# MM-TS: CHANNEL-STRUCTURED VISION-LANGUAGE MODELING FOR MULTIVARIATE TIME SERIES FORECASTING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multivariate time-series forecasting requires modeling temporal dynamics and evolving cross-channel dependencies. Recent LLM- and VLM-based forecasters adapt numerical sequences to pretrained foundation models through textualization, patch reprogramming, or visual reconstruction. However, adapting inputs to a pretrained backbone does not by itself provide explicit channel-relation guidance for temporal-token attention. We propose MM-TS, a channel-structured vision-language framework for multivariate time-series forecasting. MM-TS reuses global and local channel relations derived from each input window in two complementary ways: as visual tokens and as signed biases on temporal-token attention logits. Relations from the full-window magnitude spectrum form an image, while relations from successive frequency-axis patches form the video frames. A pretrained vision-language backbone jointly processes these visual tokens, time-series tokens, and textual prompts. To preserve channel-pair identity, local relations occupy diagonal patch blocks in the attention-bias matrix, whereas the global relation is broadcast to off-diagonal blocks. This dual use aligns each relation entry with the attention edges of its channel pair, enabling channel structure to guide information routing rather than remain an input-only representation. Evaluation covers eight standard benchmarks and ten additional TIME datasets spanning varied variables and sampling resolutions. Against nine baselines, MM-TS achieves the lowest horizon-averaged MSE on 14 of these 18 datasets. It also remains competitive with limited training data and achieves the lowest average MSE in five of six evaluated ETT transfer directions without target-domain weight updates. Controlled ablations support the complementary contributions of the two paths and the importance of channel-pair correspondence and input-dependent relations. Code is available at `https://anonymous.4open.science/r/MM-TS-DACB/`.

## 1 INTRODUCTION

Time series forecasting helps understand complex dynamical systems and supports anticipatory decision-making in traffic scheduling (Zhao et al., 2019), financial analysis (Sezer et al., 2020), and industrial monitoring (Xu et al., 2021; Deb et al., 2017). Forecasting models must capture local fluctuations and long-term trends across multivariate and high-dimensional series (Benidis et al., 2022; Zhang & Yan, 2023; Wang et al., 2024). This makes accurate cross-channel modeling central to multivariate forecasting (Zhou et al., 2021; Zhang et al., 2023).

LLM-based forecasters usually align numerical patches or summary statistics with language-model tokens through textualization, prompting, or reprogramming (Zhou et al., 2023; Jin et al., 2024). Multimodal methods further introduce visual representations to reduce the modality gap between continuous time series and discrete language and to preserve fine-grained temporal details through VLM pretraining (Shen et al., 2025; Zhong et al., 2025). However, input adaptation alone does not specify how channel-pair relations should guide interactions among temporal tokens inside the backbone. Figure 1 contrasts this input-adaptation view with the explicit representation and attention-level use of channel relations pursued in this work.



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.

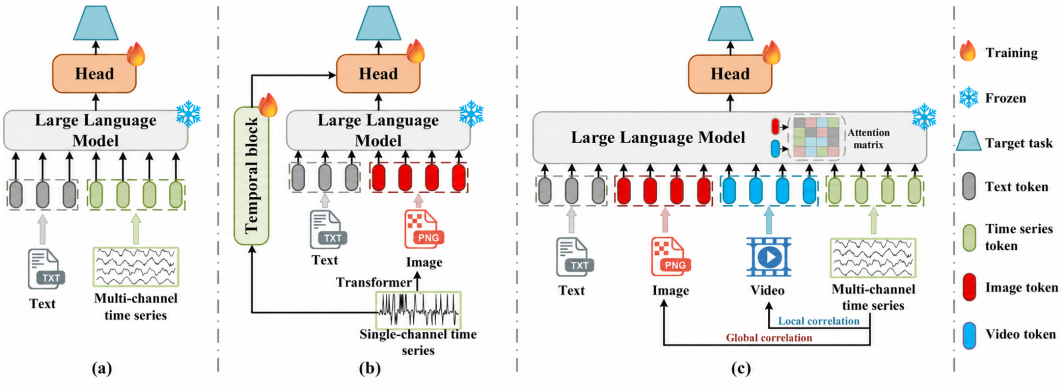



Figure 1: Comparison with existing foundation-model forecasting paradigms. (a) Text-to-time-series alignment adapts numerical sequences to language tokens. (b) Single-channel visual reconstruction treats images as auxiliary inputs. (c) MM-TS reuses channel-relation matrices to construct image/video tokens and structured attention biases that guide temporal-token interactions within the backbone.

**Explicit channel structure.** Multivariate forecasting benefits from representing which variables interact and how these interactions evolve (Ge et al., 2025; Chowdhury et al., 2026). Full-window and patch-level spectral relations provide complementary representations of channel pairs derived from the same observed window.

**Dynamic relations in attention.** Channel relations exhibit local and long-range variation (Lim & Zohren, 2021; Torres et al., 2021). Aligning each relation entry with its channel-pair attention edges allows current-window structure to guide interactions among temporal tokens inside the backbone.

To address these requirements, we propose **MM-TS**. Its central idea is to reuse the same channel relations for visual representation and attention guidance. Global and local spectral relations form image and video tokens, while their corresponding bias entries adjust attention between temporal tokens of each channel pair. This connects the structure presented to the visual encoder with the interactions used for forecasting. Our contributions are:

- We propose MM-TS, a channel-structured vision-language framework for multivariate time series forecasting that reuses time-series-derived relation matrices for both visual token construction and attention guidance. This connects explicit channel representations with information routing inside a pretrained backbone.
- We introduce a structured attention bias that preserves channel-pair identity: local spectral relations populate diagonal patch blocks, while the global relation is broadcast to off-diagonal blocks. This maps each relation entry to the attention edges of its channel pair, enabling channel structure to guide temporal-token interactions within the backbone.
- We evaluate MM-TS on eight standard benchmarks and ten additional TIME datasets, obtaining the lowest horizon-averaged MSE on 14 of 18 datasets against nine baselines. Component removal, channel permutation, and previous-window controls provide complementary evidence for the two structural paths, channel-pair correspondence, and input-dependent relations.

## 2 Related Work

Time-series foundation models such as TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), and Moirai (Woo et al., 2024) learn transferable forecasting representations from large time-series corpora. MM-TS focuses on reusing input-derived channel relations for visual representation and attention guidance within a pretrained multimodal backbone.

**Large Language Models for Time Series Forecasting.** LLM-based forecasting has progressed from numerical textualization to representation adaptation and context-enhanced prompt-

ing (Kong et al., 2025; Tang et al., 2025). Directly casting time series as text enables autoregressive prediction and zero-shot use, but suffers from quantization error, numerical instability, and recursive error accumulation (Gruver et al., 2023). Later methods align continuous signals with pretrained language representations: GPT4MTS (Jia et al., 2024) introduces prompt-based multimodal time-series forecasting by aligning numerical signals with language-model prompts, while Time-LLM freezes the LLM and uses patch reprogramming plus prompt prefixes containing dataset context, task instructions, and summary statistics to improve few-shot and zero-shot forecasting (Jin et al., 2024). These methods improve language compatibility, but cross-channel interactions are still implicitly absorbed by tokens or patches, limiting control over dynamic multivariate structure and motivating visual formulations with more explicit geometric and relational patterns.

**Visual and Multimodal Models for Time Series Forecasting.** Visual representations offer another route to time-series modeling. TimesNet shows that arranging temporal variation in two dimensions offers an inductive bias for periodicity and local structure (Wu et al., 2023). VisionTS recasts forecasting as image reconstruction with pretrained masked autoencoders, demonstrating transferable visual pretraining (Chen et al., 2025), while VisionTS++ extends this paradigm through continual pretraining and addresses modality, multivariate, and probabilistic gaps under cross-modal transfer (Shen et al., 2025). Time-VLM combines temporal features with visual and textual embeddings through cross-modal attention and gated fusion (Zhong et al., 2025). Graphormer (Ying et al., 2021) encodes graph structure as attention biases, while Multimodal Graph Transformer (He & Wang, 2023) incorporates visual and textual graph priors into multimodal attention. MM-TS reuses the same input-derived spectral channel relations as visual tokens and signed temporal-token attention biases inside a pretrained VLM. This design couples relation-based visual representation with channel-pair-specific guidance of temporal interactions.

# 3 METHOD

## 3.1 PROBLEM FORMULATION

Given a multivariate time series $X \in \mathbb{R}^{L\times C}$, we predict $\hat{Y} \in \mathbb{R}^{H\times C}$, where $L$, $H$, and $C$ denote the look-back length, forecasting horizon, and channel count (Kim et al., 2025). For batch size $B$, inputs and targets have shapes $B \times L \times C$ and $B \times H \times C$, respectively. Each channel is divided into patches of length $L_{\text{patch}}$ and spacing $\text{stride}$, giving $P = \lfloor (L - L_{\text{patch}})/\text{stride} \rfloor + 1$ patches. We use Qwen3-VL-2B (Bai et al., 2025) to jointly process the four token streams and denote its hidden dimension by $D$. MM-TS is trained by minimizing the MSE between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$ (Zeng et al., 2023).

## 3.2 FRAMEWORK OVERVIEW

MM-TS jointly processes temporal, text, image, and video tokens in Qwen3-VL's shared hidden space (Figure 2) (Bai et al., 2025). Cached spectral relations also guide temporal-token attention. Decoder states at temporal-token positions produce the forecast.

## 3.3 MULTIMODAL TOKEN CONSTRUCTION

**Time-Series Tokens.** A lightweight MLP maps the $P$ patches per channel to $\mathbf{Z}_{ts} \in \mathbb{R}^{B\times PC\times D}$ in Qwen3-VL's hidden space. In patch-major order, patch $p$ and channel $c$ have index $(p-1)C + c$ (both from 1), giving $P \times P$ attention blocks of size $C \times C$.

**Text Tokens.** The prompt template in Appendix A combines the forecasting task, dataset and sampling context, input length, horizon, channel count, and normalized-window statistics (minimum, maximum, mean, standard deviation, and trend). Qwen3-VL tokenizes it once with a 192-token budget, then embeds its $N_{\text{text}}$ tokens as $Z_{\text{text}} \in \mathbb{R}^{B\times N_{\text{text}}\times D}$.

**Image Tokens.** We apply an $L$-point FFT to each channel over the entire observed window and retain the magnitude at all $L$ frequency bins, $X_{\text{freq}} = |\mathcal{F}(X)|$. Pairwise DTW similarity, covariance, and Pearson correlation on $X_{\text{freq}}$ yield normalized matrices $\widetilde{R}_{\text{DTW}}$, $\widetilde{R}_{\text{Cov}}$, and $\widetilde{R}_{\text{Pearson}}$ in

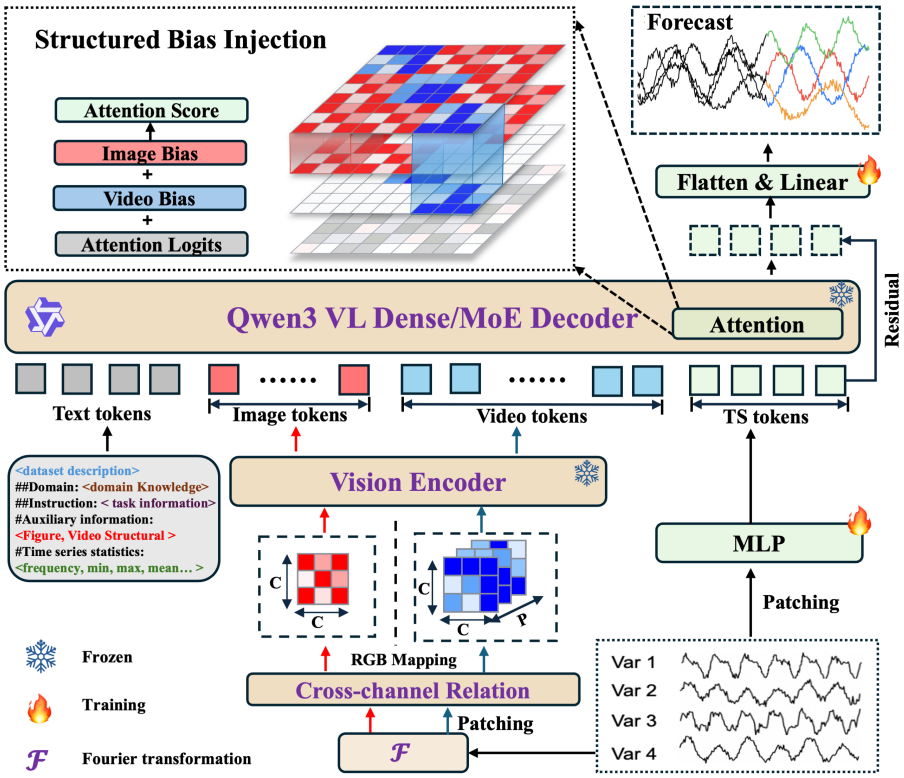



Figure 2: Overall architecture of MM-TS. Time-series, text, image, and video tokens are arranged as $[Z_{\text{text}}, Z_{\text{image}}, Z_{\text{video}}, Z_{\text{ts}}]$ and processed by Qwen3-VL, while cached image/video relation matrices are reused to form structured attention biases for temporal tokens.

$[0,1]^{B\times C\times C}$. We stack them as an RGB image:

$$S_{\text{image}} = \text{Stack}\Big(\tilde{R}_{\text{DTW}}, \tilde{R}_{\text{Cov}}, \tilde{R}_{\text{Pearson}}\Big) \in [0,1]^{B\times C\times C\times 3}, \tag{1}$$

where $\text{Stack}(\cdot)$ assigns DTW similarity, covariance, and Pearson correlation to the red, green, and blue channels, respectively. DTW aligns spectral magnitudes along frequency-bin indices to compare spectral shapes. These magnitude-based relations characterize spectral coupling without phase information, complementing temporal tokens that retain the ordered observations. The Qwen3-VL visual encoder produces $Z_{\text{image}} \in \mathbb{R}^{B\times N_{\text{image}}\times D}$. Normalization is detailed in Appendix A.

**Video Tokens.** We partition the full-window magnitude spectrum along its frequency-bin axis using the same patch length and stride as the temporal branch, yielding $P$ spectral patches. Each frame summarizes channel relations within one spectral patch. For patch $i$, the same three relations give $\widetilde{R}^i_{\text{DTW}}$, $\widetilde{R}^i_{\text{Cov}}$, and $\widetilde{R}^i_{\text{Pearson}}$ in $[0,1]^{B\times C\times C}$, forming one RGB frame:

$$S^i_{\text{video}} = \text{Stack}\Big(\widetilde{R}^i_{\text{DTW}}, \widetilde{R}^i_{\text{Cov}}, \widetilde{R}^i_{\text{Pearson}}\Big) \in [0,1]^{B\times C\times C\times 3}, \tag{2}$$

$$S_{\text{video}} = \{S^1_{\text{video}}, \ldots, S^P_{\text{video}}\}. \tag{3}$$

The visual encoder maps these $P$ frames to $Z_{\text{video}} \in \mathbb{R}^{B\times N_{\text{video}}\times D}$. We concatenate the tokens as $[Z_{\text{text}}, Z_{\text{image}}, Z_{\text{video}}, Z_{\text{ts}}]$ for Qwen3-VL decoding, placing task context and visual relations before temporal tokens. Section 3.4 reuses the cached RGB matrices as attention biases.

### 3.4 Structured Attention Bias Injection

The same cached relation matrices feed the VLM through two paths: the visual encoder produces image and video tokens, while the bias path maps relation entries to attention edges between tem-

poral tokens of the corresponding channel pairs. Decoder-layer indices are zero-based throughout. With the patch-major ordering defined above, diagonal $C \times C$ blocks connect channels within a temporal patch, and off-diagonal blocks connect different temporal patches.

Let $Q_{\text{all}}^{(\ell)}$, $K_{\text{all}}^{(\ell)}$, and $V_{\text{all}}^{(\ell)}$ be the query, key, and value projections at layer $\ell$ for the $N = N_{\text{text}} + N_{\text{image}} + N_{\text{video}} + PC$ multimodal tokens. Attention uses the full sequence, but structural bias applies only to the temporal-token submatrix. The scaled logits are

$$A_{\text{all}}^{(\ell)} = \frac{Q_{\text{all}}^{(\ell)} K_{\text{all}}^{(\ell)\top}}{\sqrt{d_k}}, \quad A_{\text{all}}^{(\ell)} \in \mathbb{R}^{B\times N\times N}, \quad \ell \in \{0, 1\}, \tag{4}$$

where $d_k$ is the key dimension. For a cached tensor $S \in [0, 1]^{B\times C\times C\times 3}$, let $S_{\text{dtw}}$, $S_{\text{cov}}$, and $S_{\text{pear}}$ denote its three channels in $[0, 1]^{B\times C\times C}$. We define the signed relation bias as

$$\mathcal{B}(S) = w_{\text{dtw}}(2S_{\text{dtw}} - 1) + w_{\text{cov}}(2S_{\text{cov}} - 1) + w_{\text{pear}}(2S_{\text{pear}} - 1), \tag{5}$$

where $w_{\text{dtw}}$, $w_{\text{cov}}$, and $w_{\text{pear}}$ are scalar weights for the relation channels. The mapping $2x - 1$ converts $[0, 1]$ to $[-1, 1]$; positive and negative biases increase and decrease the corresponding logits. For sample $b$, we place the relation bias of spectral patch $p$ in the $p$-th diagonal temporal-token block. This fixed index mapping preserves channel-pair correspondence within each block and specifies where spectral information enters temporal-token attention:

$$\text{Bias}_{\text{video}}^{(b)} = \text{blkdiag}\Big(\mathcal{B}(S_{\text{video}}^{1,(b)}), \ldots, \mathcal{B}(S_{\text{video}}^{P,(b)})\Big) \in \mathbb{R}^{PC\times PC}. \tag{6}$$

Stacking all batch elements gives $\text{Bias}_{\text{video}} \in \mathbb{R}^{B\times PC\times PC}$. The global image bias is broadcast to off-diagonal patch blocks. For the $b$-th sample, we define

$$\text{Bias}_{\text{image}}^{(b)} = (\mathbf{1}_{P\times P} - I_P) \otimes \mathcal{B}(S_{\text{image}}^{(b)}) \in \mathbb{R}^{PC\times PC}, \tag{7}$$

where $\mathbf{1}_{P\times P}$ is the all-one matrix, $I_P$ is the $P$-dimensional identity matrix, and $\otimes$ denotes the Kronecker product. Stacking over the batch dimension gives $\text{Bias}_{\text{image}} \in \mathbb{R}^{B\times PC\times PC}$. We embed $\text{Bias}_{\text{video}} + \text{Bias}_{\text{image}}$ in the temporal-token submatrix of $B_{\text{full}} \in \mathbb{R}^{B\times N\times N}$ and set all other entries to zero. The modified temporal-token attention is then computed as

$$\text{Attn}_{\text{ts}}^{(\ell)} = \Big[\text{Softmax}\Big(A_{\text{all}}^{(\ell)} + M + \gamma B_{\text{full}}\Big) V_{\text{all}}^{(\ell)}\Big]_{\text{ts}}, \quad \ell \in \{0, 1\}, \tag{8}$$

where $\gamma > 0$ is a learnable global bias scale and $[\cdot]_{\text{ts}}$ selects temporal-token rows. The mask $M$ blocks padding, keeps prefix queries causal, and lets temporal queries attend to all valid prefix and temporal tokens. It applies in every decoder layer; all temporal tokens come from the observed window. We extract the output states $Z_{ts}^{out} \in \mathbb{R}^{B\times PC\times D}$, add the input temporal tokens through a residual connection (He et al., 2016), and project this stream to the $H$-step forecast:

$$\hat{Y} = \text{Head}(Z_{ts}^{out} + Z_{ts}), \quad \hat{Y} \in \mathbb{R}^{B\times H\times C}. \tag{9}$$

## 4 Experiments

**Datasets and Metrics.** We evaluate MM-TS on eight standard benchmarks (ETTh1, ETTh2, ETTm1, ETTm2, Weather, ILI, Exchange, and Electricity) and ten additional TIME datasets (Qiao et al., 2026). The standard suite provides comparability with established forecasting methods, while TIME broadens coverage of variable combinations and temporal resolutions, assessing the same relation-construction framework under more heterogeneous forecasting conditions. For supervised forecasting, MM-TS and all nine baselines are evaluated with seeds 2022–2026 on each dataset and horizon. Reported MSE and MAE are averaged across the five runs; dataset-level scores average the evaluated horizons. Appendix B provides dataset characteristics and preprocessing; Appendix C reports complete results.

**Baselines.** The nine baselines comprise methods based on pretrained models (Time-VLM (Zhong et al., 2025), VisionTS (VTS) (Chen et al., 2025), Time-LLM (Jin et al., 2024), and TimesFM (TFM) (Das et al., 2024)) and task-specific forecasting models (iTransformer

(iTr) (Liu et al., 2024), PatchTST (PTST) (Nie et al., 2023), TimesNet (TNet) (Wu et al., 2023), TSMixer (TMix) (Ekambaram et al., 2023), and DLinear (DLin) (Zeng et al., 2023)). Within each evaluation setting, methods use the same dataset splits and forecast horizons. Details of the TimesFM adaptation are provided in Appendix B.3.

**Forecasting Protocol.** Standard benchmarks use $L = 96$ and $H \in \{96, 192, 336, 720\}$, except ILI, which uses $L = 104$ and $H \in \{24, 36, 48, 60\}$. TIME datasets use $L = 96$ and their task-specific horizons. Using a common look-back length controls the observed context within each task, so the comparisons characterize performance under matched input lengths. Few-shot and zero-shot protocols are specified below.

**Implementation Details.** MM-TS uses Qwen3-VL-2B with frozen base weights and trains its task-specific modules and selected LoRA adapters. Temporal patches have length 16 and stride 8. We optimize with AdamW, an effective batch size of 32, at most 20 epochs, and early-stopping patience 3, selecting checkpoints by validation MSE. Adapter placement and TIME-specific hyperparameters are summarized in Appendix B.

Lower MSE and MAE are better. Bold and underlined scores mark the best and second-best displayed values, respectively.

### 4.1 Supervised Forecasting

Table 1: Supervised forecasting: horizon-averaged MSE on 18 datasets.

| Dataset | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|
| *(a) Standard benchmarks* | | | | | | | | | | |
| ETTh1 | **0.434** | 0.440 | 0.465 | 0.437 | 0.464 | 0.454 | 0.470 | 0.459 | 0.441 | 0.456 |
| ETTh2 | **0.368** | 0.382 | 0.387 | 0.386 | 0.397 | 0.384 | 0.386 | 0.414 | 0.383 | 0.559 |
| ETTm1 | **0.383** | 0.390 | 0.391 | 0.398 | 0.388 | 0.410 | 0.388 | 0.399 | 0.389 | 0.403 |
| ETTm2 | **0.276** | 0.278 | 0.292 | 0.298 | 0.299 | 0.288 | 0.281 | 0.291 | 0.286 | 0.350 |
| Weather | **0.240** | 0.262 | 0.275 | 0.249 | 0.257 | 0.258 | 0.259 | 0.259 | 0.244 | 0.265 |
| ILI | 1.560 | 2.065 | 2.048 | 1.752 | 1.652 | 2.555 | **1.514** | 2.159 | 4.410 | 2.346 |
| Exchange | 0.351 | 0.370 | 0.369 | 0.358 | 0.357 | 0.378 | 0.371 | 0.464 | 0.382 | **0.336** |
| Electricity | **0.190** | 0.208 | 0.237 | 0.192 | 0.201 | 0.192 | 0.216 | 0.196 | 0.194 | 0.212 |
| *(b) TIME datasets* | | | | | | | | | | |
| SG PM2.5 | **1.023** | 1.105 | 1.133 | 1.102 | 1.100 | 1.078 | 1.108 | 1.101 | 1.029 | 1.044 |
| NEM | **0.487** | 0.608 | 0.698 | 0.637 | 0.607 | 0.572 | 0.628 | 0.516 | 0.499 | 0.695 |
| MetroPT-3 | **0.690** | 0.794 | 0.838 | 0.752 | 0.734 | 0.802 | 0.802 | 0.780 | 0.692 | 0.719 |
| NE Wind | **0.366** | 0.424 | 0.533 | 0.456 | 0.435 | 0.433 | 0.425 | 0.470 | 0.373 | 0.427 |
| EWELD | **0.290** | 0.316 | 0.414 | 0.319 | 0.295 | 0.296 | 0.322 | 0.305 | 0.327 | 0.327 |
| Oil Price | 0.065 | **0.064** | 0.128 | 0.070 | 0.071 | 0.077 | 0.070 | 0.133 | 0.091 | 0.066 |
| AU Solar | **0.248** | 0.268 | 0.449 | 0.254 | 0.249 | 0.255 | 0.304 | 0.270 | 0.259 | 0.289 |
| Coastal T-S | **0.404** | 0.501 | 0.933 | 0.656 | 0.479 | 0.455 | 0.525 | 0.517 | 0.441 | 0.467 |
| Crypto | 0.096 | **0.093** | 0.352 | 0.166 | 0.107 | 0.119 | 0.117 | 0.158 | 0.443 | 0.099 |
| Supply Chain | **0.617** | 0.731 | 1.733 | 1.404 | 0.625 | 0.684 | 0.869 | 1.122 | 0.745 | 0.672 |

**Settings.** Table 1 reports horizon-averaged MSE under the supervised protocol above. Models are trained separately for each TIME dataset and horizon, using identical task-specific horizons and evaluation masks across methods.

**Results.** MM-TS leads all four ETT datasets, Weather, and Electricity in average MSE, and ranks second on ILI (1.560) and Exchange (0.351). It also leads eight of the ten TIME tasks.

**Analysis.** MM-TS leads six of eight standard benchmarks and eight of ten TIME datasets in horizon-averaged MSE, extending its gains across different variable sets and sampling resolutions. The improvement over the strongest baseline reaches 3.7% on ETTh2 and 8.4% on Coastal T-S. The fixed-backbone ablations help explain these gains: removing visual tokens or attention bias raises ETTh2 MSE from 0.368 to 0.378 or 0.381, respectively, supporting complementary benefits from

representing channel relations and using them to guide token interactions. MM-TS also leads 14 datasets in average MAE. Section 4.3 and Appendix C provide the component controls and results by horizon, metric, and run.

### 4.2 Few-shot and Zero-shot Forecasting

Table 2: Few-shot and zero-shot forecasting: horizon-averaged MSE.

| Dataset / transfer | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|
| *(a) Few-shot forecasting: 5% training data* | | | | | | | | | | |
| ETTh1 | **0.453** | 0.459 | 0.879 | 0.464 | 0.476 | 0.540 | 0.702 | 0.837 | 0.560 | 0.462 |
| ETTm1 | **0.400** | 0.407 | 0.597 | 0.437 | 0.419 | 0.488 | 0.406 | 0.480 | 0.576 | 0.494 |
| Weather | **0.253** | 0.270 | 0.278 | 0.271 | 0.262 | 0.277 | 0.273 | 0.292 | 0.270 | 0.296 |
| Electricity | 0.219 | 0.250 | 0.239 | 0.223 | **0.202** | 0.223 | 0.234 | 0.303 | 0.257 | 0.250 |
| *(b) Zero-shot forecasting: direct ETT transfer* | | | | | | | | | | |
| ETTh1→ETTh2 | **0.371** | 0.377 | 0.385 | 0.383 | 0.384 | 0.382 | 0.377 | 0.424 | 0.379 | 0.509 |
| ETTh1→ETTm2 | **0.307** | 0.317 | 0.317 | 0.340 | 0.316 | 0.330 | 0.312 | 0.339 | 0.312 | 0.445 |
| ETTh2→ETTm2 | **0.314** | 0.322 | 0.321 | 0.336 | 0.339 | 0.331 | 0.321 | 0.366 | 0.321 | 0.434 |
| ETTm1→ETTh2 | 0.596 | 0.429 | 0.441 | **0.417** | 0.442 | 0.445 | 0.436 | 0.548 | 0.432 | 0.473 |
| ETTm1→ETTm2 | **0.288** | 0.293 | 0.301 | 0.303 | 0.300 | 0.302 | 0.295 | 0.356 | 0.295 | 0.344 |
| ETTm2→ETTh2 | **0.401** | 0.410 | 0.417 | 0.419 | 0.425 | 0.461 | 0.414 | 0.513 | 0.423 | 0.441 |

**Settings.** Few-shot forecasting uses 5% or 10% of the original training split, with validation and test splits unchanged. Zero-shot forecasting means direct ETT transfer: models trained on a source dataset predict a target dataset without target-domain weight updates. Each comparison averages its available common horizons.

**Results.** MM-TS leads three of the four 5% comparisons in Table 2(a). Across all eight datasets, it leads six at 5% and five at 10%. It also leads five of six direct-transfer directions, with ETTh1-to-ETTm2 MSE of 0.307 versus 0.312 for the strongest baselines.

**Analysis.** MM-TS computes channel relations from each observed window, so this structural information remains available when labeled training data are limited. It leads six datasets at the 5% training fraction. On the shared ETTh1 horizons of 96, 192, and 336 steps, increasing training data to 10% reduces MSE from 0.407, 0.456, and 0.497 to 0.397, 0.441, and 0.484, respectively. During direct ETT transfer, model weights stay fixed, but relation matrices are recomputed from each target window. This lets attention use channel associations in the target observations. Both MSE and MAE improve for ETTh1-to-ETTh2 and ETTh1-to-ETTm2, covering the evaluated same-resolution and cross-resolution transfers. Appendices D.1 and D.2 report all training fractions, matched horizons, and transfer directions.

### 4.3 Ablation Study

Table 3: Ablation study: mean MSE across four prediction lengths.

| Setting | ETTh1 | ETTh2 | Weather | Electricity |
|---|---|---|---|---|
| Full | **0.434** | **0.368** | **0.240** | **0.190** |
| Without visual tokens | 0.443 | 0.378 | 0.249 | 0.195 |
| Without text | 0.439 | 0.373 | 0.245 | 0.193 |
| Without attention bias | 0.446 | 0.381 | 0.250 | 0.199 |
| Without visual tokens and attention bias | 0.455 | 0.389 | 0.258 | 0.204 |
| Permuted bias | 0.451 | 0.385 | 0.253 | 0.201 |
| Previous-window bias | 0.445 | 0.379 | 0.249 | 0.197 |

**Settings.** We remove individual input or attention components and test channel-permuted and previous-window biases on four datasets, averaging MSE over four horizons.

**Results.** The full model leads all four comparisons in Table 3. On ETTh2, removing visual tokens, attention bias, or both raises MSE from 0.368 to 0.378, 0.381, and 0.389, respectively. Joint removal gives the largest degradation on every dataset.

**Analysis.** The ablations test three distinct requirements of structural reuse. Removing either path increases error, supporting their complementary contributions. Channel permutation performs worse than omitting the bias on all four datasets, showing that the assignment of relations to channel pairs matters. Previous-window bias also underperforms current-window bias, supporting recomputation from the input being forecast. Appendix E.1 reports the corresponding horizon-level and MAE comparisons.

### 4.4 Model Analysis

Table 4: Backbone and visual-pretraining comparisons: mean MSE at $H = 96$ and 720.

| Setting | ETTh1 | ETTm1 | Weather |
|---|---|---|---|
| Qwen3-VL-2B (Full) | **0.424** | **0.388** | **0.250** |
| 2B, reset VL projector | 0.438 | 0.401 | 0.260 |
| Qwen3-VL-4B | 0.426 | 0.390 | 0.252 |
| DINOv2 + decoder | 0.440 | 0.403 | 0.262 |
| iTransformer | 0.445 | 0.412 | 0.267 |
| iTransformer + relation bias | 0.435 | 0.402 | 0.259 |

**Backbone Comparison. Settings.** We compare backbones, reset the 2B visual-language projector, and add global relation bias to iTransformer, averaging MSE at $H = 96$ and 720.

**Results.** Qwen3-VL-2B leads all three dataset comparisons in Table 4. Resetting its projector raises ETTh1 MSE from 0.424 to 0.438; the 4B backbone does not improve these averages.

**Analysis.** Resetting the visual-language projector increases average MSE on all three datasets, supporting the retained pretrained interface for encoding relation images. Adding global relation bias to iTransformer improves its average MSE from 0.445 to 0.435 on ETTh1, 0.412 to 0.402 on ETTm1, and 0.267 to 0.259 on Weather. These comparisons support the usefulness of both the pretrained interface and structural attention. Component effects are assessed within a fixed backbone; comparisons between backbone families additionally reflect capacity and pretraining differences. The 4B model provides no accuracy advantage under the evaluated settings and requires greater memory and latency. Appendices E.2 and F provide the detailed accuracy and cost evidence.

**Design Choices.** The design controls support complementary relation descriptions and early structural guidance. Combining DTW, covariance, and Pearson relations gives lower error than the individual metrics, favoring multiple descriptions of spectral channel structure. Bias injection in layers 0 and 1 outperforms later placement, supporting the introduction of channel guidance before subsequent decoder transformations. The frame-count and joint patching comparisons favor multiple local spectral views with the selected temporal and spectral granularity. The tested weight-ratio perturbation changes average MSE by at most 0.004 across the three tested datasets, indicating limited sensitivity to this perturbation. Appendices E.3–E.7 connect each observation to the corresponding design choice.

**Computational Cost.** On an RTX 4090 at $L = H = 96$, MM-TS uses 6.672 million trainable parameters and 8.135 GB peak reserved training memory on ETTh1. Inference takes 25.825, 24.802, and 34.889 ms per sample on ETTh1, ETTm1, and Weather, respectively (Tables 21–23). Measurements use training batch size 4 and inference batch size 1. Training time excludes offline visual caching; latency includes input processing. Appendix F relates the component costs to the accuracy gains and distinguishes parameter count from runtime.

### 4.5 CASE STUDY

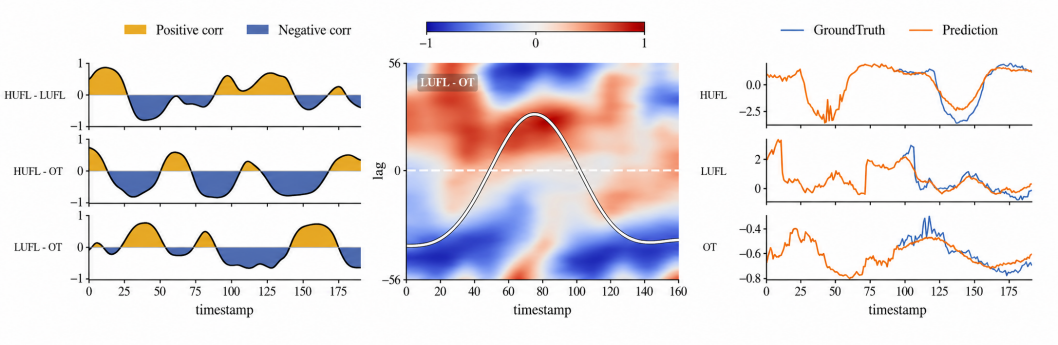



Figure 3: ETTm1 case study. Left: rolling correlations (yellow/blue: positive/negative). Middle: LUFL–OT lag correlations (red/blue: positive/negative); the white curve marks the dominant lag. Right: observations (blue) and predictions (orange).

**Settings.** Figure 3 examines ETTm1 through rolling correlations, LUFL–OT lag structure, and forecasts. HUFL and LUFL denote high and low useful load; OT denotes transformer oil temperature (Zhou, n.d.).

**Results.** HUFL–LUFL and LUFL–OT repeatedly switch between positive and negative local correlation, while HUFL–OT has extended negative intervals. The LUFL–OT lag ridge moves across zero, indicating that their temporal association changes within the window. MM-TS follows the HUFL valley and recovery, the oscillatory LUFL transition, and the broader rise and decline of OT.

**Analysis.** The load variables fluctuate rapidly, while oil temperature changes more smoothly. This difference is consistent with heat accumulation and dissipation in the transformer (Taheri et al., 2020). The correlation sign changes and moving lag ridge in Figure 3 illustrate why one simultaneous correlation summary is insufficient to describe the window. The forecast tracks the broad load recovery and the slower rise and decline of oil temperature, reflecting the distinct response scales of electrical loading and thermal accumulation. Temporal tokens retain the ordered trajectories, while spectral relation inputs summarize how channels co-vary across frequency ranges. This division of roles connects the case study to the model's joint use of temporal observations and channel structure. The ablations in Section 4.3 assess the two structural paths; Appendices G.1 and G.2 provide adjacent-window statistics and further examples.

## 5 CONCLUSION

We presented MM-TS, a channel-structured vision-language framework that reuses global and local spectral channel relations as visual tokens and signed attention biases. Evaluation on eight standard benchmarks and ten additional TIME datasets supports its applicability across varied forecasting tasks, with the lowest average MSE on 14 of 18 datasets. The expanded TIME evaluation broadens the evidence beyond the established benchmark suite, while component and structural controls connect the gains to complementary visual representation, channel-pair correspondence, and input-conditioned attention guidance. MM-TS also leads six of eight datasets at the 5% training fraction and five of six evaluated ETT transfer directions in average MSE. Together with the structural controls, these results support reusing input-derived channel relations for both multimodal representation and attention guidance across the evaluated forecasting settings.

**Limitations and Future Work.** The current pairwise relation construction incurs increasing cost as the channel count grows, making sparse and low-rank representations promising extensions. Direct transfer across ETT datasets provides evidence under a shared variable layout; extending this capability to heterogeneous channel sets motivates learning flexible channel correspondence. These directions would broaden the scale and transfer scope of the framework.

### AI USE STATEMENT

Generative AI tools assisted with searching for and reading related literature, and manuscript editing. The authors reviewed the AI-assisted work and take responsibility for the final paper.

### REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe the model and input construction. Section 4 and Appendix B specify the datasets, preprocessing, evaluation protocols, and training configuration. Appendices C–E report detailed results, run-to-run variation, and ablations, while Appendix F documents computational costs. Anonymous code repository: `https://anonymous.4open.science/r/MM-TS-DACB/`.

## REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, 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 Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=gerNCVqqtR`.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, et al. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36, 2022.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. VisionTS: Visual masked autoencoders are free-lunch zero-shot time series forecasters. In *International Conference on Machine Learning*, pp. 8979–9007. PMLR, 2025.

Abdul Monaf Chowdhury, Rabeya Akter, and Safaeid Hossain Arib. T3time: Tri-modal time series forecasting via adaptive multi-head alignment and residual fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20597–20605, 2026.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024.

Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017.

Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. TSMixer: Lightweight MLP-mixer model for multivariate time series forecasting. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 459–469, 2023.

Di Ge, Bo Liu, Yuhang Cheng, and Zhongjie Zhu. Rethinking decomposition for time series forecasting: Learning from temporal patterns. *Expert Systems with Applications*, 287:128044, 2025.

Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Xuehai He and Xin Wang. Multimodal graph transformer for multimodal question answering. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 189–200. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.15.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. GPT4MTS: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 23343–23351, 2024.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-fang Li, Shirui Pan, et al. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

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*, 58(7):216, 2025.

Xiangjie Kong, Zhenghao Chen, Weiyao Liu, Kaili Ning, Lechao Zhang, Syauqie Muhammad Marier, Yichen Liu, Yuhao Chen, and Feng Xia. Deep learning for time series forecasting: a survey. *International Journal of Machine Learning and Cybernetics*, 16(7):5079–5112, 2025.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences*, 379(2194), 2021.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

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

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's time: Towards the next generation of time series forecasting benchmarks. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90:106181, 2020.

Lefei Shen, Mouxiang Chen, Xu Liu, Han Fu, Xiaoxue Ren, Jianling Sun, Zhuo Li, and Chenghao Liu. VisionTS++: Cross-modal time series foundation model with continual pre-trained vision backbones. *arXiv preprint arXiv:2508.04379*, 2025.

Ali Asghar Taheri, Ali Abdali, and Abbas Rabiee. Indoor distribution transformers oil temperature prediction using new electro-thermal resistance model and normal cyclic overloading strategy: an experimental case study. *IET Generation, Transmission & Distribution*, 14(24):5792–5803, 2020. doi: 10.1049/iet-gtd.2020.0457.

Hua Tang, Chong Zhang, Mingyu Jin, Qinkai Yu, Zhenting Wang, Xiaobo Jin, Yongfeng Zhang, and Mengnan Du. Time series forecasting with llms: Understanding and enhancing model capabilities. *ACM SIGKDD Explorations Newsletter*, 26(2):109–118, 2025.

José F Torres, Dalil Hadjout, Abderrazak Sebaa, Francisco Martínez-Álvarez, and Alicia Troncoso. Deep learning for time series forecasting: a survey. *Big data*, 9(1):3–21, 2021.

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. *Advances in Neural Information Processing Systems*, 37:469–498, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*, 2021.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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

Yi-Fan Zhang, Qingsong Wen, Xue Wang, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Onenet: Enhancing time series forecasting models under concept drift by online ensembling. *Advances in Neural Information Processing Systems*, 36:69949–69980, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE transactions on intelligent transportation systems*, 21(9):3848–3858, 2019.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78478–78497. PMLR, 2025.

Haoyi Zhou. Electricity transformer dataset (ETDataset). https://github.com/zhouhaoyi/ETDataset, n.d. Accessed September 25, 2026.

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

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems(NeurIPS)*, 36:43322–43355, 2023.

## A Method Details

**Input preparation.** Each channel is standardized with training-split statistics. The temporal branch then normalizes each observed window with its own scale and either its mean or last observed value, according to the dataset configuration. Predictions are transformed back to the

dataset-standardized scale for evaluation. Image and video relations use mean-centered, variance-normalized windows. Learned patch and channel embeddings identify the temporal tokens; modality-specific LayerNorm parameters and scalar gates adjust the text and visual tokens before fusion.

**Text prompts.** The prompt combines the forecast task, dataset, sampling frequency, input length, horizon, channel count, and observed-window statistics. It is tokenized with a maximum of 192 tokens. Its template is:

```
Forecast the next {H} steps from {L} observed steps.
Dataset={dataset}; domain={domain};
sampling={frequency}; channels={C}.
Normalized window: min={min}, max={max}, mean={mean},
std={std}; trend={trend}.
RGB relations: DTW similarity, covariance, Pearson
correlation of spectral magnitudes.
```

**Visual relations.** For the full-window FFT magnitude and each of its local frequency patches, we compute DTW similarity, covariance, and Pearson correlation between channels. The three matrices are normalized separately to $[0, 1]$ and stacked as RGB channels. Rows and columns index variables, so each RGB pixel represents a channel pair rather than a physical location. This layout exposes joint patterns across relation measures to the visual encoder, while the bias path preserves explicit channel-pair correspondence. DTW distance $D_{ij}$ between z-normalized channel spectra is converted to similarity by $\exp(-D_{ij}/\tau)$, with $\tau = 1$ before min–max normalization. Covariance is clipped at the 1st and 99th percentiles before min–max normalization, and Pearson correlation is mapped to $[0, 1]$ by $(R_{\text{Pearson}} + 1)/2$.

**Attention bias.** The cached RGB matrices are reused for attention injection. Each channel is mapped to $[-1, 1]$ by $2S - 1$, combined with the DTW, covariance, and Pearson weights, and scaled by a learned positive factor $\gamma = \text{softplus}(\alpha)$. For the Pearson component, retaining the sign distinguishes aligned and opposing spectral variations. Negative contributions lower the corresponding logits without masking channel interactions. Attention weights are jointly determined by the learned query and key projections and the scaled relation bias. Local relations fill the diagonal $C \times C$ blocks of the patch-major temporal-token submatrix; the full-window relation fills its off-diagonal blocks. The structural bias is applied in decoder layers 0 and 1. Relation weights and the initial scale follow the dataset configurations; the fixed TIME settings are listed in Table 7.

# B Experimental Details

## B.1 Datasets and Protocol

Tables 5 and 6 list the input and prediction lengths used in the supervised evaluation. The ETT datasets follow the chronological 12/4/4-month split. Standard-benchmark evaluation follows the source implementations' chronological training, validation, and test partitions. Input and prediction lengths are in time steps; timestamp counts refer to the processed files. The ten TIME series use the processed CSV release of TIME (Qiao et al., 2026) and the forecast horizons attached to those tasks.

Table 5: Standard benchmarks: dimensions and forecasting horizons.

| Dataset | Channels | Input length | Sampling interval | Prediction horizons |
|---|---|---|---|---|
| ETTh1 | 7 | 96 | 1 hour | $\{96, 192, 336, 720\}$ |
| ETTh2 | 7 | 96 | 1 hour | $\{96, 192, 336, 720\}$ |
| ETTm1 | 7 | 96 | 15 min | $\{96, 192, 336, 720\}$ |
| ETTm2 | 7 | 96 | 15 min | $\{96, 192, 336, 720\}$ |
| Weather | 21 | 96 | 10 min | $\{96, 192, 336, 720\}$ |

Continued on next page

Table 5 (continued)

| Dataset | Channels | Input length | Sampling interval | Prediction horizons |
|---|---|---|---|---|
| ILI | 7 | 104 | 1 week | $\{24, 36, 48, 60\}$ |
| Exchange | 8 | 96 | 1 day | $\{96, 192, 336, 720\}$ |
| Electricity | 321 | 96 | 1 hour | $\{96, 192, 336, 720\}$ |

Table 6: TIME datasets: dimensions and forecasting horizons ($L = 96$).

| Dataset | Timestamps | Channels | Sampling interval | Prediction horizons |
|---|---|---|---|---|
| SG PM2.5 | 38688 | 5 | 1 hour | $\{24, 72, 168\}$ |
| OpenElectricity NEM | 43488 | 10 | 5 min | $\{24, 96, 288\}$ |
| MetroPT-3 | 17809 | 6 | 5 min | $\{48, 288, 576\}$ |
| NE China Wind | 8764 | 4 | 1 hour | $\{24, 72, 168\}$ |
| EWELD Load | 20544 | 10 | 15 min | $\{24, 96, 672\}$ |
| Oil Price | 5035 | 12 | business day | $\{20\}$ |
| Australia Solar | 35064 | 3 | 1 hour | $\{24, 72, 168\}$ |
| Coastal T-S | 8198 | 3 | 20 min | $\{9, 72, 216\}$ |
| Crypto | 2842 | 4 | 1 day | $\{30\}$ |
| Supply Chain Customer | 2007 | 36 | 1 day | $\{30\}$ |

**Evaluation scope.** Tables 5 and 6 span different channel counts, sampling intervals, and forecast lengths. Electricity has 321 channels, compared with seven in each ETT dataset; the TIME tasks range from nine-step Coastal T-S forecasts to 672-step EWELD forecasts. A fixed number of steps therefore represents different physical durations. This motivates the dataset-wise horizon averages in Section 4.1: each method is compared on the same horizons within a task, and errors are not pooled across heterogeneous tasks.

For TIME, model input is formed from the 96 observed timestamps preceding each forecast origin. Missing input values are forward-filled and then filled with the training-period median where needed. Channels are standardized using valid training observations. The loss and evaluation metrics include only originally observed target values. Each model is trained separately for each TIME dataset and prediction length using the same prepared series, valid-target masks, and chronological partitions.

### B.2 Model and Training Configuration

Table 7: MM-TS configuration; TIME settings are shared across ten datasets.

| Parameter | Value |
|---|---|
| Backbone | Qwen3-VL-2B-Instruct |
| Temporal patch length / stride | 16 / 8 |
| Attention-bias layers | 0, 1 |
| LoRA target layers | 0, 1, 26, 27 |
| LoRA target modules | q_proj, k_proj |
| LoRA rank / dropout | 8 / 0.05 |
| Maximum epochs / patience | 20 / 3 |
| Effective training batch | 32 |
| Prompt token budget | 192 |
| TIME learning rate / weight decay | $2 \times 10^{-4}$ / 0.0012 |
| TIME relation weights (DTW, Cov, Pearson) | 2.5, 2.0, 2.5 |
| TIME initial bias scale | 0.06 |

MM-TS freezes the Qwen3-VL base weights while training the temporal projection, patch and channel embeddings, modality normalization and gates, forecasting head, bias scale, and the selected

LoRA adapters. We select the checkpoint with the lowest validation MSE. Dataset-specific standard-benchmark settings follow the released performance scripts; the ten TIME tasks use one fixed MM-TS configuration. The same prediction horizons and evaluation masks are applied to all methods within each dataset.

### B.3 TimesFM Baseline

The TimesFM comparison starts from the `google/timesfm-1.0-200m-pytorch` checkpoint. Task adaptation updates its residual input and output blocks while the pretrained Transformer is frozen. The model is evaluated on the same observed input length, target horizons, splits, and original valid target values as the other methods.

## C Detailed Supervised Results

Tables 8 and 9 report all supervised results, with one table per benchmark suite. Each table compares all ten methods in separate MSE and MAE panels. $H$ is the prediction length in time steps, and Avg is the arithmetic mean over the displayed horizons of that dataset. Model names and column order are consistent across tables; Abbreviated model headers follow Section 4. Bold red and underlined blue mark the lowest and second-lowest distinct displayed values; ties receive the same marking.

**Dataset-specific comparisons.** MM-TS leads six standard datasets and eight TIME datasets in average MSE. It ranks second on ILI (1.560 versus 1.514 for PatchTST) and Exchange (0.351 versus 0.336 for DLinear). Time-VLM leads on Oil Price (0.064 versus 0.065) and Crypto (0.093 versus 0.096). The horizon and metric profiles below characterize the distribution of these gains and complementary task-specific strengths.

### C.1 Standard Benchmarks

Table 8: Standard-benchmark forecasting: MSE and MAE.

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *MSE* | | | | | | | | | | | |
| ETTh1 | 96 | **0.379** | 0.384 | 0.403 | 0.381 | 0.383 | 0.386 | 0.415 | 0.384 | 0.381 | 0.387 |
| | 192 | **0.425** | 0.433 | 0.463 | 0.427 | 0.441 | 0.442 | 0.461 | 0.437 | 0.433 | 0.437 |
| | 336 | **0.462** | 0.470 | 0.495 | 0.470 | 0.497 | 0.486 | 0.501 | 0.492 | 0.473 | 0.481 |
| | 720 | **0.469** | 0.474 | 0.498 | 0.470 | 0.533 | 0.503 | 0.502 | 0.522 | 0.474 | 0.519 |
| | *Avg* | **0.434** | 0.440 | 0.465 | 0.437 | 0.464 | 0.454 | 0.470 | 0.459 | 0.441 | 0.456 |
| ETTh2 | 96 | **0.283** | 0.291 | 0.299 | 0.310 | 0.310 | 0.296 | 0.302 | 0.340 | 0.293 | 0.333 |
| | 192 | **0.362** | 0.375 | 0.383 | 0.383 | 0.404 | 0.380 | 0.388 | 0.403 | 0.372 | 0.477 |
| | 336 | **0.407** | 0.428 | 0.425 | 0.418 | 0.439 | 0.429 | 0.425 | 0.452 | 0.422 | 0.594 |
| | 720 | **0.421** | 0.432 | 0.442 | 0.434 | 0.434 | 0.432 | 0.429 | 0.461 | 0.443 | 0.831 |
| | *Avg* | **0.368** | 0.382 | 0.387 | 0.386 | 0.397 | 0.384 | 0.386 | 0.414 | 0.383 | 0.559 |
| ETTm1 | 96 | 0.325 | 0.328 | 0.330 | 0.334 | **0.322** | 0.334 | 0.330 | 0.338 | 0.327 | 0.345 |
| | 192 | **0.363** | 0.368 | 0.370 | 0.381 | 0.368 | 0.387 | 0.368 | 0.374 | 0.367 | 0.380 |
| | 336 | **0.392** | 0.401 | 0.400 | 0.410 | 0.400 | 0.427 | 0.398 | 0.410 | 0.404 | 0.413 |
| | 720 | **0.451** | 0.462 | 0.465 | 0.466 | 0.463 | 0.490 | 0.453 | 0.477 | 0.458 | 0.474 |
| | *Avg* | **0.383** | 0.390 | 0.391 | 0.398 | 0.388 | 0.410 | 0.388 | 0.399 | 0.389 | 0.403 |

Continued on next page

Table 8 (continued)

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ETTm2 | 96 | **0.173** | 0.176 | 0.186 | 0.183 | 0.183 | 0.180 | 0.175 | 0.187 | 0.177 | 0.193 |
| | 192 | **0.238** | 0.240 | 0.251 | 0.254 | 0.252 | 0.250 | 0.241 | 0.249 | 0.243 | 0.284 |
| | 336 | **0.297** | 0.298 | 0.313 | 0.327 | 0.325 | 0.311 | 0.305 | 0.322 | 0.310 | 0.369 |
| | 720 | **0.396** | 0.400 | 0.416 | 0.429 | 0.436 | 0.412 | 0.402 | 0.408 | 0.416 | 0.554 |
| | *Avg* | **0.276** | 0.278 | 0.292 | 0.298 | 0.299 | 0.288 | 0.281 | 0.291 | 0.286 | 0.350 |
| Weather | 96 | **0.160** | 0.180 | 0.196 | 0.168 | 0.172 | 0.174 | 0.177 | 0.172 | 0.167 | 0.196 |
| | 192 | **0.201** | 0.228 | 0.242 | 0.212 | 0.221 | 0.221 | 0.225 | 0.219 | 0.206 | 0.237 |
| | 336 | **0.260** | 0.282 | 0.293 | 0.268 | 0.277 | 0.277 | 0.278 | 0.280 | 0.263 | 0.283 |
| | 720 | **0.340** | 0.359 | 0.368 | 0.346 | 0.356 | 0.358 | 0.353 | 0.366 | 0.341 | 0.345 |
| | *Avg* | **0.240** | 0.262 | 0.275 | 0.249 | 0.257 | 0.258 | 0.259 | 0.259 | 0.244 | 0.265 |
| ILI | 24 | 1.355 | 2.083 | 2.094 | 1.926 | 1.548 | 2.441 | **1.300** | 2.139 | 4.351 | 2.276 |
| | 36 | **1.547** | 2.006 | 2.139 | 1.764 | 1.660 | 2.448 | 1.655 | 2.276 | 4.533 | 2.239 |
| | 48 | 1.642 | 2.072 | 2.006 | **1.617** | 1.676 | 2.593 | 1.659 | 2.017 | 4.441 | 2.302 |
| | 60 | 1.697 | 2.098 | 1.954 | 1.700 | 1.724 | 2.738 | **1.443** | 2.205 | 4.316 | 2.569 |
| | *Avg* | 1.560 | 2.065 | 2.048 | 1.752 | 1.652 | 2.555 | **1.514** | 2.159 | 4.410 | 2.346 |
| Exchange | 96 | **0.077** | 0.088 | 0.092 | 0.084 | 0.087 | 0.104 | 0.086 | 0.117 | 0.086 | 0.078 |
| | 192 | **0.164** | 0.182 | 0.183 | 0.176 | 0.179 | 0.201 | 0.183 | 0.223 | 0.179 | 0.166 |
| | 336 | **0.308** | 0.347 | 0.340 | 0.335 | 0.322 | 0.341 | 0.338 | 0.405 | 0.338 | 0.317 |
| | 720 | 0.856 | 0.861 | 0.862 | 0.839 | 0.842 | 0.866 | 0.879 | 1.110 | 0.925 | **0.783** |
| | *Avg* | 0.351 | 0.370 | 0.369 | 0.358 | 0.357 | 0.378 | 0.371 | 0.464 | 0.382 | **0.336** |
| Electricity | 96 | **0.163** | 0.188 | 0.222 | 0.166 | 0.173 | 0.168 | 0.195 | 0.168 | 0.164 | 0.197 |
| | 192 | **0.175** | 0.191 | 0.220 | 0.176 | 0.184 | 0.181 | 0.199 | 0.184 | 0.178 | 0.196 |
| | 336 | **0.190** | 0.204 | 0.236 | 0.193 | 0.203 | 0.192 | 0.215 | 0.194 | 0.193 | 0.209 |
| | 720 | 0.230 | 0.250 | 0.271 | 0.231 | 0.245 | **0.225** | 0.256 | 0.238 | 0.238 | 0.245 |
| | *Avg* | **0.190** | 0.208 | 0.237 | 0.192 | 0.201 | 0.192 | 0.216 | 0.196 | 0.194 | 0.212 |
| *MAE* | | | | | | | | | | | |
| ETTh1 | 96 | **0.392** | 0.395 | 0.415 | 0.397 | 0.400 | 0.405 | 0.420 | 0.403 | 0.394 | 0.400 |
| | 192 | **0.422** | 0.424 | 0.449 | 0.423 | 0.433 | 0.436 | 0.446 | 0.429 | 0.423 | 0.432 |
| | 336 | **0.441** | 0.449 | 0.465 | 0.445 | 0.466 | 0.457 | 0.466 | 0.471 | 0.444 | 0.459 |
| | 720 | **0.462** | 0.475 | 0.489 | 0.467 | 0.506 | 0.491 | 0.489 | 0.501 | 0.465 | 0.516 |
| | *Avg* | **0.429** | 0.436 | 0.455 | 0.433 | 0.451 | 0.448 | 0.455 | 0.451 | 0.432 | 0.452 |
| ETTh2 | 96 | **0.337** | 0.339 | 0.348 | 0.353 | 0.353 | 0.348 | 0.348 | 0.374 | 0.343 | 0.387 |
| | 192 | **0.388** | 0.392 | 0.399 | 0.395 | 0.409 | 0.400 | 0.400 | 0.414 | 0.393 | 0.476 |
| | 336 | **0.424** | 0.437 | 0.434 | 0.428 | 0.442 | 0.432 | 0.433 | 0.452 | 0.434 | 0.541 |
| | 720 | **0.441** | 0.448 | 0.451 | 0.450 | 0.449 | 0.446 | 0.445 | 0.468 | 0.454 | 0.657 |
| | *Avg* | **0.398** | 0.404 | 0.408 | 0.406 | 0.413 | 0.407 | 0.407 | 0.427 | 0.406 | 0.515 |

Continued on next page

Table 8 (continued)

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ETTm1 | 96 | **0.360** | 0.365 | 0.372 | 0.369 | 0.364 | 0.368 | 0.368 | 0.375 | 0.361 | 0.372 |
| | 192 | **0.383** | 0.385 | 0.390 | 0.396 | 0.388 | 0.391 | 0.385 | 0.387 | 0.384 | 0.389 |
| | 336 | **0.404** | 0.408 | 0.411 | 0.417 | 0.412 | 0.420 | 0.410 | 0.411 | 0.407 | 0.413 |
| | 720 | **0.431** | 0.440 | 0.445 | 0.445 | 0.448 | 0.458 | 0.439 | 0.449 | 0.435 | 0.453 |
| | *Avg* | **0.395** | 0.399 | 0.404 | 0.407 | 0.403 | 0.409 | 0.400 | 0.405 | 0.397 | 0.407 |
| ETTm2 | 96 | **0.252** | 0.261 | 0.271 | 0.274 | 0.263 | 0.263 | 0.259 | 0.267 | 0.260 | 0.292 |
| | 192 | **0.300** | 0.302 | 0.311 | 0.319 | 0.310 | 0.309 | 0.303 | 0.309 | 0.303 | 0.362 |
| | 336 | **0.336** | 0.338 | 0.351 | 0.365 | 0.355 | 0.348 | 0.343 | 0.351 | 0.344 | 0.427 |
| | 720 | **0.396** | **0.396** | 0.408 | 0.424 | 0.422 | 0.407 | 0.400 | 0.403 | 0.406 | 0.522 |
| | *Avg* | **0.321** | 0.325 | 0.335 | 0.346 | 0.337 | 0.332 | 0.326 | 0.333 | 0.328 | 0.401 |
| Weather | 96 | **0.206** | 0.220 | 0.234 | 0.209 | 0.212 | 0.214 | 0.218 | 0.220 | 0.208 | 0.255 |
| | 192 | **0.248** | 0.260 | 0.272 | 0.250 | 0.255 | 0.254 | 0.259 | 0.261 | 0.249 | 0.296 |
| | 336 | **0.287** | 0.298 | 0.309 | 0.288 | 0.296 | 0.296 | 0.298 | 0.306 | 0.289 | 0.335 |
| | 720 | 0.341 | 0.348 | 0.354 | **0.339** | 0.347 | 0.349 | 0.348 | 0.359 | 0.341 | 0.381 |
| | *Avg* | **0.271** | 0.281 | 0.292 | 0.272 | 0.277 | 0.278 | 0.281 | 0.286 | 0.272 | 0.317 |
| ILI | 24 | 0.828 | 0.938 | 0.950 | 0.895 | 0.771 | 1.072 | **0.738** | 0.995 | 1.535 | 1.059 |
| | 36 | **0.797** | 0.922 | 0.979 | 0.860 | 0.806 | 1.088 | 0.897 | 0.980 | 1.565 | 1.058 |
| | 48 | 0.841 | 0.954 | 0.937 | **0.819** | 0.822 | 1.118 | 0.877 | 0.969 | 1.541 | 1.081 |
| | 60 | 0.883 | 0.948 | 0.948 | 0.850 | 0.816 | 1.153 | **0.788** | 1.022 | 1.510 | 1.158 |
| | *Avg* | 0.837 | 0.940 | 0.953 | 0.856 | **0.804** | 1.108 | 0.825 | 0.992 | 1.538 | 1.089 |
| Exchange | 96 | **0.197** | 0.204 | 0.213 | 0.202 | 0.205 | 0.229 | 0.204 | 0.244 | 0.206 | 0.199 |
| | 192 | **0.284** | 0.303 | 0.305 | 0.299 | 0.302 | 0.323 | 0.303 | 0.341 | 0.300 | 0.296 |
| | 336 | **0.409** | 0.424 | 0.423 | 0.418 | 0.412 | 0.425 | 0.422 | 0.469 | 0.417 | 0.419 |
| | 720 | 0.714 | 0.698 | 0.700 | 0.688 | 0.685 | 0.709 | 0.705 | 0.794 | 0.735 | **0.670** |
| | *Avg* | 0.401 | 0.407 | 0.410 | 0.402 | 0.401 | 0.422 | 0.408 | 0.462 | 0.414 | **0.396** |
| Electricity | 96 | **0.254** | 0.274 | 0.311 | 0.258 | 0.257 | 0.266 | 0.285 | 0.272 | 0.256 | 0.282 |
| | 192 | 0.262 | 0.277 | 0.310 | **0.260** | 0.269 | 0.263 | 0.289 | 0.289 | 0.267 | 0.285 |
| | 336 | **0.286** | 0.291 | 0.325 | 0.287 | 0.287 | 0.289 | 0.305 | 0.300 | 0.288 | 0.301 |
| | 720 | 0.320 | 0.326 | 0.350 | **0.310** | 0.322 | 0.317 | 0.337 | 0.328 | 0.335 | 0.333 |
| | *Avg* | 0.281 | 0.292 | 0.324 | **0.279** | 0.284 | 0.284 | 0.304 | 0.297 | 0.287 | 0.300 |

**Settings.** Table 8 expands Table 1(a) into individual horizons, with MSE and MAE reported separately.

**Results and analysis.** The detailed results characterize where the aggregate gains arise. On ETTm2, MM-TS leads at all four horizons, supporting effectiveness across the evaluated forecast lengths. Exchange presents a more horizon-dependent pattern: MM-TS leads at 96, 192, and 336 steps, while DLinear leads at 720 steps and obtains the lower average MSE. ILI gives MM-TS the second-lowest average MSE, with its strongest ranking at 36 steps. These patterns identify forecast horizon as an informative dimension of the comparison. The two metrics provide a complementary view: Electricity favors MM-TS in MSE and VisionTS in MAE, reflecting their different weighting of prediction errors. Section 4.1 therefore reports dataset-level performance alongside the horizon and metric profiles provided here.

### C.2 TIME DATASETS

Table 9: TIME forecasting: MSE and MAE.

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *MSE* | | | | | | | | | | | |
| SG PM2.5 | 24 | **0.902** | 0.950 | 0.974 | 0.959 | 0.958 | 0.927 | 0.953 | 0.948 | 0.906 | 0.961 |
| | 72 | **1.036** | 1.107 | 1.132 | 1.096 | 1.094 | 1.079 | 1.109 | 1.099 | 1.041 | **1.036** |
| | 168 | **1.130** | 1.259 | 1.294 | 1.249 | 1.248 | 1.229 | 1.263 | 1.255 | 1.139 | 1.134 |
| | *Avg* | **1.023** | 1.105 | 1.133 | 1.102 | 1.100 | 1.078 | 1.108 | 1.101 | 1.029 | 1.044 |
| NEM | 24 | 0.236 | 0.290 | 0.342 | 0.281 | 0.279 | 0.258 | 0.298 | 0.282 | **0.234** | 0.340 |
| | 96 | **0.566** | 0.745 | 0.876 | 0.768 | 0.750 | 0.697 | 0.776 | 0.603 | 0.578 | 0.905 |
| | 288 | **0.658** | 0.791 | 0.876 | 0.863 | 0.792 | 0.761 | 0.808 | 0.663 | 0.686 | 0.839 |
| | *Avg* | **0.487** | 0.608 | 0.698 | 0.637 | 0.607 | 0.572 | 0.628 | 0.516 | 0.499 | 0.695 |
| MetroPT-3 | 48 | 0.682 | 0.766 | 0.821 | 0.712 | 0.703 | 0.756 | 0.769 | 0.738 | **0.675** | 0.709 |
| | 288 | **0.685** | 0.788 | 0.831 | 0.765 | 0.751 | 0.810 | 0.801 | 0.780 | 0.693 | 0.721 |
| | 576 | **0.702** | 0.827 | 0.863 | 0.780 | 0.748 | 0.841 | 0.836 | 0.821 | 0.706 | 0.728 |
| | *Avg* | **0.690** | 0.794 | 0.838 | 0.752 | 0.734 | 0.802 | 0.802 | 0.780 | 0.692 | 0.719 |
| NE Wind | 24 | **0.192** | 0.204 | 0.339 | 0.233 | 0.212 | 0.213 | 0.205 | 0.257 | 0.195 | 0.203 |
| | 72 | **0.398** | 0.453 | 0.569 | 0.481 | 0.465 | 0.461 | 0.455 | 0.514 | 0.405 | 0.451 |
| | 168 | **0.509** | 0.616 | 0.692 | 0.655 | 0.629 | 0.624 | 0.615 | 0.638 | 0.518 | 0.625 |
| | *Avg* | **0.366** | 0.424 | 0.533 | 0.456 | 0.435 | 0.433 | 0.425 | 0.470 | 0.373 | 0.427 |
| EWELD | 24 | **0.226** | 0.271 | 0.346 | 0.272 | 0.228 | 0.239 | 0.278 | 0.256 | 0.244 | 0.283 |
| | 96 | **0.307** | 0.343 | 0.426 | 0.344 | 0.311 | 0.311 | 0.338 | 0.317 | 0.325 | 0.332 |
| | 672 | 0.338 | **0.333** | 0.472 | 0.342 | 0.347 | 0.336 | 0.350 | 0.344 | 0.412 | 0.365 |
| | *Avg* | **0.290** | 0.316 | 0.414 | 0.319 | 0.295 | 0.296 | 0.322 | 0.305 | 0.327 | 0.327 |
| Oil Price | 20 | 0.065 | **0.064** | 0.128 | 0.070 | 0.071 | 0.077 | 0.070 | 0.133 | 0.091 | 0.066 |
| | *Avg* | 0.065 | **0.064** | 0.128 | 0.070 | 0.071 | 0.077 | 0.070 | 0.133 | 0.091 | 0.066 |
| AU Solar | 24 | **0.188** | 0.224 | 0.389 | 0.201 | 0.190 | 0.199 | 0.260 | 0.224 | 0.198 | 0.246 |
| | 72 | **0.257** | 0.276 | 0.448 | 0.263 | 0.261 | 0.262 | 0.310 | 0.275 | 0.269 | 0.292 |
| | 168 | 0.298 | 0.304 | 0.512 | **0.297** | **0.297** | 0.306 | 0.343 | 0.311 | 0.310 | 0.330 |
| | *Avg* | **0.248** | 0.268 | 0.449 | 0.254 | 0.249 | 0.255 | 0.304 | 0.270 | 0.259 | 0.289 |
| Coastal T-S | 9 | **0.089** | 0.097 | 0.363 | 0.337 | 0.091 | 0.102 | 0.114 | 0.156 | 0.119 | 0.093 |
| | 72 | **0.367** | 0.393 | 0.868 | 0.501 | 0.407 | 0.372 | 0.415 | 0.387 | 0.464 | 0.374 |
| | 216 | 0.754 | 1.014 | 1.570 | 1.130 | 0.941 | 0.891 | 1.047 | 1.010 | **0.739** | 0.933 |
| | *Avg* | **0.404** | 0.501 | 0.933 | 0.656 | 0.479 | 0.455 | 0.525 | 0.517 | 0.441 | 0.467 |
| Crypto | 30 | 0.096 | **0.093** | 0.352 | 0.166 | 0.107 | 0.119 | 0.117 | 0.158 | 0.443 | 0.099 |
| | *Avg* | 0.096 | **0.093** | 0.352 | 0.166 | 0.107 | 0.119 | 0.117 | 0.158 | 0.443 | 0.099 |
| Supply Chain | 30 | **0.617** | 0.731 | 1.733 | 1.404 | 0.625 | 0.684 | 0.869 | 1.122 | 0.745 | 0.672 |
| | *Avg* | **0.617** | 0.731 | 1.733 | 1.404 | 0.625 | 0.684 | 0.869 | 1.122 | 0.745 | 0.672 |

Continued on next page

Table 9 (continued)

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *MAE* | | | | | | | | | | | |
| SG PM2.5 | 24 | **0.704** | 0.722 | 0.737 | 0.736 | 0.723 | 0.707 | 0.726 | 0.725 | 0.709 | 0.730 |
| | 72 | **0.759** | 0.780 | 0.789 | 0.779 | 0.778 | 0.767 | 0.782 | 0.781 | **0.759** | 0.760 |
| | 168 | 0.799 | 0.828 | 0.841 | 0.825 | 0.824 | 0.819 | 0.834 | 0.830 | 0.806 | **0.796** |
| | *Avg* | **0.754** | 0.777 | 0.789 | 0.780 | 0.775 | 0.764 | 0.781 | 0.778 | 0.758 | 0.762 |
| NEM | 24 | **0.294** | 0.328 | 0.365 | 0.307 | 0.301 | 0.295 | 0.322 | 0.332 | 0.296 | 0.351 |
| | 96 | **0.531** | 0.614 | 0.680 | 0.608 | 0.588 | 0.569 | 0.622 | 0.548 | 0.537 | 0.668 |
| | 288 | **0.574** | 0.635 | 0.683 | 0.665 | 0.622 | 0.602 | 0.643 | 0.575 | 0.597 | 0.652 |
| | *Avg* | **0.466** | 0.526 | 0.576 | 0.526 | 0.504 | 0.489 | 0.529 | 0.485 | 0.477 | 0.557 |
| MetroPT-3 | 48 | 0.664 | 0.696 | 0.739 | 0.684 | 0.669 | 0.687 | 0.701 | 0.686 | **0.662** | 0.683 |
| | 288 | **0.676** | 0.719 | 0.741 | 0.724 | 0.708 | 0.728 | 0.724 | 0.713 | 0.681 | 0.705 |
| | 576 | **0.689** | 0.736 | 0.755 | 0.734 | 0.710 | 0.735 | 0.740 | 0.731 | 0.694 | 0.711 |
| | *Avg* | **0.676** | 0.717 | 0.745 | 0.714 | 0.696 | 0.717 | 0.722 | 0.710 | 0.679 | 0.700 |
| NE Wind | 24 | **0.287** | 0.301 | 0.418 | 0.335 | 0.301 | 0.315 | 0.303 | 0.360 | 0.289 | 0.306 |
| | 72 | **0.437** | 0.483 | 0.565 | 0.508 | 0.489 | 0.495 | 0.484 | 0.528 | 0.440 | 0.494 |
| | 168 | **0.527** | 0.602 | 0.649 | 0.626 | 0.605 | 0.611 | 0.602 | 0.616 | 0.532 | 0.621 |
| | *Avg* | **0.417** | 0.462 | 0.544 | 0.490 | 0.465 | 0.474 | 0.463 | 0.502 | 0.420 | 0.474 |
| EWELD | 24 | **0.297** | 0.335 | 0.402 | 0.332 | 0.300 | 0.311 | 0.343 | 0.332 | 0.337 | 0.346 |
| | 96 | **0.361** | 0.389 | 0.457 | 0.382 | 0.366 | 0.364 | 0.386 | 0.375 | 0.412 | 0.402 |
| | 672 | 0.390 | **0.387** | 0.495 | 0.389 | 0.404 | 0.389 | 0.404 | 0.396 | 0.488 | 0.459 |
| | *Avg* | **0.349** | 0.371 | 0.452 | 0.368 | 0.357 | 0.355 | 0.377 | 0.368 | 0.412 | 0.403 |
| Oil Price | 20 | **0.175** | 0.176 | 0.263 | 0.188 | 0.190 | 0.196 | 0.189 | 0.263 | 0.224 | 0.178 |
| | *Avg* | **0.175** | 0.176 | 0.263 | 0.188 | 0.190 | 0.196 | 0.189 | 0.263 | 0.224 | 0.178 |
| AU Solar | 24 | **0.267** | 0.324 | 0.472 | 0.287 | 0.269 | 0.288 | 0.353 | 0.319 | 0.293 | 0.338 |
| | 72 | **0.338** | 0.364 | 0.506 | 0.343 | 0.340 | 0.344 | 0.389 | 0.356 | 0.357 | 0.372 |
| | 168 | **0.358** | 0.374 | 0.541 | 0.360 | 0.362 | 0.369 | 0.407 | 0.372 | 0.384 | 0.387 |
| | *Avg* | **0.321** | 0.354 | 0.506 | 0.330 | 0.324 | 0.333 | 0.383 | 0.349 | 0.345 | 0.365 |
| Coastal T-S | 9 | **0.154** | 0.167 | 0.376 | 0.345 | 0.155 | 0.177 | 0.186 | 0.235 | 0.203 | 0.159 |
| | 72 | **0.349** | 0.354 | 0.613 | 0.412 | 0.369 | 0.354 | 0.367 | 0.382 | 0.453 | 0.385 |
| | 216 | 0.544 | 0.566 | 0.822 | 0.631 | 0.557 | **0.540** | 0.578 | 0.581 | 0.574 | 0.617 |
| | *Avg* | **0.349** | 0.362 | 0.603 | 0.463 | 0.361 | 0.357 | 0.377 | 0.400 | 0.410 | 0.387 |
| Crypto | 30 | 0.222 | **0.221** | 0.472 | 0.309 | 0.238 | 0.254 | 0.249 | 0.302 | 0.549 | 0.228 |
| | *Avg* | 0.222 | **0.221** | 0.472 | 0.309 | 0.238 | 0.254 | 0.249 | 0.302 | 0.549 | 0.228 |
| Supply Chain | 30 | **0.529** | 0.598 | 0.996 | 0.898 | 0.534 | 0.578 | 0.674 | 0.784 | 0.611 | 0.562 |
| | *Avg* | **0.529** | 0.598 | 0.996 | 0.898 | 0.534 | 0.578 | 0.674 | 0.784 | 0.611 | 0.562 |

**Settings.** Table 9 expands Table 1(b) using each TIME task's horizon set and valid-target evaluation mask.

**Results and analysis.** The TIME evaluation tests the same structural framework across additional variable sets and temporal resolutions, complementing the established benchmarks. MM-TS leads eight of ten tasks in average MSE and nine in average MAE. Coastal T-S provides a concrete example: relative to TSMixer, MSE decreases from 0.119 to 0.089 at 9 steps and from 0.464 to 0.367 at 72 steps. These improvements produce an 8.4% lower average MSE, with TSMixer retaining a small advantage at 216 steps. On Oil Price, MM-TS has lower MAE, while Time-VLM has slightly lower MSE. Reporting both metrics makes their different error profiles visible. Across the additional TIME tasks, the results extend the supervised evidence in Section 4.1 to more variable sets and temporal resolutions, with gains that vary by task and horizon.

### C.3 Run-to-run Variation

Table 10: Run-to-run variation of horizon-averaged MSE and MAE.

| Dataset | Best baseline | MM-TS MSE | Baseline MSE | MM-TS MAE | Baseline MAE |
|---|---|---|---|---|---|
| *Standard benchmarks* | | | | | |
| ETTh1 | VisionTS | $0.434 \pm 0.0018$ | $0.437 \pm 0.0011$ | $0.429 \pm 0.0016$ | $0.433 \pm 0.0007$ |
| ETTh2 | Time-VLM | $0.368 \pm 0.0004$ | $0.382 \pm 0.0005$ | $0.398 \pm 0.0000$ | $0.404 \pm 0.0004$ |
| ETTm1 | PatchTST | $0.383 \pm 0.0023$ | $0.388 \pm 0.0008$ | $0.395 \pm 0.0013$ | $0.400 \pm 0.0005$ |
| ETTm2 | Time-VLM | $0.276 \pm 0.0018$ | $0.278 \pm 0.0005$ | $0.321 \pm 0.0015$ | $0.325 \pm 0.0005$ |
| Weather | TSMixer | $0.240 \pm 0.0015$ | $0.244 \pm 0.0004$ | $0.271 \pm 0.0011$ | $0.272 \pm 0.0000$ |
| ILI | PatchTST | $1.560 \pm 0.0073$ | $1.514 \pm 0.0041$ | $0.837 \pm 0.0023$ | $0.825 \pm 0.0046$ |
| Exchange | DLinear | $0.351 \pm 0.0054$ | $0.336 \pm 0.0007$ | $0.401 \pm 0.0021$ | $0.396 \pm 0.0000$ |
| Electricity | VisionTS | $0.190 \pm 0.0013$ | $0.192 \pm 0.0005$ | $0.281 \pm 0.0013$ | $0.279 \pm 0.0004$ |
| *TIME datasets* | | | | | |
| SG PM2.5 | TSMixer | $1.023 \pm 0.0059$ | $1.029 \pm 0.0054$ | $0.754 \pm 0.0036$ | $0.758 \pm 0.0031$ |
| NEM | TSMixer | $0.487 \pm 0.0029$ | $0.499 \pm 0.0019$ | $0.466 \pm 0.0023$ | $0.477 \pm 0.0011$ |
| MetroPT-3 | TSMixer | $0.690 \pm 0.0018$ | $0.692 \pm 0.0046$ | $0.676 \pm 0.0015$ | $0.679 \pm 0.0030$ |
| NE Wind | TSMixer | $0.366 \pm 0.0019$ | $0.373 \pm 0.0017$ | $0.417 \pm 0.0011$ | $0.420 \pm 0.0012$ |
| EWELD | TimesFM | $0.290 \pm 0.0015$ | $0.295 \pm 0.0011$ | $0.349 \pm 0.0011$ | $0.357 \pm 0.0009$ |
| Oil Price | Time-VLM | $0.065 \pm 0.0007$ | $0.064 \pm 0.0004$ | $0.175 \pm 0.0008$ | $0.176 \pm 0.0011$ |
| AU Solar | TimesFM | $0.248 \pm 0.0011$ | $0.249 \pm 0.0027$ | $0.321 \pm 0.0008$ | $0.324 \pm 0.0026$ |
| Coastal T-S | TSMixer | $0.404 \pm 0.0009$ | $0.441 \pm 0.0065$ | $0.349 \pm 0.0004$ | $0.410 \pm 0.0035$ |
| Crypto | Time-VLM | $0.096 \pm 0.0009$ | $0.093 \pm 0.0015$ | $0.222 \pm 0.0015$ | $0.221 \pm 0.0019$ |
| Supply Chain | TimesFM | $0.617 \pm 0.0036$ | $0.625 \pm 0.0123$ | $0.529 \pm 0.0016$ | $0.534 \pm 0.0060$ |

**Settings.** For each model and dataset, we first average MSE and MAE over the evaluated horizons within each run, then report the mean $\pm$ sample standard deviation across five runs in Table 10. The comparator is selected by the lowest five-run mean MSE among the nine baselines, and the same comparator is used for both metrics.

**Results and analysis.** On ETTh2, the mean MSE gap is 0.014, compared with run-to-run standard deviations of 0.0004 and 0.0005; Coastal T-S also shows a separation larger than the reported dispersions. MetroPT-3 is more closely ranked, with means of 0.690 and 0.692. The repeated-run results therefore distinguish clearly separated mean errors from close comparisons within the overall dataset rankings. These results complement the dataset rankings in Section 4.1.

## D Few-shot and Zero-shot Forecasting Results

### D.1 Few-shot Forecasting

We use 5% or 10% of the original training split, preserving the validation and test splits. Tables 11 and 12 report the common evaluated horizons for the eight standard datasets; each Avg follows those horizons.

Table 11: Few-shot forecasting with 5% training data: MSE and MAE.

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *MSE* | | | | | | | | | | | |
| ETTh1 | 96 | **0.407** | 0.413 | 0.568 | 0.423 | 0.417 | 0.502 | 0.678 | 0.822 | 0.699 | 0.410 |
| | 192 | **0.456** | 0.464 | 0.704 | 0.460 | 0.475 | 0.566 | 0.714 | 0.820 | 0.473 | 0.462 |
| | 336 | **0.497** | 0.499 | 1.365 | 0.510 | 0.535 | 0.552 | 0.714 | 0.868 | 0.509 | 0.514 |
| | Avg | **0.453** | 0.459 | 0.879 | 0.464 | 0.476 | 0.540 | 0.702 | 0.837 | 0.560 | 0.462 |
| ETTh2 | 96 | **0.302** | 0.307 | 0.352 | 0.334 | 0.316 | 0.364 | 0.364 | 0.392 | 0.453 | 0.390 |
| | 192 | **0.394** | 0.395 | 0.465 | 0.402 | 0.422 | 0.452 | 0.444 | 0.463 | 0.397 | 0.502 |
| | 336 | **0.441** | 0.443 | 0.912 | 0.448 | 0.442 | 0.488 | 0.447 | 0.494 | 0.472 | 0.621 |
| | Avg | **0.379** | 0.382 | 0.576 | 0.395 | 0.393 | 0.435 | 0.418 | 0.450 | 0.441 | 0.504 |
| ETTm1 | 96 | 0.353 | 0.355 | 0.522 | 0.391 | **0.352** | 0.416 | 0.363 | 0.377 | 0.597 | 0.447 |
| | 192 | **0.382** | 0.384 | 0.539 | 0.417 | 0.393 | 0.466 | 0.385 | 0.429 | 0.534 | 0.480 |
| | 336 | **0.403** | 0.412 | 0.569 | 0.439 | 0.434 | 0.502 | 0.405 | 0.543 | 0.510 | 0.499 |
| | 720 | **0.462** | 0.475 | 0.759 | 0.499 | 0.498 | 0.567 | 0.469 | 0.570 | 0.663 | 0.550 |
| | Avg | **0.400** | 0.407 | 0.597 | 0.437 | 0.419 | 0.488 | 0.406 | 0.480 | 0.576 | 0.494 |
| ETTm2 | 96 | **0.176** | 0.184 | 0.200 | 0.208 | 0.193 | 0.201 | 0.178 | 0.196 | 0.193 | 0.211 |
| | 192 | **0.241** | 0.246 | 0.275 | 0.266 | 0.266 | 0.265 | 0.245 | 0.300 | 0.249 | 0.303 |
| | 336 | **0.301** | 0.304 | 0.325 | 0.323 | 0.338 | 0.326 | 0.307 | 0.333 | 0.318 | 0.378 |
| | 720 | **0.402** | 0.405 | 0.479 | 0.418 | 0.442 | 0.425 | 0.406 | 0.461 | 0.425 | 0.484 |
| | Avg | **0.280** | 0.285 | 0.320 | 0.304 | 0.310 | 0.304 | 0.284 | 0.323 | 0.296 | 0.344 |
| Weather | 96 | **0.173** | 0.190 | 0.197 | 0.190 | 0.178 | 0.200 | 0.197 | 0.203 | 0.185 | 0.231 |
| | 192 | **0.219** | 0.236 | 0.241 | 0.238 | 0.226 | 0.245 | 0.242 | 0.278 | 0.242 | 0.270 |
| | 336 | **0.272** | 0.290 | 0.300 | 0.291 | 0.282 | 0.296 | 0.292 | 0.300 | 0.292 | 0.312 |
| | 720 | **0.347** | 0.363 | 0.375 | 0.364 | 0.361 | 0.368 | 0.362 | 0.386 | 0.361 | 0.371 |
| | Avg | **0.253** | 0.270 | 0.278 | 0.271 | 0.262 | 0.277 | 0.273 | 0.292 | 0.270 | 0.296 |
| ILI | 24 | 2.527 | 3.837 | 4.275 | 2.117 | **1.691** | 4.451 | 2.273 | 3.407 | 8.485 | 8.276 |
| | Avg | 2.527 | 3.837 | 4.275 | 2.117 | **1.691** | 4.451 | 2.273 | 3.407 | 8.485 | 8.276 |
| Exchange | 96 | **0.082** | 0.086 | 0.134 | 0.085 | 0.091 | 0.125 | 0.131 | 0.137 | 0.158 | 0.131 |
| | 192 | **0.176** | 0.184 | 0.319 | 0.180 | 0.180 | 0.227 | 0.236 | 0.229 | 0.183 | 0.214 |
| | Avg | **0.129** | 0.135 | 0.227 | 0.133 | 0.135 | 0.176 | 0.184 | 0.183 | 0.171 | 0.173 |
| Electricity | 96 | 0.190 | 0.240 | 0.220 | 0.192 | **0.173** | 0.200 | 0.204 | 0.285 | 0.258 | 0.235 |
| | 192 | 0.205 | 0.237 | 0.215 | 0.206 | **0.185** | 0.210 | 0.216 | 0.308 | 0.256 | 0.234 |
| | 336 | 0.222 | 0.241 | 0.235 | 0.224 | **0.204** | 0.218 | 0.232 | 0.283 | 0.252 | 0.248 |
| | 720 | 0.260 | 0.283 | 0.286 | 0.268 | **0.245** | 0.263 | 0.285 | 0.335 | 0.261 | 0.282 |
| | Avg | 0.219 | 0.250 | 0.239 | 0.223 | **0.202** | 0.223 | 0.234 | 0.303 | 0.257 | 0.250 |

Continued on next page

Table 11 (continued)

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *MAE* | | | | | | | | | | | |
| ETTh1 | 96 | **0.402** | 0.421 | 0.506 | 0.407 | 0.411 | 0.477 | 0.549 | 0.609 | 0.558 | 0.420 |
| | 192 | 0.444 | 0.449 | 0.582 | 0.447 | **0.443** | 0.509 | 0.567 | 0.608 | 0.455 | 0.452 |
| | 336 | **0.463** | 0.466 | 0.804 | 0.465 | 0.477 | 0.500 | 0.574 | 0.639 | 0.474 | 0.484 |
| | Avg | **0.436** | 0.445 | 0.631 | 0.440 | 0.444 | 0.495 | 0.563 | 0.619 | 0.496 | 0.452 |
| ETTh2 | 96 | **0.351** | 0.354 | 0.383 | 0.356 | 0.354 | 0.396 | 0.396 | 0.412 | 0.448 | 0.423 |
| | 192 | **0.402** | 0.407 | 0.448 | 0.407 | 0.413 | 0.444 | 0.440 | 0.454 | 0.406 | 0.485 |
| | 336 | **0.434** | 0.442 | 0.660 | 0.436 | 0.440 | 0.474 | 0.435 | 0.479 | 0.465 | 0.551 |
| | Avg | **0.396** | 0.401 | 0.497 | 0.400 | 0.402 | 0.438 | 0.424 | 0.448 | 0.440 | 0.486 |
| ETTm1 | 96 | **0.379** | 0.381 | 0.449 | 0.394 | 0.380 | 0.416 | 0.396 | 0.402 | 0.501 | 0.438 |
| | 192 | **0.392** | 0.395 | 0.465 | 0.406 | 0.402 | 0.442 | 0.396 | 0.430 | 0.474 | 0.455 |
| | 336 | **0.408** | 0.410 | 0.482 | 0.421 | 0.427 | 0.463 | 0.409 | 0.490 | 0.464 | 0.469 |
| | 720 | **0.440** | 0.442 | 0.590 | 0.455 | 0.467 | 0.498 | 0.445 | 0.505 | 0.540 | 0.500 |
| | Avg | **0.405** | 0.407 | 0.497 | 0.419 | 0.419 | 0.455 | 0.412 | 0.457 | 0.495 | 0.466 |
| ETTm2 | 96 | **0.262** | 0.270 | 0.286 | 0.293 | 0.272 | 0.285 | 0.263 | 0.279 | 0.277 | 0.311 |
| | 192 | **0.303** | 0.306 | 0.332 | 0.322 | 0.319 | 0.323 | 0.305 | 0.353 | 0.309 | 0.378 |
| | 336 | **0.349** | 0.351 | 0.356 | 0.356 | 0.366 | 0.360 | 0.354 | 0.368 | 0.352 | 0.411 |
| | 720 | **0.400** | 0.401 | 0.448 | 0.409 | 0.426 | 0.413 | 0.403 | 0.441 | 0.413 | 0.477 |
| | Avg | **0.329** | 0.332 | 0.356 | 0.345 | 0.346 | 0.345 | 0.331 | 0.360 | 0.338 | 0.394 |
| Weather | 96 | 0.223 | 0.233 | 0.252 | 0.235 | **0.218** | 0.237 | 0.236 | 0.254 | 0.232 | 0.304 |
| | 192 | 0.263 | 0.267 | 0.281 | 0.272 | **0.258** | 0.273 | 0.272 | 0.310 | 0.279 | 0.336 |
| | 336 | 0.301 | 0.304 | 0.322 | 0.305 | **0.299** | 0.310 | 0.307 | 0.320 | 0.307 | 0.365 |
| | 720 | **0.347** | 0.352 | 0.370 | 0.354 | 0.351 | 0.357 | 0.353 | 0.376 | 0.353 | 0.408 |
| | Avg | 0.284 | 0.289 | 0.306 | 0.292 | **0.281** | 0.294 | 0.292 | 0.315 | 0.293 | 0.353 |
| ILI | 24 | 1.108 | 1.455 | 1.537 | 0.944 | **0.821** | 1.549 | 1.150 | 1.325 | 2.202 | 2.130 |
| | Avg | 1.108 | 1.455 | 1.537 | 0.944 | **0.821** | 1.549 | 1.150 | 1.325 | 2.202 | 2.130 |
| Exchange | 96 | **0.198** | 0.208 | 0.251 | 0.203 | 0.209 | 0.254 | 0.261 | 0.265 | 0.287 | 0.272 |
| | 192 | **0.296** | 0.305 | 0.398 | 0.301 | 0.303 | 0.345 | 0.354 | 0.342 | 0.304 | 0.355 |
| | Avg | **0.247** | 0.257 | 0.325 | 0.252 | 0.256 | 0.300 | 0.308 | 0.304 | 0.296 | 0.314 |
| Electricity | 96 | 0.286 | 0.327 | 0.308 | 0.288 | **0.257** | 0.298 | 0.299 | 0.365 | 0.343 | 0.328 |
| | 192 | 0.296 | 0.328 | 0.305 | 0.297 | **0.270** | 0.299 | 0.301 | 0.386 | 0.345 | 0.330 |
| | 336 | 0.300 | 0.331 | 0.325 | 0.304 | **0.288** | 0.306 | 0.317 | 0.367 | 0.342 | 0.344 |
| | 720 | 0.360 | 0.362 | 0.365 | 0.363 | **0.322** | 0.362 | 0.361 | 0.403 | 0.364 | 0.372 |
| | Avg | 0.311 | 0.337 | 0.326 | 0.313 | **0.284** | 0.316 | 0.320 | 0.380 | 0.349 | 0.344 |

Table 12: Few-shot forecasting with 10% training data: MSE and MAE.

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *MSE* | | | | | | | | | | | |
| ETTh1 | 96 | **0.397** | 0.399 | 0.530 | 0.408 | 0.415 | 0.442 | 0.563 | 0.618 | 0.550 | 0.414 |
| | 192 | **0.441** | 0.443 | 0.546 | 0.449 | 0.463 | 0.497 | 0.611 | 0.644 | 0.458 | 0.450 |
| | 336 | **0.484** | 0.486 | 0.602 | 0.498 | 0.524 | 0.510 | 0.626 | 0.726 | 0.490 | 0.493 |
| | 720 | 0.707 | 0.486 | 0.863 | 0.510 | 0.525 | 0.532 | 0.613 | 0.720 | **0.481** | 0.528 |
| | Avg | 0.507 | **0.454** | 0.635 | 0.466 | 0.482 | 0.495 | 0.603 | 0.677 | 0.495 | 0.471 |
| ETTh2 | 96 | **0.294** | 0.296 | 0.306 | 0.327 | 0.309 | 0.342 | 0.348 | 0.374 | 0.387 | 0.349 |
| | 192 | **0.385** | 0.390 | 0.402 | 0.399 | 0.406 | 0.431 | 0.430 | 0.451 | 0.386 | 0.456 |
| | 336 | **0.422** | 0.426 | 0.513 | 0.436 | 0.439 | 0.469 | 0.427 | 0.492 | 0.446 | 0.605 |
| | 720 | 0.457 | 0.457 | 0.879 | 0.466 | **0.455** | 0.471 | 0.462 | 0.513 | 0.463 | 0.759 |
| | Avg | **0.390** | 0.392 | 0.525 | 0.407 | 0.403 | 0.428 | 0.417 | 0.458 | 0.421 | 0.542 |
| ETTm1 | 96 | 0.344 | 0.346 | 0.520 | 0.379 | **0.343** | 0.376 | 0.345 | 0.350 | 0.552 | 0.382 |
| | 192 | **0.370** | 0.375 | 0.528 | 0.407 | 0.383 | 0.423 | 0.386 | 0.407 | 0.463 | 0.414 |
| | 336 | **0.399** | 0.405 | 0.543 | 0.430 | 0.414 | 0.459 | 0.403 | 0.481 | 0.455 | 0.443 |
| | 720 | **0.467** | 0.473 | 0.594 | 0.484 | 0.486 | 0.530 | 0.486 | 0.517 | 0.556 | 0.501 |
| | Avg | **0.395** | 0.400 | 0.546 | 0.425 | 0.406 | 0.447 | 0.405 | 0.439 | 0.507 | 0.435 |
| ETTm2 | 96 | **0.175** | 0.181 | 0.189 | 0.202 | 0.186 | 0.191 | 0.176 | 0.187 | 0.189 | 0.193 |
| | 192 | **0.240** | 0.244 | 0.262 | 0.261 | 0.259 | 0.256 | 0.243 | 0.290 | 0.247 | 0.282 |
| | 336 | **0.301** | 0.303 | 0.321 | 0.318 | 0.329 | 0.318 | 0.303 | 0.314 | 0.313 | 0.370 |
| | 720 | **0.391** | 0.407 | 0.431 | 0.416 | 0.434 | 0.417 | 0.401 | 0.443 | 0.421 | 0.488 |
| | Avg | **0.277** | 0.284 | 0.301 | 0.299 | 0.302 | 0.296 | 0.281 | 0.309 | 0.293 | 0.333 |
| Weather | 96 | **0.160** | 0.189 | 0.183 | 0.186 | 0.174 | 0.192 | 0.198 | 0.196 | 0.168 | 0.217 |
| | 192 | **0.219** | 0.235 | 0.238 | 0.231 | 0.222 | 0.240 | 0.243 | 0.246 | 0.233 | 0.255 |
| | 336 | **0.271** | 0.287 | 0.292 | 0.283 | 0.279 | 0.291 | 0.293 | 0.290 | 0.284 | 0.300 |
| | 720 | **0.349** | 0.360 | 0.366 | 0.355 | 0.355 | 0.365 | 0.362 | 0.371 | 0.355 | 0.360 |
| | Avg | **0.250** | 0.268 | 0.270 | 0.264 | 0.258 | 0.272 | 0.274 | 0.276 | 0.260 | 0.283 |
| ILI | 24 | 2.631 | 3.621 | 3.472 | 2.046 | **1.758** | 4.090 | 1.959 | 2.770 | 7.921 | 2.721 |
| | 36 | 2.794 | 3.345 | 4.973 | **1.896** | 1.910 | 4.255 | 2.001 | 3.124 | 7.756 | 2.789 |
| | 48 | 2.673 | 3.657 | 5.286 | **1.821** | 1.864 | 4.259 | 2.231 | 3.225 | 7.614 | 2.681 |
| | Avg | 2.699 | 3.541 | 4.577 | 1.921 | **1.844** | 4.201 | 2.064 | 3.040 | 7.764 | 2.730 |
| Exchange | 96 | **0.082** | 0.088 | 0.107 | 0.083 | 0.087 | 0.112 | 0.119 | 0.122 | 0.138 | 0.109 |
| | 192 | **0.170** | 0.176 | 0.215 | 0.173 | 0.182 | 0.213 | 0.224 | 0.220 | 0.179 | 0.189 |
| | 336 | **0.309** | 0.349 | 0.326 | 0.344 | 0.330 | 0.379 | 0.372 | 0.391 | 0.344 | 0.387 |
| | Avg | **0.187** | 0.204 | 0.216 | 0.200 | 0.200 | 0.235 | 0.238 | 0.244 | 0.220 | 0.228 |

Continued on next page

Table 12 (continued)

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Electricity | 96 | 0.190 | 0.222 | 0.220 | 0.197 | **0.173** | 0.193 | 0.196 | 0.261 | 0.213 | 0.212 |
| | 192 | 0.201 | 0.222 | 0.216 | 0.203 | **0.185** | 0.204 | 0.200 | 0.260 | 0.214 | 0.212 |
| | 336 | 0.210 | 0.228 | 0.231 | 0.211 | **0.203** | 0.213 | 0.215 | 0.268 | 0.218 | 0.224 |
| | 720 | **0.245** | 0.271 | 0.280 | 0.252 | **0.245** | 0.260 | 0.258 | 0.303 | 0.247 | 0.260 |
| | Avg | 0.212 | 0.236 | 0.237 | 0.216 | **0.201** | 0.218 | 0.217 | 0.273 | 0.223 | 0.227 |
| *MAE* | | | | | | | | | | | |
| ETTh1 | 96 | **0.408** | 0.411 | 0.480 | 0.410 | **0.408** | 0.442 | 0.502 | 0.534 | 0.492 | 0.416 |
| | 192 | **0.429** | 0.433 | 0.494 | 0.435 | 0.441 | 0.473 | 0.525 | 0.545 | 0.447 | 0.443 |
| | 336 | **0.452** | 0.454 | 0.521 | 0.453 | 0.474 | 0.474 | 0.536 | 0.586 | 0.462 | 0.467 |
| | 720 | 0.595 | 0.480 | 0.654 | 0.485 | 0.489 | 0.509 | 0.546 | 0.604 | **0.471** | 0.519 |
| | Avg | 0.471 | **0.445** | 0.537 | 0.446 | 0.453 | 0.475 | 0.527 | 0.567 | 0.468 | 0.461 |
| ETTh2 | 96 | **0.337** | 0.343 | 0.352 | 0.354 | 0.350 | 0.382 | 0.385 | 0.404 | 0.410 | 0.396 |
| | 192 | **0.399** | 0.402 | 0.408 | 0.407 | 0.407 | 0.431 | 0.430 | 0.448 | 0.402 | 0.457 |
| | 336 | **0.437** | 0.439 | 0.471 | 0.443 | 0.439 | 0.463 | 0.439 | 0.482 | 0.450 | 0.544 |
| | 720 | 0.463 | 0.467 | 0.654 | 0.466 | **0.456** | 0.472 | 0.464 | 0.493 | 0.474 | 0.618 |
| | Avg | **0.409** | 0.413 | 0.471 | 0.418 | 0.413 | 0.437 | 0.430 | 0.457 | 0.434 | 0.504 |
| ETTm1 | 96 | **0.374** | 0.375 | 0.454 | 0.389 | 0.375 | 0.392 | 0.381 | 0.385 | 0.482 | 0.399 |
| | 192 | **0.383** | 0.386 | 0.452 | 0.401 | 0.396 | 0.416 | 0.402 | 0.419 | 0.438 | 0.415 |
| | 336 | **0.408** | 0.409 | 0.480 | 0.418 | 0.418 | 0.439 | 0.414 | 0.457 | 0.437 | 0.435 |
| | 720 | **0.441** | 0.450 | 0.495 | 0.447 | 0.461 | 0.478 | 0.468 | 0.475 | 0.490 | 0.471 |
| | Avg | **0.402** | 0.405 | 0.470 | 0.414 | 0.413 | 0.431 | 0.416 | 0.434 | 0.462 | 0.430 |
| ETTm2 | 96 | **0.259** | 0.266 | 0.271 | 0.290 | 0.267 | 0.277 | 0.260 | 0.270 | 0.274 | 0.289 |
| | 192 | **0.300** | 0.307 | 0.316 | 0.322 | 0.316 | 0.316 | 0.303 | 0.346 | 0.306 | 0.358 |
| | 336 | **0.340** | 0.341 | 0.351 | 0.355 | 0.360 | 0.353 | 0.342 | 0.349 | 0.349 | 0.412 |
| | 720 | **0.395** | 0.401 | 0.415 | 0.409 | 0.421 | 0.407 | 0.397 | 0.427 | 0.408 | 0.485 |
| | Avg | **0.324** | 0.329 | 0.338 | 0.344 | 0.341 | 0.338 | 0.326 | 0.348 | 0.334 | 0.386 |
| Weather | 96 | **0.210** | 0.229 | 0.230 | 0.226 | 0.214 | 0.231 | 0.237 | 0.246 | 0.216 | 0.283 |
| | 192 | 0.260 | 0.266 | 0.275 | 0.263 | **0.255** | 0.270 | 0.272 | 0.284 | 0.270 | 0.317 |
| | 336 | **0.297** | 0.302 | 0.313 | 0.299 | **0.297** | 0.307 | 0.308 | 0.311 | 0.303 | 0.350 |
| | 720 | **0.340** | 0.349 | 0.360 | 0.347 | 0.348 | 0.355 | 0.353 | 0.364 | 0.348 | 0.395 |
| | Avg | **0.277** | 0.287 | 0.295 | 0.284 | 0.278 | 0.291 | 0.293 | 0.301 | 0.284 | 0.336 |
| ILI | 24 | 1.079 | 1.416 | 1.405 | 0.933 | **0.804** | 1.478 | 0.941 | 1.177 | 2.115 | 1.175 |
| | 36 | 1.052 | 1.349 | 1.683 | 0.896 | **0.851** | 1.504 | 1.044 | 1.245 | 2.107 | 1.199 |
| | 48 | 1.072 | 1.387 | 1.692 | 0.881 | **0.860** | 1.498 | 1.109 | 1.236 | 2.071 | 1.177 |
| | Avg | 1.068 | 1.384 | 1.593 | 0.903 | **0.838** | 1.493 | 1.031 | 1.219 | 2.098 | 1.184 |
| Exchange | 96 | **0.198** | 0.207 | 0.229 | 0.201 | 0.207 | 0.238 | 0.247 | 0.248 | 0.267 | 0.246 |
| | 192 | **0.288** | 0.298 | 0.332 | 0.295 | 0.304 | 0.332 | 0.343 | 0.335 | 0.301 | 0.328 |
| | 336 | **0.401** | 0.428 | 0.419 | 0.423 | 0.415 | 0.452 | 0.446 | 0.459 | 0.426 | 0.479 |
| | Avg | **0.296** | 0.311 | 0.327 | 0.306 | 0.309 | 0.341 | 0.345 | 0.347 | 0.331 | 0.351 |

Continued on next page

Table 12 (continued)

| Data | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Electricity | 96 | 0.291 | 0.311 | 0.309 | 0.299 | **0.257** | 0.293 | 0.296 | 0.344 | 0.299 | 0.302 |
| | 192 | 0.292 | 0.313 | 0.306 | 0.295 | **0.269** | 0.302 | 0.294 | 0.344 | 0.302 | 0.305 |
| | 336 | 0.296 | 0.319 | 0.320 | 0.303 | **0.288** | 0.310 | 0.299 | 0.351 | 0.305 | 0.319 |
| | 720 | **0.310** | 0.350 | 0.356 | 0.317 | 0.322 | 0.347 | 0.334 | 0.376 | 0.314 | 0.350 |
| | Avg | 0.297 | 0.323 | 0.323 | 0.304 | **0.284** | 0.313 | 0.306 | 0.354 | 0.305 | 0.319 |

**Results.** At 5% training data, MM-TS leads six of eight datasets in average MSE and five in average MAE; at 10%, it leads five in both metrics (Tables 11 and 12). ETTm1 provides a matched-horizon comparison: its average MSE decreases from 0.400 to 0.395 and MAE from 0.405 to 0.402 as the training fraction increases.

**Analysis.** The few-shot results examine whether input-derived channel structure remains useful when supervised examples are reduced. The ETTm1 comparison is directly comparable across fractions because both settings contain all four horizons; its improvement is consistent with observed-input structure complementing additional supervision. TimesFM retains the leading MSE on ILI and Electricity at both fractions, highlighting task-dependent strengths among the approaches. For ETTh1, ETTh2, ILI, and Exchange, the reported horizon sets differ, so comparisons across training fractions should use shared horizons. At the shared ETTh1 horizons of 96, 192, and 336 steps, increasing the training fraction from 5% to 10% reduces MM-TS MSE from 0.407, 0.456, and 0.497 to 0.397, 0.441, and 0.484. This matched-horizon comparison separates the benefit of additional supervision from changes in the set of reported horizons. These details extend the limited-supervision assessment in Section 4.2.

### D.2 Zero-shot Forecasting: Direct ETT Transfer

Each model is trained on the named source ETT dataset and applied to the target ETT dataset without updating its weights on the target. The following table reports all six directions and all four prediction lengths. ETTm1→ETTm2, ETTh1→ETTh2, ETTh1→ETTm2, ETTh2→ETTm2, and ETTm2→ETTh2 attain the best average MSE with MM-TS.

Table 13: Zero-shot forecasting by direct ETT transfer: MSE and MAE.

| Pair | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *MSE* | | | | | | | | | | | |
| ETTh1→ETTh2 | 96 | **0.295** | 0.297 | 0.299 | 0.313 | 0.301 | 0.296 | 0.297 | 0.333 | 0.296 | 0.319 |
| | 192 | **0.372** | 0.378 | 0.390 | 0.383 | 0.390 | 0.379 | 0.381 | 0.409 | 0.378 | 0.413 |
| | 336 | **0.409** | 0.415 | 0.422 | 0.418 | 0.422 | 0.424 | 0.413 | 0.477 | 0.426 | 0.571 |
| | 720 | **0.406** | 0.417 | 0.427 | 0.419 | 0.423 | 0.429 | 0.418 | 0.475 | 0.414 | 0.734 |
| | Avg | **0.371** | 0.377 | 0.385 | 0.383 | 0.384 | 0.382 | 0.377 | 0.424 | 0.379 | 0.509 |
| ETTh1→ETTm2 | 96 | **0.211** | 0.216 | 0.215 | 0.248 | 0.217 | 0.238 | 0.212 | 0.244 | 0.213 | 0.246 |
| | 192 | **0.270** | 0.278 | 0.281 | 0.303 | 0.276 | 0.299 | 0.275 | 0.291 | 0.275 | 0.323 |
| | 336 | **0.325** | 0.339 | 0.336 | 0.356 | 0.338 | 0.342 | 0.333 | 0.361 | 0.331 | 0.491 |
| | 720 | **0.422** | 0.434 | 0.434 | 0.452 | 0.435 | 0.442 | 0.429 | 0.460 | 0.429 | 0.720 |
| | Avg | **0.307** | 0.317 | 0.317 | 0.340 | 0.316 | 0.330 | 0.312 | 0.339 | 0.312 | 0.445 |

Continued on next page

Table 13 (continued)

| Pair | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ETTh2→ETTm2 | 96 | **0.205** | 0.214 | 0.223 | 0.226 | 0.227 | 0.237 | 0.213 | 0.253 | 0.213 | 0.258 |
| | 192 | **0.275** | 0.282 | 0.284 | 0.304 | 0.305 | 0.292 | 0.278 | 0.341 | 0.277 | 0.342 |
| | 336 | **0.336** | 0.347 | 0.340 | 0.360 | 0.370 | 0.349 | 0.343 | 0.380 | 0.346 | 0.374 |
| | 720 | 0.438 | 0.444 | **0.435** | 0.455 | 0.455 | 0.445 | 0.450 | 0.488 | 0.449 | 0.760 |
| | Avg | **0.314** | 0.322 | 0.321 | 0.336 | 0.339 | 0.331 | 0.321 | 0.366 | 0.321 | 0.434 |
| ETTm1→ETTh2 | 96 | **0.355** | 0.358 | 0.369 | 0.359 | 0.361 | 0.359 | 0.371 | 0.418 | 0.362 | 0.362 |
| | 192 | 0.621 | 0.435 | 0.444 | **0.415** | 0.455 | 0.443 | 0.446 | 0.632 | 0.438 | 0.456 |
| | 336 | 0.656 | 0.464 | 0.480 | 0.449 | 0.485 | 0.490 | **0.446** | 0.550 | 0.471 | 0.514 |
| | 720 | 0.751 | 0.459 | 0.469 | **0.443** | 0.466 | 0.486 | 0.480 | 0.591 | 0.458 | 0.561 |
| | Avg | 0.596 | 0.429 | 0.441 | **0.417** | 0.442 | 0.445 | 0.436 | 0.548 | 0.432 | 0.473 |
| ETTm1→ETTm2 | 96 | **0.189** | 0.194 | 0.202 | 0.200 | 0.199 | 0.202 | 0.194 | 0.230 | 0.192 | 0.219 |
| | 192 | **0.245** | 0.253 | 0.262 | 0.263 | 0.261 | 0.260 | 0.256 | 0.345 | 0.256 | 0.292 |
| | 336 | **0.305** | 0.313 | 0.322 | 0.323 | 0.319 | 0.322 | 0.316 | 0.361 | 0.315 | 0.370 |
| | 720 | **0.411** | 0.412 | 0.417 | 0.427 | 0.422 | 0.422 | 0.414 | 0.488 | 0.416 | 0.493 |
| | Avg | **0.288** | 0.293 | 0.301 | 0.303 | 0.300 | 0.302 | 0.295 | 0.356 | 0.295 | 0.344 |
| ETTm2→ETTh2 | 96 | **0.319** | 0.320 | 0.335 | 0.361 | 0.321 | 0.353 | 0.330 | 0.475 | 0.326 | 0.369 |
| | 192 | **0.399** | 0.403 | 0.411 | 0.409 | 0.426 | 0.447 | 0.419 | 0.470 | 0.409 | 0.428 |
| | 336 | **0.452** | 0.458 | 0.463 | 0.465 | 0.473 | 0.517 | 0.463 | 0.546 | 0.471 | 0.478 |
| | 720 | **0.434** | 0.457 | 0.458 | 0.439 | 0.480 | 0.528 | 0.445 | 0.560 | 0.485 | 0.489 |
| | Avg | **0.401** | 0.410 | 0.417 | 0.419 | 0.425 | 0.461 | 0.414 | 0.513 | 0.423 | 0.441 |
| *MAE* | | | | | | | | | | | |
| ETTh1→ETTh2 | 96 | **0.341** | 0.342 | 0.344 | 0.354 | 0.345 | 0.344 | 0.343 | 0.373 | 0.342 | 0.376 |
| | 192 | **0.385** | 0.391 | 0.400 | 0.396 | 0.398 | 0.395 | 0.394 | 0.412 | 0.391 | 0.433 |
| | 336 | **0.424** | 0.425 | 0.428 | 0.430 | 0.429 | 0.431 | 0.429 | 0.461 | 0.435 | 0.521 |
| | 720 | **0.422** | 0.437 | 0.444 | 0.442 | 0.443 | 0.444 | 0.440 | 0.469 | 0.435 | 0.613 |
| | Avg | **0.393** | 0.399 | 0.404 | 0.406 | 0.404 | 0.404 | 0.402 | 0.429 | 0.401 | 0.486 |
| ETTh1→ETTm2 | 96 | **0.297** | 0.302 | 0.301 | 0.309 | **0.297** | 0.321 | 0.299 | 0.321 | 0.299 | 0.350 |
| | 192 | **0.327** | 0.337 | 0.341 | 0.332 | 0.333 | 0.353 | 0.333 | 0.345 | 0.334 | 0.400 |
| | 336 | **0.357** | 0.374 | 0.370 | 0.365 | 0.370 | 0.375 | 0.367 | 0.383 | 0.366 | 0.494 |
| | 720 | **0.408** | 0.424 | 0.424 | 0.421 | 0.422 | 0.428 | 0.418 | 0.433 | 0.418 | 0.606 |
| | Avg | **0.347** | 0.359 | 0.359 | 0.357 | 0.356 | 0.369 | 0.354 | 0.371 | 0.354 | 0.463 |
| ETTh2→ETTm2 | 96 | **0.288** | 0.298 | 0.306 | 0.290 | 0.305 | 0.321 | 0.299 | 0.330 | 0.300 | 0.354 |
| | 192 | **0.326** | 0.340 | 0.341 | 0.331 | 0.352 | 0.350 | 0.335 | 0.381 | 0.338 | 0.410 |
| | 336 | **0.362** | 0.378 | 0.373 | 0.365 | 0.390 | 0.381 | 0.369 | 0.401 | 0.373 | 0.418 |
| | 720 | **0.420** | 0.429 | 0.423 | 0.430 | 0.434 | 0.430 | 0.440 | 0.460 | 0.439 | 0.626 |
| | Avg | **0.349** | 0.361 | 0.361 | 0.354 | 0.370 | 0.371 | 0.361 | 0.393 | 0.363 | 0.452 |

Continued on next page

Table 13 (continued)

| Pair | $H$ | MM-TS | Time-VLM | Time-LLM | VTS | TFM | iTr | PTST | TNet | TMix | DLin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ETTm1→ETTh2 | 96 | 0.383 | 0.385 | 0.394 | 0.395 | **0.381** | 0.396 | 0.396 | 0.440 | 0.397 | 0.410 |
| | 192 | 0.534 | 0.427 | 0.432 | **0.413** | 0.431 | 0.436 | 0.433 | 0.539 | 0.426 | 0.460 |
| | 336 | 0.563 | 0.455 | 0.463 | **0.444** | 0.459 | 0.469 | 0.446 | 0.508 | 0.456 | 0.501 |
| | 720 | 0.605 | 0.461 | 0.465 | **0.454** | 0.461 | 0.476 | 0.473 | 0.538 | 0.460 | 0.537 |
| | Avg | 0.521 | 0.432 | 0.439 | **0.427** | 0.433 | 0.444 | 0.437 | 0.506 | 0.435 | 0.477 |
| ETTm1→ETTm2 | 96 | **0.267** | 0.270 | 0.272 | 0.284 | 0.274 | 0.279 | 0.270 | 0.307 | 0.269 | 0.311 |
| | 192 | **0.299** | 0.306 | 0.309 | 0.320 | 0.312 | 0.314 | 0.308 | 0.376 | 0.305 | 0.362 |
| | 336 | **0.341** | 0.344 | 0.347 | 0.355 | 0.349 | 0.352 | 0.347 | 0.383 | 0.343 | 0.416 |
| | 720 | **0.398** | 0.400 | 0.401 | 0.414 | 0.407 | 0.407 | 0.404 | 0.453 | 0.407 | 0.484 |
| | Avg | **0.326** | 0.330 | 0.332 | 0.343 | 0.335 | 0.338 | 0.332 | 0.380 | 0.331 | 0.393 |
| ETTm2→ETTh2 | 96 | **0.359** | 0.364 | 0.379 | 0.377 | **0.359** | 0.395 | 0.367 | 0.459 | 0.371 | 0.423 |
| | 192 | **0.402** | 0.412 | 0.420 | 0.418 | 0.418 | 0.445 | 0.418 | 0.458 | 0.419 | 0.442 |
| | 336 | **0.455** | 0.460 | 0.465 | 0.462 | **0.455** | 0.491 | 0.474 | 0.500 | 0.470 | 0.477 |
| | 720 | **0.448** | 0.449 | 0.466 | 0.450 | 0.476 | 0.505 | 0.454 | 0.511 | 0.479 | 0.493 |
| | Avg | **0.416** | 0.421 | 0.433 | 0.427 | 0.427 | 0.459 | 0.428 | 0.482 | 0.435 | 0.459 |

**Results.** Table 13 resolves the averages in Section 4.2 by transfer direction and forecast length. From ETTh1 to ETTh2, MM-TS achieves average MSE/MAE of 0.371/0.393, compared with the best baseline scores of 0.377/0.399. From ETTh1 to ETTm2, the corresponding scores are 0.307/0.347 versus 0.312/0.354. These comparisons show that the gains in the selected same-resolution and cross-resolution transfers appear under both metrics.

**Analysis.** Direct transfer separates fixed model parameters from input-dependent structural guidance. Although parameters are learned on the source dataset, MM-TS recomputes channel relations from each target window, allowing the structural input to reflect target observations. The leading average MSE in five directions supports this design within the shared ETT variable layout. The ETTm1-to-ETTh2 direction provides a contrasting case, with MSE of 0.596 for MM-TS and 0.417 for VisionTS, indicating that the source-target pairing remains consequential. Together, these results characterize transfer across domains and sampling resolutions with corresponding variables, and motivate extending the framework to learned channel correspondence across heterogeneous variable sets.

## E Ablations and Design Choices

Table 14 reports $H \in \{96, 192, 336, 720\}$; Tables 15–20 report $H \in \{96, 720\}$. All component and design-control results in this appendix average five runs with seeds 2022–2026. Each Avg uses only its table's displayed horizons, so averages from the two-horizon controls can differ from the four-horizon supervised results. Layer indices start at zero; LoRA layers remain 0, 1, 26, and 27.

### E.1 Component Ablations

Table 14: Component ablations: MSE and MAE.

| Data | $H$ | Metric | MM-TS (full) | Without visual tokens | Without text | Without attention bias | Without visual tokens or bias | Permuted channel bias | Previous window bias |
|---|---|---|---|---|---|---|---|---|---|
| ETTh1 | 96 | MSE | **0.379** | 0.387 | 0.382 | 0.389 | 0.397 | 0.392 | 0.387 |
| | | MAE | **0.392** | 0.397 | 0.395 | 0.398 | 0.403 | 0.400 | 0.397 |
| | 192 | MSE | **0.425** | 0.434 | 0.430 | 0.436 | 0.446 | 0.440 | 0.435 |
| | | MAE | **0.422** | 0.427 | 0.425 | 0.429 | 0.434 | 0.431 | 0.428 |
| | 336 | MSE | **0.462** | 0.472 | 0.469 | 0.474 | 0.484 | 0.479 | 0.472 |
| | | MAE | **0.441** | 0.448 | 0.445 | 0.449 | 0.454 | 0.450 | 0.448 |
| | 720 | MSE | **0.469** | 0.480 | 0.475 | 0.484 | 0.494 | 0.490 | 0.482 |
| | | MAE | **0.462** | 0.468 | 0.466 | 0.471 | 0.476 | 0.473 | 0.469 |
| | *Avg* | MSE | **0.434** | 0.443 | 0.439 | 0.446 | 0.455 | 0.451 | 0.445 |
| | | MAE | **0.429** | 0.435 | 0.433 | 0.437 | 0.442 | 0.439 | 0.436 |
| ETTh2 | 96 | MSE | **0.283** | 0.289 | 0.287 | 0.292 | 0.298 | 0.295 | 0.289 |
| | | MAE | **0.337** | 0.341 | 0.339 | 0.342 | 0.346 | 0.344 | 0.342 |
| | 192 | MSE | **0.362** | 0.371 | 0.367 | 0.375 | 0.383 | 0.379 | 0.373 |
| | | MAE | **0.388** | 0.394 | 0.391 | 0.395 | 0.401 | 0.398 | 0.394 |
| | 336 | MSE | **0.407** | 0.418 | 0.413 | 0.422 | 0.431 | 0.426 | 0.420 |
| | | MAE | **0.424** | 0.431 | 0.429 | 0.433 | 0.439 | 0.435 | 0.433 |
| | 720 | MSE | **0.421** | 0.432 | 0.426 | 0.436 | 0.443 | 0.438 | 0.432 |
| | | MAE | **0.441** | 0.447 | 0.444 | 0.450 | 0.454 | 0.451 | 0.446 |
| | *Avg* | MSE | **0.368** | 0.378 | 0.373 | 0.381 | 0.389 | 0.385 | 0.379 |
| | | MAE | **0.398** | 0.403 | 0.401 | 0.405 | 0.410 | 0.407 | 0.404 |
| Weather | 96 | MSE | **0.160** | 0.166 | 0.163 | 0.166 | 0.172 | 0.168 | 0.165 |
| | | MAE | **0.206** | 0.211 | 0.208 | 0.210 | 0.214 | 0.212 | 0.210 |
| | 192 | MSE | **0.201** | 0.209 | 0.205 | 0.210 | 0.216 | 0.212 | 0.208 |
| | | MAE | **0.248** | 0.253 | 0.251 | 0.252 | 0.257 | 0.255 | 0.253 |
| | 336 | MSE | **0.260** | 0.269 | 0.264 | 0.271 | 0.279 | 0.275 | 0.270 |
| | | MAE | **0.287** | 0.294 | 0.290 | 0.294 | 0.299 | 0.297 | 0.293 |
| | 720 | MSE | **0.340** | 0.351 | 0.345 | 0.355 | 0.364 | 0.358 | 0.354 |
| | | MAE | **0.341** | 0.349 | 0.345 | 0.350 | 0.356 | 0.354 | 0.350 |
| | *Avg* | MSE | **0.240** | 0.249 | 0.245 | 0.250 | 0.258 | 0.253 | 0.249 |
| | | MAE | **0.271** | 0.277 | 0.274 | 0.277 | 0.282 | 0.280 | 0.276 |
| Electricity | 96 | MSE | **0.163** | 0.167 | 0.166 | 0.169 | 0.174 | 0.171 | 0.168 |
| | | MAE | **0.254** | 0.258 | 0.257 | 0.260 | 0.262 | 0.261 | 0.258 |
| | 192 | MSE | **0.175** | 0.180 | 0.178 | 0.182 | 0.188 | 0.185 | 0.181 |
| | | MAE | **0.262** | 0.266 | 0.264 | 0.268 | 0.271 | 0.269 | 0.266 |
| | 336 | MSE | **0.190** | 0.196 | 0.193 | 0.200 | 0.206 | 0.203 | 0.198 |
| | | MAE | **0.286** | 0.291 | 0.289 | 0.293 | 0.297 | 0.296 | 0.292 |
| | 720 | MSE | **0.230** | 0.237 | 0.235 | 0.242 | 0.249 | 0.246 | 0.241 |
| | | MAE | **0.320** | 0.326 | 0.322 | 0.328 | 0.333 | 0.330 | 0.328 |
| | *Avg* | MSE | **0.190** | 0.195 | 0.193 | 0.199 | 0.204 | 0.201 | 0.197 |
| | | MAE | **0.281** | 0.285 | 0.283 | 0.287 | 0.291 | 0.289 | 0.286 |

**Settings.** Table 14 varies the input components and the source or channel alignment of the attention bias, extending Table 3 to four horizons and both metrics.

**Results.** On ETTh2, joint removal of visual tokens and attention bias raises MSE from 0.283 to 0.298 at $H = 96$ and from 0.421 to 0.443 at $H = 720$. MAE rises from 0.337 to 0.346 and from 0.441 to 0.454, respectively.

**Analysis.** The ablations distinguish access to relation information from its use during token interaction. Without visual tokens, the bias still supplies channel guidance; without bias, the model still receives the visual relation representation. Their higher errors support complementary access through both paths. On ETTh2, joint removal increases MSE by 0.015 at 96 steps and 0.022 at 720 steps, showing that the contribution persists at both evaluated endpoints. Channel permutation changes which token pairs receive the structural guidance, and its poorer performance than no bias emphasizes the importance of correct correspondence. The previous-window comparison further supports updating that guidance from the current observation. These controls substantiate the three-part interpretation in Section 4.3.

### E.2 Backbone Comparison

Table 15: Backbone capacity and visual pretraining: MSE and MAE.

| Data | $H$ | Metric | Qwen3-VL 2B (full) | 2B, reset visual projector | Qwen3-VL 4B | DINOv2 + decoder | iTr | iTr + relation bias |
|---|---|---|---|---|---|---|---|---|
| ETTh1 | 96 | MSE | **0.379** | 0.388 | 0.380 | 0.385 | 0.386 | 0.383 |
| | | MAE | **0.392** | 0.400 | 0.393 | 0.402 | 0.405 | 0.399 |
| | 720 | MSE | **0.469** | 0.488 | 0.472 | 0.495 | 0.503 | 0.486 |
| | | MAE | **0.462** | 0.476 | 0.464 | 0.481 | 0.491 | 0.478 |
| | *Avg* | MSE | **0.424** | 0.438 | 0.426 | 0.440 | 0.445 | 0.435 |
| | | MAE | **0.427** | 0.438 | 0.429 | 0.442 | 0.448 | 0.439 |
| ETTm1 | 96 | MSE | **0.325** | 0.333 | 0.326 | 0.331 | 0.334 | 0.329 |
| | | MAE | **0.360** | 0.367 | 0.361 | 0.366 | 0.368 | 0.364 |
| | 720 | MSE | **0.451** | 0.468 | 0.454 | 0.475 | 0.490 | 0.474 |
| | | MAE | **0.431** | 0.444 | 0.433 | 0.449 | 0.458 | 0.446 |
| | *Avg* | MSE | **0.388** | 0.401 | 0.390 | 0.403 | 0.412 | 0.402 |
| | | MAE | **0.396** | 0.406 | 0.398 | 0.408 | 0.413 | 0.405 |
| Weather | 96 | MSE | **0.160** | 0.166 | 0.162 | 0.170 | 0.174 | 0.169 |
| | | MAE | **0.206** | 0.211 | 0.207 | 0.212 | 0.214 | 0.210 |
| | 720 | MSE | **0.340** | 0.353 | 0.341 | 0.354 | 0.358 | 0.347 |
| | | MAE | **0.341** | 0.350 | 0.342 | 0.347 | 0.349 | 0.344 |
| | *Avg* | MSE | **0.250** | 0.260 | 0.252 | 0.262 | 0.267 | 0.259 |
| | | MAE | **0.274** | 0.281 | 0.275 | 0.279 | 0.282 | 0.277 |

**Settings.** Table 15 compares backbone capacity, the retained visual-language projector, and relation bias in a time-series backbone at $H = 96$ and 720. The projector-reset control randomizes only the visual-language projector, retaining the Full-model optimization policy and all other pretrained weights. The iTransformer control adds the same global channel-relation bias to channel attention while retaining its architecture and training configuration; MM-TS's local patch bias is not included. DINOv2 with a six-layer decoder is evaluated as a complete backbone alternative.

**Results.** Resetting the 2B projector raises average MSE from 0.424 to 0.438 on ETTh1 and from 0.388 to 0.401 on ETTm1. Adding global relation bias to iTransformer reduces the corresponding averages from 0.445 to 0.435 and from 0.412 to 0.402; Weather also improves from 0.267 to 0.259. The 4B backbone gives slightly higher average MSE than 2B on all three datasets.

**Analysis.** The projector reset increases MSE and MAE at both horizons on all three datasets, suggesting that the pretrained visual interface remains useful when the visual content consists of synthetic relation matrices. The iTransformer comparison examines a different part of the framework: global relation bias improves forecasting without introducing the Qwen visual-language backbone. These observations support both the visual interface and the structural attention mechanism. Meanwhile, increasing the backbone from 2B to 4B yields slightly higher average error in this evaluation. Read together with Appendix F, the results favor the selected combination of pretrained mapping and explicit structure at its measured computational cost, supporting Section 4.4.

### E.3 Relation-Metric Selection

Cov. denotes covariance; MI denotes mutual information.

Table 16: Relation-metric selection: MSE and MAE.

| Data | $H$ | Metric | DTW + Cov. + Pearson (full) | DTW only | Cov. only | Pearson only | Granger + Cov. + Pearson | DTW + MI + Pearson | DTW + Cov. + MI |
|---|---|---|---|---|---|---|---|---|---|
| ETTh1 | 96 | MSE | **0.379** | 0.389 | 0.394 | 0.392 | 0.385 | 0.381 | 0.382 |
| | | MAE | **0.392** | 0.400 | 0.402 | 0.402 | 0.397 | 0.393 | 0.395 |
| | 720 | MSE | **0.469** | 0.483 | 0.491 | 0.487 | 0.479 | 0.473 | 0.474 |
| | | MAE | **0.462** | 0.474 | 0.477 | 0.475 | 0.470 | 0.465 | 0.467 |
| | *Avg* | MSE | **0.424** | 0.436 | 0.443 | 0.440 | 0.433 | 0.428 | 0.429 |
| | | MAE | **0.427** | 0.438 | 0.440 | 0.439 | 0.434 | 0.429 | 0.431 |
| Weather | 96 | MSE | **0.160** | 0.168 | 0.172 | 0.167 | 0.165 | 0.163 | 0.164 |
| | | MAE | **0.206** | 0.213 | 0.215 | 0.212 | 0.210 | 0.207 | 0.209 |
| | 720 | MSE | **0.340** | 0.352 | 0.357 | 0.352 | 0.348 | 0.343 | 0.344 |
| | | MAE | **0.341** | 0.351 | 0.354 | 0.352 | 0.347 | 0.343 | 0.345 |
| | *Avg* | MSE | **0.250** | 0.260 | 0.265 | 0.259 | 0.256 | 0.253 | 0.254 |
| | | MAE | **0.274** | 0.282 | 0.285 | 0.282 | 0.279 | 0.276 | 0.277 |
| Electricity | 96 | MSE | **0.163** | 0.172 | 0.170 | 0.169 | 0.167 | 0.165 | 0.167 |
| | | MAE | **0.254** | 0.261 | 0.260 | 0.259 | 0.258 | 0.255 | 0.257 |
| | 720 | MSE | **0.230** | 0.244 | 0.240 | 0.238 | 0.237 | 0.233 | 0.234 |
| | | MAE | **0.320** | 0.330 | 0.329 | 0.326 | 0.326 | 0.322 | 0.323 |
| | *Avg* | MSE | **0.197** | 0.208 | 0.205 | 0.204 | 0.203 | 0.199 | 0.201 |
| | | MAE | **0.288** | 0.296 | 0.295 | 0.293 | 0.292 | 0.289 | 0.290 |
| Exchange | 96 | MSE | **0.077** | 0.083 | 0.087 | 0.083 | 0.081 | 0.078 | 0.081 |
| | | MAE | **0.197** | 0.203 | 0.204 | 0.202 | 0.200 | 0.198 | 0.199 |
| | 720 | MSE | **0.856** | 0.929 | 0.945 | 0.937 | 0.917 | 0.905 | 0.909 |
| | | MAE | **0.714** | 0.736 | 0.742 | 0.741 | 0.728 | 0.719 | 0.724 |
| | *Avg* | MSE | **0.467** | 0.506 | 0.516 | 0.511 | 0.499 | 0.492 | 0.495 |
| | | MAE | **0.456** | 0.470 | 0.473 | 0.472 | 0.464 | 0.459 | 0.462 |

**Settings.** Table 16 compares the three-metric representation with individual metrics and alternative combinations at two horizons.

**Results.** On ETTh1, average MSE is 0.424 for the combined representation, versus 0.436, 0.443, and 0.440 for DTW, covariance, and Pearson alone. The combined representation also has lower average MSE and MAE than each single metric on Weather, Electricity, and Exchange.

**Analysis.** DTW compares spectral shapes under flexible frequency alignment, covariance measures joint variation, and Pearson correlation measures normalized linear association. Using all three gives lower MSE and MAE than each individual metric on all four datasets. This result favors describing channel relations from multiple perspectives. MM-TS makes that combined description available to both the visual encoder and temporal-token attention, supporting the relation design in Section 4.4.

### E.4 ATTENTION-BIAS PLACEMENT

Table 17: Attention-bias injection layers: MSE and MAE.

| Data | $H$ | Metric | Layers 0–1 (full) | Layer 0 | Layers 0–2 | Layers 13–14 | Layers 26–27 | All 28 layers: matched energy | All 28 layers: learned gates |
|---|---|---|---|---|---|---|---|---|---|
| ETTh1 | 96 | MSE | **0.379** | 0.383 | 0.381 | 0.390 | 0.396 | 0.389 | 0.382 |
| | | MAE | **0.392** | 0.395 | 0.394 | 0.400 | 0.405 | 0.399 | 0.395 |
| | 720 | MSE | **0.469** | 0.476 | 0.473 | 0.486 | 0.493 | 0.484 | 0.475 |
| | | MAE | **0.462** | 0.466 | 0.465 | 0.473 | 0.478 | 0.472 | 0.466 |
| | *Avg* | MSE | **0.424** | 0.430 | 0.427 | 0.438 | 0.445 | 0.436 | 0.429 |
| | | MAE | **0.427** | 0.431 | 0.429 | 0.437 | 0.441 | 0.436 | 0.431 |
| Weather | 96 | MSE | **0.160** | 0.163 | 0.162 | 0.166 | 0.170 | 0.166 | 0.162 |
| | | MAE | **0.206** | 0.209 | 0.209 | 0.211 | 0.214 | 0.211 | 0.208 |
| | 720 | MSE | **0.340** | 0.347 | 0.343 | 0.354 | 0.357 | 0.351 | 0.344 |
| | | MAE | **0.341** | 0.345 | 0.344 | 0.350 | 0.353 | 0.350 | 0.344 |
| | *Avg* | MSE | **0.250** | 0.255 | 0.253 | 0.260 | 0.264 | 0.259 | 0.253 |
| | | MAE | **0.274** | 0.277 | 0.276 | 0.281 | 0.284 | 0.280 | 0.276 |
| Electricity | 96 | MSE | **0.163** | 0.166 | 0.165 | 0.170 | 0.173 | 0.170 | 0.165 |
| | | MAE | **0.254** | 0.257 | 0.256 | 0.260 | 0.262 | 0.260 | 0.256 |
| | 720 | MSE | **0.230** | 0.237 | 0.234 | 0.244 | 0.250 | 0.242 | 0.234 |
| | | MAE | **0.320** | 0.325 | 0.323 | 0.330 | 0.334 | 0.329 | 0.323 |
| | *Avg* | MSE | **0.197** | 0.202 | 0.200 | 0.207 | 0.212 | 0.206 | 0.200 |
| | | MAE | **0.288** | 0.291 | 0.290 | 0.295 | 0.298 | 0.294 | 0.290 |

**Settings.** Table 17 varies bias placement. We vary only the bias-injection layers, keeping LoRA at layers 0, 1, 26, and 27 and all other training settings unchanged. Layer indices are zero-based, and Full injects bias at layers 0 and 1. Injection configurations are initialized with matched total bias energy; the learned-gate variant additionally learns a bias coefficient for each layer.

**Results.** Layers 0 and 1 give the lowest average MSE on all three datasets. On ETTh1, the average is 0.424, compared with 0.430 for layer 0 alone, 0.438 for layers 13 and 14, and 0.445 for layers 26 and 27. Extending injection to all 28 layers yields 0.436 with matched energy and 0.429 with learned gates.

**Analysis.** Early injection makes channel guidance available when temporal-token representations first interact in the decoder, so subsequent transformations can operate on representations already influenced by the structural prior. The observed ordering is consistent with this rationale: early-layer injection outperforms the tested middle- and late-layer placements, while extending injection to all layers also gives higher average error. Thus, the tested configurations favor focused early guidance, supporting the placement in Section 4.4. This comparison varies bias locations while retaining the stated LoRA placement.

### E.5 Relation-Fusion Weights

Table 18: Relation-fusion weights: MSE and MAE.

| Data | $H$ | Metric | MM-TS (full) | Equal weights | Fixed weights and scale | Perturbed ratios | Learned weights per dataset | Learned weights per layer |
|---|---|---|---|---|---|---|---|---|
| ETTh1 | 96 | MSE | **0.379** | 0.383 | 0.388 | 0.381 | 0.380 | 0.381 |
| | | MAE | **0.392** | 0.395 | 0.398 | 0.394 | 0.393 | 0.393 |
| | 720 | MSE | **0.469** | 0.478 | 0.485 | 0.473 | 0.471 | 0.472 |
| | | MAE | **0.462** | 0.468 | 0.473 | 0.465 | 0.464 | 0.464 |
| | *Avg* | MSE | **0.424** | 0.430 | 0.437 | 0.428 | 0.426 | 0.427 |
| | | MAE | **0.427** | 0.432 | 0.436 | 0.430 | 0.428 | 0.429 |
| Weather | 96 | MSE | **0.160** | 0.163 | 0.165 | 0.161 | 0.161 | 0.161 |
| | | MAE | **0.206** | 0.208 | 0.210 | 0.208 | 0.207 | 0.207 |
| | 720 | MSE | **0.340** | 0.345 | 0.351 | 0.343 | 0.341 | 0.342 |
| | | MAE | **0.341** | 0.345 | 0.349 | 0.343 | 0.342 | 0.343 |
| | *Avg* | MSE | **0.250** | 0.254 | 0.258 | 0.252 | 0.251 | 0.252 |
| | | MAE | **0.274** | 0.277 | 0.280 | 0.275 | 0.275 | 0.276 |
| Electricity | 96 | MSE | **0.163** | 0.165 | 0.168 | 0.165 | 0.164 | 0.164 |
| | | MAE | **0.254** | 0.256 | 0.258 | 0.255 | 0.255 | 0.255 |
| | 720 | MSE | **0.230** | 0.234 | 0.239 | 0.232 | 0.231 | 0.231 |
| | | MAE | **0.320** | 0.324 | 0.328 | 0.322 | 0.321 | 0.322 |
| | *Avg* | MSE | **0.197** | 0.200 | 0.204 | 0.198 | 0.198 | 0.198 |
| | | MAE | **0.288** | 0.291 | 0.293 | 0.289 | 0.288 | 0.289 |

**Settings.** Table 18 compares fusion-weight controls. These controls modify bias-side fusion while keeping the image and video inputs unchanged. Equal weights preserve the Full-model weight sum. The ratio perturbation multiplies the DTW, covariance, and Pearson weights by 1.2, 0.8, and 1.0, respectively, then restores their original sum. The fixed variant uses weights (1, 0.5, 2) and a constant bias scale of 0.05. Learned variants start from Full and optimize dataset-level or layer-level weight ratios with the sum preserved.

**Results.** The weight-ratio perturbation changes average MSE from 0.424 to 0.428 on ETTh1, 0.250 to 0.252 on Weather, and 0.197 to 0.198 on Electricity. Fixing both weights and scale produces larger increases to 0.437, 0.258, and 0.204. Learned weights per dataset or per layer remain close to the selected configuration.

**Analysis.** Fusion weights determine the relative contribution of the relation descriptors, while the bias scale controls their influence on attention logits. The perturbation results show that the selected configuration tolerates the tested change in relation composition: the weight-ratio perturbation increases average MSE by 0.004, 0.002, and 0.001 on ETTh1, Weather, and Electricity. Fixing both weights and scale produces larger increases, favoring the default combination of relation composition and scaling. Learned weights per dataset or layer give nearby results, suggesting that the selected configuration captures a useful balance within the explored alternatives. These findings support the practical configuration adopted in Section 4.4.

### E.6 Visual Sequence Length

Table 19: Number of visual relation frames: MSE and MAE.

| Data | $H$ | Metric | 11 frames (full) | 0 frames | 1 frame | 3 frames | 6 frames |
|---|---|---|---|---|---|---|---|
| ETTh1 | 96 | MSE | **0.379** | 0.389 | 0.386 | 0.383 | 0.381 |
| | | MAE | **0.392** | 0.400 | 0.398 | 0.395 | 0.394 |
| | 720 | MSE | **0.469** | 0.490 | 0.484 | 0.477 | 0.472 |
| | | MAE | **0.462** | 0.477 | 0.473 | 0.468 | 0.465 |
| | *Avg* | MSE | **0.424** | 0.440 | 0.435 | 0.430 | 0.427 |
| | | MAE | **0.427** | 0.439 | 0.436 | 0.432 | 0.430 |
| Weather | 96 | MSE | **0.160** | 0.165 | 0.164 | 0.163 | 0.162 |
| | | MAE | **0.206** | 0.211 | 0.210 | 0.208 | 0.207 |
| | 720 | MSE | **0.340** | 0.353 | 0.350 | 0.345 | 0.342 |
| | | MAE | **0.341** | 0.351 | 0.348 | 0.345 | 0.343 |
| | *Avg* | MSE | **0.250** | 0.259 | 0.257 | 0.254 | 0.252 |
| | | MAE | **0.274** | 0.281 | 0.279 | 0.277 | 0.275 |
| Electricity | 96 | MSE | **0.163** | 0.168 | 0.167 | 0.165 | 0.164 |
| | | MAE | **0.254** | 0.258 | 0.258 | 0.256 | 0.255 |
| | 720 | MSE | **0.230** | 0.240 | 0.237 | 0.234 | 0.232 |
| | | MAE | **0.320** | 0.328 | 0.326 | 0.324 | 0.322 |
| | *Avg* | MSE | **0.197** | 0.204 | 0.202 | 0.200 | 0.199 |
| | | MAE | **0.288** | 0.293 | 0.292 | 0.290 | 0.289 |

**Settings.** Table 19 compares 0, 1, 3, 6, and 11 frames. We fix patch length and stride at 16/8 and vary only the visual video sequence. The zero-frame variant retains the global image, text, temporal tokens, and original attention bias. The 1-, 3-, and 6-frame variants pool the local relation frames in frequency order.

**Results.** Average MSE decreases progressively with the tested frame counts on all three datasets. On ETTh1, the sequence is 0.440, 0.435, 0.430, 0.427, and 0.424; Weather decreases from 0.259 to 0.250 and Electricity from 0.204 to 0.197 between the endpoints.

**Analysis.** With temporal patching and structural bias fixed, the lower error at larger frame counts supports preserving finer local spectral information in the visual path. This result supports the multi-frame visual representation in Section 4.4.

### E.7 Patch Configuration

Table 20: Joint temporal and visual patch configurations: MSE and MAE.

| Data | $H$ | Metric | Length 16 stride 8 (full) | Length 8 stride 4 | Length 24 stride 12 | Length 32 stride 16 |
|---|---|---|---|---|---|---|
| ETTh1 | 96 | MSE | **0.379** | 0.380 | 0.382 | 0.387 |
| | | MAE | **0.392** | 0.393 | 0.395 | 0.398 |
| | 720 | MSE | **0.469** | 0.474 | 0.477 | 0.483 |
| | | MAE | **0.462** | 0.465 | 0.468 | 0.471 |
| | *Avg* | MSE | **0.424** | 0.427 | 0.430 | 0.435 |
| | | MAE | **0.427** | 0.429 | 0.431 | 0.435 |

Continued on next page

Table 20 (continued)

| Data | $H$ | Metric | Length 16 stride 8 (full) | Length 8 stride 4 | Length 24 stride 12 | Length 32 stride 16 |
|---|---|---|---|---|---|---|
| Weather | 96 | MSE | **0.160** | 0.161 | 0.163 | 0.166 |
| | | MAE | **0.206** | 0.207 | 0.208 | 0.210 |
| | 720 | MSE | **0.340** | 0.343 | 0.345 | 0.350 |
| | | MAE | **0.341** | 0.344 | 0.346 | 0.348 |
| | *Avg* | MSE | **0.250** | 0.252 | 0.254 | 0.258 |
| | | MAE | **0.274** | 0.276 | 0.277 | 0.279 |
| Electricity | 96 | MSE | **0.163** | 0.164 | 0.166 | 0.168 |
| | | MAE | **0.254** | 0.255 | 0.256 | 0.258 |
| | 720 | MSE | **0.230** | 0.233 | 0.235 | 0.238 |
| | | MAE | **0.320** | 0.323 | 0.324 | 0.327 |
| | *Avg* | MSE | **0.197** | 0.198 | 0.201 | 0.204 |
| | | MAE | **0.288** | 0.289 | 0.290 | 0.293 |

**Settings.** Table 20 jointly varies temporal and spectral patch length and stride. At $L = 96$, configurations 8/4, 16/8, 24/12, and 32/16 yield 23, 11, 7, and 5 patches, respectively. Every local relation matrix is retained as a video frame, so temporal patch count and video frame count change together.

**Results.** The selected 16/8 pair has the lowest average MSE on all three datasets: 0.424 on ETTh1, 0.250 on Weather, and 0.197 on Electricity. Using 32/16 raises these values to 0.435, 0.258, and 0.204. The 8/4 pair is closer, with averages of 0.427, 0.252, and 0.198.

**Analysis.** The 16/8 configuration gives the lowest average MSE among the joint configurations on all three datasets, while both finer 8/4 and coarser 32/16 settings give higher error. Patch length and stride affect temporal segmentation, local spectral support, and visual sequence length together. The results favor the selected balance across these coupled components. Together with the fixed-bias frame comparison in Appendix E.6, this experiment supports the design choices in Section 4.4.

# F Computational Cost

Table 21: Computational cost on ETTh1 ($L = H = 96$).

| Model or setting | Total parameters (M) | Trainable parameters (M) | Training time (s/epoch) | Training memory (GB) | Inference memory (GB) | Inference latency (ms/sample) |
|---|---|---|---|---|---|---|
| *MM-TS and component variants* | | | | | | |
| MM-TS (full) | 2058.664 | 6.672 | 222.992 | 8.135 | 3.748 | 25.825 |
| Without visual tokens | 2058.664 | 6.672 | 197.109 | 7.665 | 3.668 | 18.223 |
| Without text prompt | 2058.664 | 6.672 | 84.698 | 4.563 | 3.733 | 19.077 |
| Without attention bias | 2058.664 | 6.672 | 213.907 | 8.135 | 3.748 | 24.787 |
| *Forecasting baselines* | | | | | | |
| Time-VLM | 151.425 | 0.148 | 59.959 | 0.696 | 0.688 | 14.800 |
| VisionTS | 111.908 | 0.056 | 39.197 | 1.653 | 0.541 | 6.435 |
| TimesFM | 203.569 | 6.725 | 61.310 | 1.451 | 0.908 | 10.022 |
| iTransformer | 0.842 | 0.842 | 12.978 | 0.042 | 0.027 | 1.552 |
| Time-LLM | 132.804 | 50.891 | 104.787 | 3.976 | 0.784 | 11.692 |
| TSMixer | 0.223 | 0.223 | 6.335 | 0.027 | 0.025 | 0.597 |
| PatchTST | 0.035 | 0.035 | 18.583 | 0.025 | 0.023 | 2.000 |
| TimesNet | 0.605 | 0.605 | 54.495 | 0.038 | 0.027 | 8.040 |
| DLinear | 0.019 | 0.019 | 2.856 | 0.023 | 0.023 | 0.279 |

Continued on next page

Table 21 (continued)

| Model or setting | Total parameters (M) | Trainable parameters (M) | Training time (s/epoch) | Training memory (GB) | Inference memory (GB) | Inference latency (ms/sample) |
|---|---|---|---|---|---|---|
| *Backbone and relation-bias variants* | | | | | | |
| 2B, reset visual projector | 2058.664 | 6.672 | 220.337 | 8.135 | 3.748 | 25.537 |
| Qwen3-VL-4B | 4365.675 | 9.692 | 494.526 | 18.635 | 8.319 | 40.894 |
| DINOv2 + 6-layer decoder | 112.608 | 26.027 | 41.095 | 0.958 | 0.535 | 37.290 |
| iTransformer + relation bias | 0.842 | 0.842 | 12.747 | 0.042 | 0.027 | 27.247 |

Table 22: Computational cost on ETTm1 ($L = H = 96$).

| Model or setting | Total parameters (M) | Trainable parameters (M) | Training time (s/epoch) | Training memory (GB) | Inference memory (GB) | Inference latency (ms/sample) |
|---|---|---|---|---|---|---|
| *MM-TS and component variants* | | | | | | |
| MM-TS (full) | 2058.664 | 6.672 | 871.364 | 8.133 | 3.748 | 24.802 |
| Without visual tokens | 2058.664 | 6.672 | 780.760 | 7.665 | 3.668 | 17.825 |
| Without text prompt | 2058.664 | 6.672 | 344.409 | 4.565 | 3.733 | 18.942 |
| Without attention bias | 2058.664 | 6.672 | 859.757 | 8.135 | 3.748 | 24.607 |
| *Forecasting baselines* | | | | | | |
| Time-VLM | 151.425 | 0.148 | 241.172 | 0.696 | 0.688 | 14.229 |
| VisionTS | 111.908 | 0.056 | 153.085 | 1.653 | 0.541 | 6.407 |
| TimesFM | 203.569 | 6.725 | 247.481 | 1.451 | 0.908 | 10.014 |
| iTransformer | 0.842 | 0.842 | 51.015 | 0.042 | 0.027 | 1.564 |
| Time-LLM | 132.804 | 50.891 | 421.785 | 4.157 | 0.784 | 11.509 |
| TSMixer | 0.223 | 0.223 | 26.434 | 0.029 | 0.025 | 0.548 |
| PatchTST | 0.035 | 0.035 | 68.743 | 0.025 | 0.023 | 1.919 |
| TimesNet | 0.605 | 0.605 | 200.883 | 0.036 | 0.027 | 8.127 |
| DLinear | 0.019 | 0.019 | 11.361 | 0.023 | 0.023 | 0.307 |
| *Backbone and relation-bias variants* | | | | | | |
| 2B, reset visual projector | 2058.664 | 6.672 | 883.998 | 8.133 | 3.748 | 25.662 |
| Qwen3-VL-4B | 4365.675 | 9.692 | 2034.012 | 18.633 | 8.319 | 41.588 |
| DINOv2 + 6-layer decoder | 112.608 | 26.027 | 164.715 | 0.958 | 0.535 | 36.508 |
| iTransformer + relation bias | 0.842 | 0.842 | 51.627 | 0.042 | 0.027 | 26.884 |

Table 23: Computational cost on Weather ($L = H = 96$).

| Model or setting | Total parameters (M) | Trainable parameters (M) | Training time (s/epoch) | Training memory (GB) | Inference memory (GB) | Inference latency (ms/sample) |
|---|---|---|---|---|---|---|
| *MM-TS and component variants* | | | | | | |
| MM-TS (full) | 2058.693 | 6.701 | 1529.215 | 10.119 | 3.798 | 34.889 |
| Without visual tokens | 2058.693 | 6.701 | 1390.460 | 9.372 | 3.712 | 27.747 |
| Without text prompt | 2058.693 | 6.701 | 807.814 | 6.862 | 3.748 | 28.064 |
| Without attention bias | 2058.693 | 6.701 | 1542.475 | 10.117 | 3.779 | 35.655 |
| *Forecasting baselines* | | | | | | |
| Time-VLM | 151.425 | 0.148 | 263.367 | 0.696 | 0.688 | 14.669 |
| VisionTS | 111.908 | 0.056 | 415.231 | 3.441 | 0.612 | 7.018 |
| TimesFM | 203.569 | 6.725 | 267.157 | 1.474 | 0.914 | 9.965 |
| iTransformer | 0.842 | 0.842 | 53.846 | 0.042 | 0.029 | 1.456 |

Continued on next page

Table 23 (continued)

| Model or setting | Total parameters (M) | Trainable parameters (M) | Training time (s/epoch) | Training memory (GB) | Inference memory (GB) | Inference latency (ms/sample) |
|---|---|---|---|---|---|---|
| Time-LLM | 132.804 | 50.891 | 1315.729 | 12.598 | 1.065 | 25.834 |
| TSMixer | 0.251 | 0.251 | 25.894 | 0.029 | 0.025 | 0.597 |
| PatchTST | 0.035 | 0.035 | 72.352 | 0.029 | 0.023 | 2.139 |
| TimesNet | 0.606 | 0.606 | 250.386 | 0.038 | 0.027 | 8.854 |
| DLinear | 0.019 | 0.019 | 12.012 | 0.023 | 0.023 | 0.270 |
| *Backbone and relation-bias variants* | | | | | | |
| 2B, reset visual projector | 2058.693 | 6.701 | 1544.509 | 10.119 | 3.798 | 34.979 |
| Qwen3-VL-4B | 4365.710 | 9.728 | 3734.995 | 22.750 | 8.393 | 59.111 |
| DINOv2 + 6-layer decoder | 113.412 | 26.831 | 172.110 | 0.981 | 0.535 | 253.626 |
| iTransformer + relation bias | 0.842 | 0.842 | 57.207 | 0.042 | 0.029 | 242.896 |

**Settings.** Tables 21–23 compare parameter count, training time, peak GPU memory, and inference latency at $L = H = 96$. Measurements use an RTX 4090 with BF16 autocast, training batch size 4 and inference batch size 1. After one warm-up training epoch and 20 warm-up inference batches, we time three full training epochs and 200 inference batches. Cost measurements use seed 2026. Epoch times include the final partial batch and exclude preparation of the frozen visual cache; inference latency includes the required input processing. The measurement platform shares CPU, RAM, and storage resources across concurrent GPU activity. The tables combine updated measurements for 18 MM-TS configurations with 33 previously measured model-setting entries. GPU memory is peak reserved memory in decimal GB ($10^9$ bytes). In the computational-cost evaluation, Time-LLM uses a GPT-2 backbone.

**Results.** Tables 21–23 report full-model latency of 25.825, 24.802, and 34.889 ms per sample on ETTh1, ETTm1, and Weather. On ETTh1, the 2B model has 6.672 million trainable parameters out of 2058.664 million total parameters, approximately 0.32%. The 4B variant increases training memory from 8.135 to 18.635 GB and latency from 25.825 to 40.894 ms without improving the average forecasting accuracy in Table 15.

**Component costs.** Removing visual tokens reduces ETTh1 latency from 25.825 to 18.223 ms, accompanied by higher forecasting error in Table 14. Removing attention bias gives 24.787 ms on ETTh1, alongside a larger accuracy loss than removing visual tokens in Table 3. Removing text gives 19.077 ms. On Weather, the bias-disabled variant is slightly slower than the full model (35.655 versus 34.889 ms), indicating that the measured runtime effect also depends on the dataset. These comparisons help assess each path's contribution at its observed cost.

**Interpretation.** MM-TS adapts a small fraction of its parameters, while inference uses the full pretrained backbone and multimodal sequence. The 2B model achieves lower average error, memory use, and latency than 4B in these experiments. Alongside the component accuracy and runtime comparisons, these measurements explain the choice of the 2B configuration in Section 4.4.

# G Relation Stability and Case Studies

## G.1 Relation Stability

We divide each training segment into non-overlapping windows of length 96 (104 for ILI), discarding an incomplete final window. For each window, we form a global signed relation matrix using DTW/covariance/Pearson weights of 1/0.5/2 and no further normalization. Let $v_t$ contain its strictly upper-triangular entries. We define $S_{\text{rel}}$ as the mean cosine similarity between $v_t$ and $v_{t+1}$. These fixed diagnostic weights are separate from the horizon-specific training configurations. The diagnostic uses DTW band 8 and $\tau = 1$, with covariance normalized using the 1st and 99th percentiles. Table 24 reports its 95% interval from 10,000 block-bootstrap resamples alongside forecasting accuracy.

Table 24: Training-period relation stability and forecasting accuracy.

| Dataset | $C$ | MM-TS MSE | MM-TS MAE | Best baseline | Baseline MSE | MSE gain | $S_{\text{rel}}$ | 95% interval |
|---|---|---|---|---|---|---|---|---|
| ETTh1 | 7 | 0.434 | 0.429 | VisionTS | 0.437 | +0.69% | 0.677 | [0.637, 0.714] |
| ETTh2 | 7 | 0.368 | 0.398 | Time-VLM | 0.382 | +3.66% | 0.687 | [0.627, 0.742] |
| ETTm1 | 7 | 0.383 | 0.395 | PatchTST | 0.388 | +1.29% | 0.713 | [0.679, 0.742] |
| ETTm2 | 7 | 0.276 | 0.321 | Time-VLM | 0.278 | +0.72% | 0.724 | [0.693, 0.755] |
| Weather | 21 | 0.240 | 0.271 | TSMixer | 0.244 | +1.64% | 0.868 | [0.855, 0.880] |
| ILI | 7 | 1.560 | 0.837 | PatchTST | 1.514 | -3.04% | 0.952* | [0.944*, 0.961*] |
| Exchange | 8 | 0.351 | 0.401 | DLinear | 0.336 | -4.46% | 0.725 | [0.678, 0.771] |
| Electricity | 321 | 0.190 | 0.281 | VisionTS | 0.192 | +1.04% | 0.913 | [0.909, 0.918] |

* ILI uses five adjacent training-window pairs. MSE gain is computed from the displayed means as $100(\text{MSE}_{\text{base}} - \text{MSE}_{\text{MM-TS}})/\text{MSE}_{\text{base}}$.

**Results and interpretation.** Channel relations show recurring patterns across adjacent windows, with reported stability values of 0.713 on ETTm1, 0.868 on Weather, and 0.913 on Electricity. The case studies also show local changes in individual associations. These observations motivate representing the recurring structure explicitly while recomputing it for each current input. Forecasting utility also depends on how that structure interacts with the task: ILI has a high stability estimate but a higher MSE than PatchTST. Its estimate is based on five window pairs. Table 24 therefore characterizes structural persistence, while the current-window and permutation controls in Appendix E.1 assess its use in forecasting, complementing Section 4.5.

## G.2 Additional Case Studies

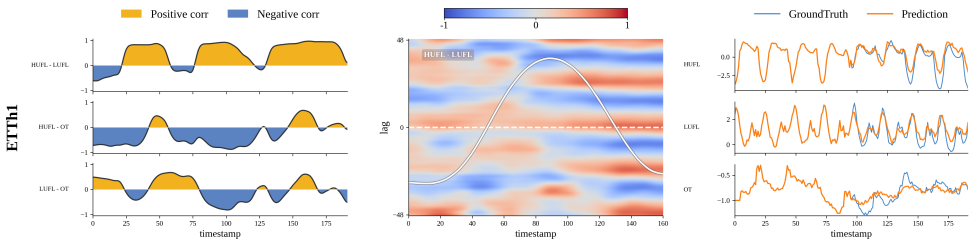



Figure 4: ETTh1 case study. Panels and colors follow Figure 3. Positive lag means the first variable leads the second; negative lag indicates the reverse.

**ETTh1.** Figure 4 contrasts rapid load fluctuations with smoother oil-temperature evolution. HUFL–LUFL is locally positive, whereas load–OT pairs show longer negative intervals. Forecasts

preserve these response scales. Different ETTh1 and ETTm1 correlation patterns despite shared variables motivate window-specific relations.

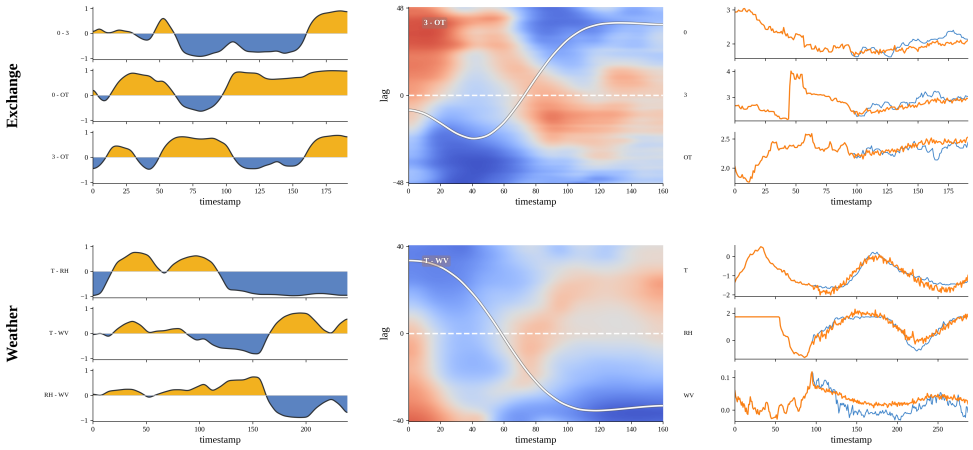



Figure 5: Exchange (top) and Weather (bottom) case studies. Panels, colors, and lag signs follow Figures 3 and 4.

**Exchange and Weather.** Figure 5 extends this comparison across variable semantics. Exchange shows changing co-movement and a lag ridge crossing zero. Weather combines evolving temperature-humidity relations and temperature-wind lags; forecasts preserve smooth temperature and humidity trajectories alongside the wind-speed envelope. Window-specific structure contextualizes each trajectory (Section 4.5), complementing component evidence in Appendix E.1.

## H Asset Licenses

The materials released with the paper are licensed under CC-BY 4.0. Public datasets, baseline implementations, and the Qwen3-VL backbone remain subject to their original licenses and terms of use. Their creators are credited in the bibliography and released source.

## I Broader Impacts

MM-TS supports multivariate forecasting in energy, weather, finance, and industrial monitoring. Its explicit relation representations can also help inspect how channel organization changes across observed windows.