# Multi-Scale Finetuning for Encoder-based Time Series Foundation Models

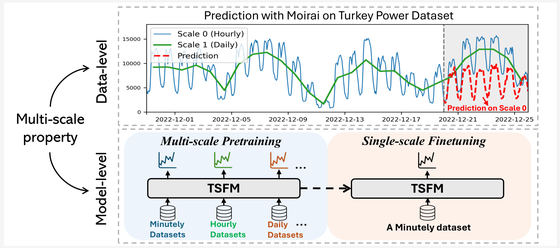
**Zhongzheng Qiao**[1,2,3], **Chenghao Liu**[4], **Yiming Zhang**[1], **Ming Jin**[5], **Quang Pham**[4], **Qingsong Wen**[6], **P.N.Suganthan**[7], **Xudong Jiang**[1], **Savitha Ramasamy**[2,3]

[1]NTU, [2]I2R A*STAR, [3]CNRS@CREATE, [4]Salesforce AI Research, [5]Griffith Univ, [6]Squirrel Ai Learning, [7]Qatar Univ

**NeurIPS 2025, San Diego, USA**

## 1. The Challenge: Naive Finetuning

**Time Series Foundation Models (TSFMs)** show impressive zero-shot performance. However, adapting them to downstream tasks remains challenging.

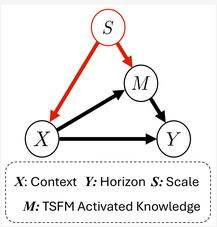



**Why Naive Finetuning Fails:**

- Direct finetuning (or linear probing) focuses only on the **original scale**.
- It ignores the inherent **multi-scale nature** of time series data and TSFMs.
- **Consequence:** Overfitting to patterns at the original scale and underutilizing the pre-trained capabilities.

## 2. A Causal Perspective

We model the finetuning process using a **Structural Causal Model (SCM)**:




- **Scale ($S$)** acts as a **confounder** affecting both Input ($X$) and Activated Knowledge ($M$).
- Naive methods model $P(Y|X)$, capturing spurious correlations.
- **Our Goal:** Estimate the interventional distribution $\mathbf{P(Y|do(X))}$ via backdoor adjustment.

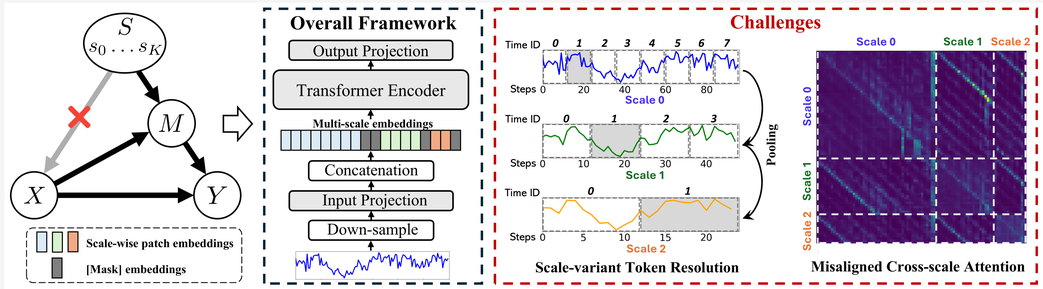
## Key Contributions

- **Novel Insight:** First work to introduce multi-scale modeling into TSFM finetuning from a causal view.
- **Method (MSFT):** A general framework compatible with encoder-based TSFMs (MOIRAI, MOMENT, UNITS).
- **SOTA Results:** Surpasses naive finetuning and deep learning baselines on Long Sequence & Probabilistic Forecasting.

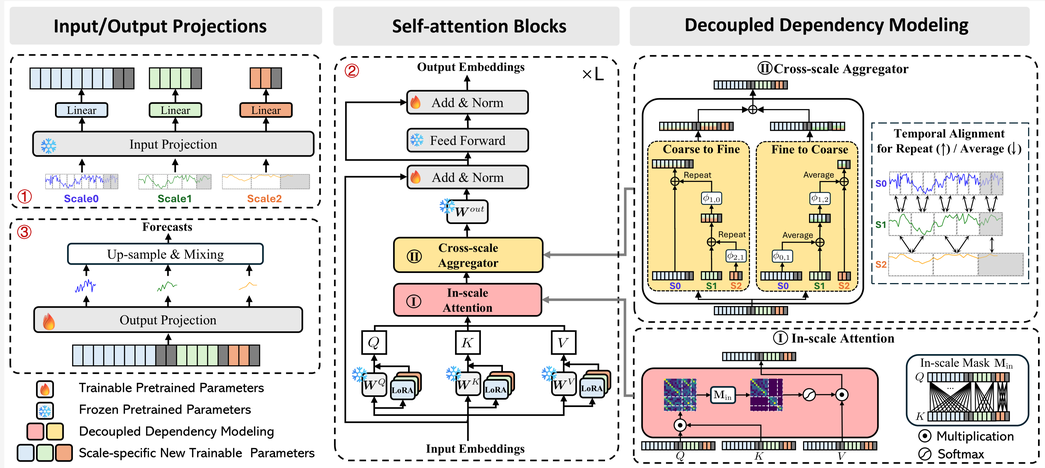
## 3. MSFT: Multi-Scale Finetuning Framework

We propose **MSFT**, a plug-and-play framework that explicitly models multiple scales to eliminate the confounding effect of scale.

**Intervention-based Framework & Scale-Induced Challenges**




**Full Design of MSFT:**




**Three Key Components:**

1. **Scale-Specific Knowledge Activation:**
   - Instead of a shared projection, we use **Scale-Specific Adapters** (Linear layers) for input projection.
   - Independent **LoRA** modules for each scale to adapt attention weights without forgetting pretrained knowledge.
2. **Decoupled Dependency Modeling (The Core):**
   - *Challenge:* Tokens at different scales have misaligned time indices. Direct attention causes spurious correlations.
   - *Solution:*
     - **In-Scale Attention:** Masking $M_{in}$ ensures tokens only attend to their own scale.
     - **Cross-Scale Aggregator:** Explicitly fuses corss-scale information via **Coarse-to-Fine** (Repeat) and **Fine-to-Coarse** (AvgPool) branches.
3. **Multi-Scale Mixing:**
   - Learnable weights $w_i$ combine predictions from all scales (Ensembling effect).

## 4. Long Sequence Forecasting

Comparison on ETT, Electricity, Weather datasets. MSFT consistently improves over other finetuning methods.

| Method | ETTm1 | | ETTm2 | | ETTh1 | | ETTh2 | | Electricity | | Weather | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| DLinear[2023] | 0.403 | 0.419 | 0.350 | 0.401 | 0.456 | 0.452 | 0.559 | 0.515 | 0.212 | 0.365 | 0.265 | 0.317 |
| PatchTST[2023] | 0.387 | 0.400 | 0.281 | 0.326 | 0.469 | 0.455 | 0.387 | 0.407 | 0.216 | 0.304 | 0.259 | 0.281 |
| iTransformer[2024a] | 0.407 | 0.410 | 0.288 | 0.332 | 0.454 | 0.448 | 0.383 | 0.407 | 0.178 | 0.270 | 0.258 | 0.278 |
| TimeMixer[2024] | 0.381 | 0.395 | 0.275 | 0.323 | 0.447 | 0.440 | 0.364 | 0.395 | 0.182 | 0.272 | 0.240 | 0.271 |
| SimpleTM [2025] | 0.381 | 0.396 | 0.275 | 0.322 | 0.422 | 0.428 | 0.353 | 0.391 | 0.166 | 0.260 | 0.243 | 0.271 |
| MOIRAI$_{\text{Small}}$ | 0.448 | 0.409 | 0.300 | 0.341 | 0.416 | 0.428 | 0.355 | 0.381 | 0.233 | 0.320 | 0.268 | 0.279 |
| + *Full finetuning* | 0.367 | 0.382 | 0.273 | 0.316 | 0.415 | 0.429 | 0.352 | 0.378 | 0.193 | 0.279 | 0.228 | 0.251 |
| + *Linear probing* | 0.388 | 0.392 | 0.295 | 0.337 | 0.414 | 0.427 | 0.354 | 0.380 | 0.212 | 0.299 | 0.237 | 0.260 |
| + *Prompt tuning* | 0.384 | 0.391 | 0.292 | 0.334 | 0.414 | 0.428 | 0.354 | 0.381 | 0.217 | 0.304 | 0.235 | 0.258 |
| + *LoRA* | 0.370 | 0.383 | 0.272 | 0.314 | 0.414 | 0.427 | 0.354 | 0.380 | 0.192 | 0.279 | 0.225 | 0.252 |
| + *AdaLoRA* | 0.381 | 0.386 | 0.273 | 0.319 | 0.414 | 0.427 | 0.354 | 0.380 | 0.191 | 0.279 | 0.226 | 0.252 |
| + ***MSFT*** | **0.353** | **0.377** | **0.250** | **0.301** | **0.412** | **0.426** | **0.349** | **0.375** | **0.187** | **0.275** | **0.216** | **0.248** |
| MOIRAI$_{\text{Base}}$ | 0.381 | 0.388 | 0.281 | 0.326 | 0.412 | 0.424 | 0.356 | 0.388 | 0.188 | 0.274 | 0.246 | 0.265 |
| + *Full finetuning* | 0.368 | 0.371 | 0.258 | 0.307 | 0.409 | 0.424 | 0.357 | 0.384 | 0.173 | 0.263 | 0.232 | 0.258 |
| + *Linear probing* | 0.388 | 0.387 | 0.277 | 0.319 | 0.409 | 0.424 | 0.356 | 0.387 | 0.182 | 0.269 | 0.229 | 0.253 |
| + *Prompt tuning* | 0.378 | 0.386 | 0.280 | 0.325 | 0.412 | 0.423 | 0.360 | 0.387 | 0.183 | 0.271 | 0.230 | 0.255 |
| + *LoRA* | 0.361 | 0.371 | 0.259 | 0.308 | 0.409 | 0.423 | 0.358 | 0.384 | 0.173 | 0.263 | 0.230 | 0.258 |
| + *AdaLoRA* | 0.359 | 0.371 | 0.258 | 0.307 | 0.410 | 0.423 | 0.356 | 0.384 | 0.173 | 0.264 | 0.236 | 0.260 |
| + ***MSFT*** | **0.332** | **0.369** | **0.247** | **0.305** | **0.407** | **0.422** | **0.352** | **0.383** | **0.169** | **0.260** | **0.213** | **0.245** |
| MOMENT | - | - | - | - | - | - | - | - | - | - | - | - |
| + *Full finetuning* | 0.352 | 0.380 | 0.260 | 0.320 | 0.425 | 0.440 | 0.347 | 0.394 | 0.224 | 0.311 | 0.336 | 0.310 |
| + *Linear probing* | 0.355 | 0.381 | 0.261 | 0.321 | 0.429 | 0.441 | 0.347 | 0.395 | 0.226 | 0.313 | 0.338 | 0.312 |
| + *Prompt tuning* | 0.356 | 0.381 | 0.261 | 0.320 | 0.427 | 0.440 | 0.348 | 0.395 | 0.226 | 0.312 | 0.336 | 0.310 |
| + *LoRA* | 0.356 | 0.381 | 0.260 | 0.320 | 0.425 | 0.439 | 0.347 | 0.395 | 0.225 | 0.312 | 0.335 | 0.309 |
| + *AdaLoRA* | 0.355 | 0.381 | 0.259 | 0.319 | 0.426 | 0.440 | 0.347 | 0.394 | 0.224 | 0.311 | 0.336 | 0.311 |
| + ***MSFT*** | **0.344** | **0.377** | **0.255** | **0.316** | **0.422** | **0.436** | **0.345** | **0.392** | **0.221** | **0.309** | **0.332** | **0.307** |
| UNITS | 0.713 | 0.553 | 0.321 | 0.355 | 0.527 | 0.491 | 0.406 | 0.418 | 0.432 | 0.488 | 0.291 | 0.313 |
| + *Full finetuning* | 0.395 | 0.405 | 0.297 | 0.338 | 0.442 | 0.435 | 0.386 | 0.409 | 0.190 | 0.283 | 0.257 | 0.283 |
| + *Linear probing* | 0.399 | 0.409 | 0.301 | 0.343 | 0.445 | 0.437 | 0.392 | 0.412 | 0.200 | 0.291 | 0.274 | 0.293 |
| + *Prompt tuning* | 0.431 | 0.430 | 0.299 | 0.341 | 0.438 | 0.433 | 0.386 | **0.405** | 0.191 | 0.287 | 0.247 | 0.276 |
| + *LoRA* | 0.393 | 0.405 | 0.296 | 0.338 | 0.437 | 0.434 | 0.384 | 0.407 | 0.188 | 0.282 | 0.250 | 0.279 |
| + ***MSFT*** | **0.390** | **0.403** | **0.288** | **0.334** | **0.434** | **0.430** | **0.380** | **0.405** | **0.184** | **0.279** | **0.242** | **0.273** |

## 5. Probabilistic Forecasting

MSFT is also effective for probabilistic tasks.

| Method | Electricity | | Solar | | Weather | | Istanbul Traffic | | Turkey Power | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CRPS | MSIS | CRPS | MSIS | CRPS | MSIS | CRPS | MSIS | CRPS | MSIS |
| DeepAR[2020] | 0.065 | 6.893 | 0.431 | 11.181 | 0.132 | 21.651 | 0.108 | 4.094 | 0.066 | 13.520 |
| TFT[2021] | 0.050 | 6.278 | 0.446 | 8.057 | 0.043 | 7.791 | 0.110 | 4.057 | 0.039 | 7.943 |
| PatchTST[2023] | 0.052 | 5.744 | 0.518 | 8.447 | 0.059 | 7.759 | 0.112 | 3.813 | 0.054 | 8.978 |
| TiDE[2023] | 0.048 | 5.672 | 0.420 | 13.754 | 0.054 | 8.095 | 0.110 | 4.752 | 0.046 | 8.579 |
| MOIRAI$_{\text{Small}}$ | 0.072 | 7.999 | 0.471 | 8.425 | 0.049 | 5.236 | 0.173 | 5.937 | 0.048 | 7.127 |
| + *Full finetuning* | 0.055 | 6.009 | 0.395 | 6.947 | 0.039 | 4.477 | 0.151 | 6.735 | 0.040 | 6.887 |
| + *Linear probing* | 0.062 | 6.438 | 0.369 | **5.865** | 0.049 | 4.785 | 0.154 | 4.645 | 0.047 | 6.912 |
| + *Prompt tuning* | 0.066 | 6.595 | 0.421 | 6.936 | 0.050 | 4.901 | 0.154 | 4.733 | 0.045 | 7.042 |
| + *LoRA* | 0.064 | 6.753 | 0.372 | 6.582 | 0.039 | 4.386 | 0.154 | 4.753 | 0.042 | 7.051 |
| + *AdaLoRA* | 0.064 | 6.892 | 0.366 | 8.015 | 0.040 | 4.496 | 0.152 | 4.670 | 0.041 | 7.127 |
| + ***MSFT*** | **0.047** | **5.327** | **0.353** | 7.706 | **0.036** | **4.178** | **0.141** | **4.447** | **0.038** | **6.810** |
| MOIRAI$_{\text{Base}}$ | 0.055 | 6.172 | 0.419 | 7.011 | 0.041 | 5.136 | 0.116 | 4.461 | 0.040 | 6.766 |
| + *Full finetuning* | 0.049 | 5.414 | 0.188 | 4.292 | 0.038 | 5.282 | 0.120 | 7.272 | 0.036 | 6.712 |
| + *Linear probing* | 0.055 | 5.951 | 0.379 | 5.645 | 0.039 | 4.544 | 0.104 | 3.736 | 0.042 | 7.259 |
| + *Prompt tuning* | 0.054 | 6.024 | 0.412 | 6.885 | 0.040 | 5.274 | 0.105 | 3.987 | 0.040 | 6.698 |
| + *LoRA* | 0.051 | 5.651 | 0.382 | 6.745 | 0.037 | 4.904 | 0.113 | 4.752 | 0.036 | 6.744 |
| + *AdaLoRA* | 0.054 | 5.937 | 0.383 | 8.825 | 0.038 | 4.802 | 0.110 | 3.895 | 0.037 | 6.762 |
| + ***MSFT*** | **0.046** | **5.199** | **0.142** | **3.464** | **0.035** | **4.603** | **0.098** | **3.685** | **0.034** | **6.419** |

## 6. Takeaway

**Scale Matters: Explicitly Model Scale for Robust Forecasting.**

By treating scale as a causal confounder and explicitly modeling it via MSFT, we unlock the true potential of Time Series Foundation Models.

**Get the Code:**

github.com/zqiao11/MSFT