFM (Task-Specific)
36	ETTh1	0.0866 / 0.0889 / 0.0948	0.0849 / 0.0852 / 0.0916	0.0885 / 0.0879 / 0.0911	0.0872 / 0.0872 / 0.0878
	Traffic	0.0408 / 0.0430 / 0.0615	0.0395 / 0.0397 / 0.0448	0.0399 / 0.0412 / 0.0664	0.0402 / 0.0403 / 0.0407
	Electricity	0.1578 / 0.1587 / 0.1824	0.1451 / 0.1452 / 0.1536	0.1543 / 0.1547 / 0.1782	0.1470 / 0.1470 / 0.1473
	Average	0.0951 / 0.0969 / 0.1129	0.0898 / 0.0900 / 0.0967	0.0942 / 0.0946 / 0.1119	0.0915 / 0.0915 / 0.0919
60	ETTh1	0.1217 / 0.1155 / 0.1245	0.1127 / 0.1138 / 0.1205	0.1193 / 0.1206 / 0.1191	0.1177 / 0.1178 / 0.1183
	Traffic	0.0442 / 0.0413 / 0.0715	0.0401 / 0.0406 / 0.0457	0.0403 / 0.0414 / 0.0702	0.0409 / 0.0409 / 0.0413
	Electricity	0.1810 / 0.1795 / 0.2041	0.1650 / 0.1650 / 0.1752	0.1755 / 0.1761 / 0.1885	0.1672 / 0.1672 / 0.1674
	Average	0.1156 / 0.1121 / 0.1334	0.1059 / 0.1065 / 0.1138	0.1117 / 0.1127 / 0.1259	0.1086 / 0.1086 / 0.1090
96	ETTh1	0.1383 / 0.1407 / 0.1458	0.1334 / 0.1354 / 0.1435	0.1413 / 0.1417 / 0.1478	0.1401 / 0.1411 / 0.1435
	Traffic	0.0431 / 0.0438 / 0.0670	0.0437 / 0.0439 / 0.0481	0.0420 / 0.0422 / 0.0490	0.0441 / 0.0444 / 0.0500
	Electricity	0.1945 / 0.1953 / 0.2141	0.1845 / 0.1852 / 0.1947	0.1919 / 0.1920 / 0.2027	0.1881 / 0.1887 / 0.2054
	Average	0.1253 / 0.1266 / 0.1423	0.1205 / 0.1215 / 0.1288	0.1251 / 0.1253 / 0.1332	0.1241 / 0.1247 / 0.1330
128	ETTh1	0.1549 / 0.1629 / 0.7284	0.1446 / 0.1478 / 0.1564	0.1599 / 0.1664 / 0.6989	0.1537 / 0.1538 / 0.1543
	Traffic	0.0443 / 0.0470 / 0.1743	0.0445 / 0.0447 / 0.0493	0.0430 / 0.0439 / 0.0957	0.0451 / 0.0452 / 0.0456
	Electricity	0.2140 / 0.2120 / 0.6063	0.1985 / 0.1990 / 0.2082	0.2096 / 0.2102 / 0.4996	0.2023 / 0.2025 / 0.2027
	Average	0.1377 / 0.1406 / 0.5030	0.1292 / 0.1305 / 0.1380	0.1375 / 0.1402 / 0.4314	0.1337 / 0.1338 / 0.1342
192	ETTh1	0.1865 / 0.1869 / 0.9497	0.1594 / 0.1624 / 0.1721	0.1834 / 0.1890 / 0.9698	0.1695 / 0.1696 / 0.1703
	Traffic	0.0471 / 0.0473 / 0.1835	0.0459 / 0.0467 / 0.0512	0.0442 / 0.0447 / 0.0956	0.0464 / 0.0464 / 0.0468
	Electricity	0.2336 / 0.2335 / 0.6946	0.2170 / 0.2178 / 0.2269	0.2280 / 0.2289 / 0.5592	0.2211 / 0.2213 / 0.2216
	Average	0.1557 / 0.1559 / 0.6093	0.1408 / 0.1423 / 0.1501	0.1519 / 0.1542 / 0.5415	0.1457 / 0.1458 / 0.1462
336	ETTh1	0.2003 / 0.1964 / 2.0529	0.1767 / 0.1791 / 0.1926	0.1942 / 0.1914 / 1.7342	0.1863 / 0.1864 / 0.1873
	Traffic	0.0476 / 0.0447 / 0.2063	0.0460 / 0.0468 / 0.0519	0.0438 / 0.0434 / 0.1004	0.0474 / 0.0474 / 0.0477
	Electricity	0.2698 / 0.2677 / 0.8700	0.2460 / 0.2463 / 0.2555	0.2614 / 0.2624 / 0.7083	0.2468 / 0.2470 / 0.2492
	Average	0.1726 / 0.1696 / 1.0431	0.1562 / 0.1574 / 0.1667	0.1665 / 0.1657 / 0.8476	0.1602 / 0.1603 / 0.1614
720	ETTh1	0.3074 / 0.3049 / 6.0517	0.2400 / 0.2446 / 0.2622	0.3152 / 0.3100 / 9.0011	0.2474 / 0.2481 / 0.2501
	Traffic	0.0488 / 0.0466 / 0.3201	0.0525 / 0.0523 / 0.0574	0.0446 / 0.0436 / 0.1354	0.0527 / 0.0527 / 0.0529
	Electricity	0.3538 / 0.3504 / 1.4421	0.3162 / 0.3160 / 0.3274	0.3396 / 0.3429 / 1.0141	0.3184 / 0.3188 / 0.3191
	Average	0.2367 / 0.2340 / 2.6046	0.2029 / 0.2043 / 0.2157	0.2331 / 0.2322 / 3.3835	0.2062 / 0.2065 / 0.2074

and MV, while the magnitude of the benefit remains model-, dataset-, and horizon-dependent. The result supports that diversified scaling inference expands the predictive support of TSFMs under a fixed compute budget.

I Compute Latency

Scaling the number of samples (N) linearly increases total computation (FLOPs). However, in practical deployments, efficiency is primarily determined by latency rather than raw operation counts. Since our approach is inherently parallelizable, multiple samples can be generated simultaneously, meaning that increasing N does not lead to a proportional increase in inference time. The actual latency overhead depends on the perturbation design. Task-agnostic perturbations (e.g., prefix) with $\mathcal{O}(1)$ time complexity introduce negligible computational overhead, while task-specific perturbations may incur a small additional cost due to their structural characteristic, trading minor latency increases for improved predictive performance.

To quantify this trade-off, we conduct end-to-end latency benchmarks on the Chronos family. As shown in Table 12, diversified scaling inference achieves substantial performance gains (up to 25% error reduction) with only marginal latency overhead (typically 5%–10% for $N = 16$). These results demonstrate that inference scaling provides a compute-efficient alternative to model scaling, which often incurs significantly higher latency costs.

Table 12: Latency-performance trade-off under diversified inference on Chronos models. Increasing the number of samples (N) yields significant error reduction with marginal latency overhead.

Dataset	Model	Strategy	Latency (s)	MV
ETTh1	tiny	Standard (N=1)	326.66 s	0.1949
	tiny	Scaling (N=16)	353.82 s	0.1544
	tiny	Scaling (N=32)	397.32 s	0.1540
	base	Standard (N=1)	739.37 s	0.1767
	base	Scaling (N=16)	791.31 s	0.1427
	base	Scaling (N=32)	823.23 s	0.1424
Exchange Rate	tiny	Standard (N=1)	132.50 s	0.1177
	tiny	Scaling (N=16)	145.43 s	0.0944
	tiny	Scaling (N=32)	172.30 s	0.0932
	base	Standard (N=1)	309.11 s	0.1091
	base	Scaling (N=16)	325.42 s	0.0814
	base	Scaling (N=32)	360.14 s	0.0804

J Oracle Perturbation Selection via Prior-Guided Bandit

In practice, noise-based perturbations provide a robust default choice, as MV typically decreases with increasing N and converges to a stable regime (often around $N = 2^5$). But more importantly, the effectiveness of perturbations is inherently dataset-dependent. Our empirical analysis reveals consistent patterns across datasets. For smooth datasets (e.g., ETTh1), structure-preserving perturbations such as prefix tend to perform best, as aggressive perturbations may disrupt stable temporal patterns. For noisy or heterogeneous datasets (e.g., Traffic and Electricity), task-specific perturbations are more effective, as they introduce meaningful diversity without amplifying unstructured noise. Thus, to provide a practical and automated procedure for selecting perturbation strategies, we formalize the selection process as a two-stage framework that combines empirical priors with adaptive exploration.

Step 1: Rule-Based Search Space Before performing any automated exploration, we leverage empirical observations to prune the perturbation space. Specifically, we apply the following constraints: (i) *Hard constraints*: perturbations that violate temporal structure are excluded (e.g., Suffix perturbations break causality and Missing Data perturbations disrupt continuity), as identified in Table 1; (ii) *Magnitude control*: candidate perturbations are restricted to a high relevance regime (e.g., cosine similarity above a predefined threshold, Figure 3); (iii) *Soft priors*: perturbations are initialized with dataset-dependent preferences, where structure-preserving perturbations are prioritized for smooth datasets (e.g., ETTh1), while task-specific perturbations are favored for noisy or heterogeneous datasets (e.g., Traffic).

Step 2: Extreme-Value Bandit Exploration Given a fixed candidate set, we can adopt a multi-armed bandit for automatic perturbation selection. Crucially, as emphasized in our principle, the goal of diversified perturbation is not to improve the average quality of individual samples, but to expand the support set to capture highly accurate extremes (best-case candidates) for the aggregator. Therefore, inspired by the standard Upper Confidence Bound (UCB) algorithm (Auer et al., 2002) which balances exploration and exploitation, instead of optimizing the empirical mean, the algorithm minimizes a dynamically penalized proxy loss based on the EM (historical minimum MSE) by selecting a perturbation action a that minimizes the Lower Confidence Bound (LCB_a):

$$LCB_a = EM_{n_a} - c(t) \cdot \sqrt{\frac{\ln t}{n_a}}. \tag{17}$$

Here, t is the current total number of trials (budget spent, e.g., $t < 32$), n_a is the number of times perturbation action a has been selected, and $c(t)$ is an exploration coefficient. Furthermore, guided by Prop. 4.3 and Fig. 3, we couple the exploration constant $c(t)$ with the current budget t. For early steps, $c(t)$ is kept small to favor conservative, high-fidelity perturbations (exploitation). As the budget increases, $c(t)$ scales up to encourage exploring more diverse perturbations (exploration). Finally, providing a principled stopping criterion based on our findings, the automated allocation naturally terminates at a total inference budget of $N \approx 32$, beyond which the marginal gain diminishes. The

Table 13: Performance comparison of perturbation selection strategies across datasets and TSFMs.

Dataset	Model	None Perturbation	Heuristic (Noise EM_{32})	Prior-Guided MAB (UCB EM_{32})	Upper Bound (Best EM_{128})
Electricity	TimeMoE	0.9102	0.8999 (gaussian)	0.8533 (reconstruction)	0.8496 (reconstruction)
	Chronos	0.6481	0.4870 (gaussian)	0.4870 (gaussian)	0.4712 (reconstruction)
ETTh1	TimeMoE	0.2412	0.2321 (gaussian)	0.2321 (gaussian)	0.2300 (random)
	Chronos	0.1957	0.1416 (gaussian)	0.1416 (gaussian)	0.1401 (gaussian)
Traffic	TimeMoE	0.3202	0.3156 (gaussian)	0.2807 (reconstruction)	0.2801 (reconstruction)
	Chronos	0.3001	0.1284 (gaussian)	0.1284 (gaussian)	0.1284 (gaussian)
Exchange Rate	TimeMoE	0.1648	0.1637 (gaussian)	0.1602 (prefix)	0.1602 (prefix)
	Chronos	0.1033	0.0831 (gaussian)	0.0780 (random)	0.0775 (insert)

remaining budget is then entirely dedicated to the single winning action to efficiently expand the valid support set for robust inference scaling.

Empirical Validation We evaluate the proposed selection strategy across multiple datasets and TSFMs. As Table 13 below, we empirically validate our selector against the baseline and our general heuristic. We provide results under three different settings: A default heuristic strategy (Heuristic (Noise EM_{32})); the adaptive selection strategy (Prior-Guided MAB (UCB EM_{32})); and the headroom setting (Upper Bound (Best EM_{128})). The results demonstrate that our automated UCB procedure consistently identifies the optimal or near-optimal perturbation type within a limited budget, eliminating the need for manual configuration. These results demonstrate that automated perturbation selection is an effective and practical solution for deploying diversified inference in real-world scenarios.

K Limitation and Impact Statement

Limitations This work provides a systematic study of inference-time scaling and diversified sampling for time series foundation models. However, this work has several limitations. First, this work does not aim to identify a universally optimal perturbation that improves every TSFM on every dataset. A fixed default policy can be applied zero-shot, but model- or dataset-specific optimization may require additional validation or inference-time configuration. For users with additional computational resources or a stronger demand for optimal performance, selecting a better inference configuration can further improve results. Secondly, RobustMSE measures the performance headroom obtained under the evaluated inference-time sampling strategies, rather than an intrinsic, strategy-independent property of a TSFM. Lastly, although our experiments cover multiple representative TSFMs, datasets, perturbation strategies, and aggregation functions, it is still difficult to exhaustively cover all possible application domains and deployment scenarios in time series forecasting. Future work may further validate diversified scaling inference in more specialized domains, where temporal patterns and evaluation requirements can be highly domain-specific.

Another limitation concerns the computational trade-off at inference time. Our method is naturally parallelizable because candidate predictions generated from different perturbations can be computed independently. This makes diversified scaling inference efficient in parallel computing environments. However, compared with single-pass inference, it requires additional computational resources due to the generation of multiple candidates. In this work, we mitigate this issue by analyzing the relationship between inference budget and performance, and identifying moderate sample sizes where performance tends to stabilize. Nevertheless, selecting the most appropriate inference budget remains application-dependent and should consider accuracy gains and resource constraints.

Impact Statement Our work enhances forecasting reliability in critical domains such as energy, transportation, and healthcare, thereby improving resource allocation and operational decision-making. However, we acknowledge that incorrect predictions in sensitive applications can lead to suboptimal actions, and advanced forecasting on user-level data may introduce privacy risks. To mitigate these concerns, we strongly recommend deploying diversified scaling inference alongside domain-specific validation, uncertainty assessment, and human oversight. Furthermore, practitioners should transparently report their inference budgets, perturbation strategies, and computational costs when using diversified sampling or RobustMSE. Ultimately, this work is intended to support more reliable and compute-aware use of TSFMs, while recognizing that deployment in sensitive domains should be accompanied by careful validation and responsible data governance.