# VIFO: VISUAL FEATURE EMPOWERED MULTIVARIATE TIME SERIES FORECASTING WITH CROSS-MODAL FUSION

*Yanlong Wang*[1,2⋆], *Hang Yu*[3⋆], *Jian Xu*[1], *Fei Ma*[4], *Hongkang Zhang*[1], *Tongtong Feng*[1]
*Zijian Zhang*[5], *Shao-Lun Huang*[1†], *Danny Dongning Sun*[2†], *Xiao-Ping Zhang*[1†]

[1] Tsinghua University [2] Pengcheng Laboratory [3] Ant Group
[4] Guangming Laboratory [5] University of Pennsylvania

## ABSTRACT

Large time series foundation models often adopt channel-independent architectures to handle varying data dimensions, but this design ignores crucial cross-channel dependencies. Concurrently, existing multimodal approaches have not fully exploited the power of large vision models (LVMs) to interpret spatiotemporal data. Additionally, there remains significant unexplored potential in leveraging the advantages of information extraction from different modalities to enhance time series forecasting performance. To address these gaps, we propose the VIFO, a cross-modal forecasting model. VIFO uniquely renders multivariate time series into image, enabling pre-trained LVM to extract complex cross-channel patterns that are invisible to channel-independent models. These visual features are then aligned and fused with representations from the time series modality. By freezing the LVM and training only 7.45% of its parameters, VIFO achieves competitive performance on multiple benchmarks, offering an efficient and effective solution for capturing cross-variable relationships in time series forecasting.

***Index Terms*—** Time Series Forecasting, Spatiotemporal Representation, Cross-modal Fusion

## 1. INTRODUCTION

Time series forecasting has been widely applied in diverse settings such as weather, power systems, transportation, and finance [1–4]. These scenarios often involve a wide variety of temporal data, where different time series frequently exhibit intricate interrelationships. Early time series forecasting methods focus on statistical models and signal processing techniques like decomposition [5, 6] and frequency analysis [7]; with the development of deep learning, many studies improve prediction by fusing information across both variable and temporal dimensions of multivariate time series [8,9], and enhance the model's ability to process local features [10, 11]; recently, the field has shifted towards large foundation models [12–16] and further improves performance by incorporating cross-modal information fusion [17].

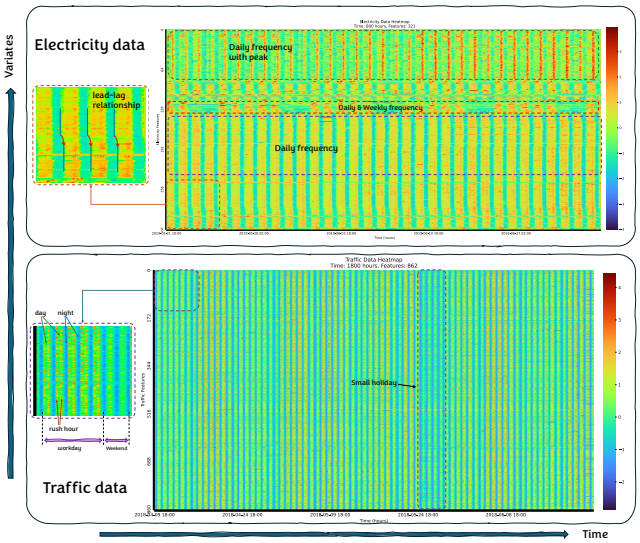


**Fig. 1**. Clear Patterns in Electricity and Traffic Time Series Visualization: Periodicity, Lead-Lag Relationships, Anomalous Events.

However, several issues remain unresolved [18]. Firstly, the channel-independent architecture is commonly adopted in large time series models. While this allows the model to handle varying numbers of variables across datasets, it overlooks valuable information provided by other time series. Although many studies [19, 20] employ channel-dependent designs, they require full-parameter training on large scale time series datasets, which increases both data volume and training parameters, and also face challenges in effectively learning the distinct cross-variable dependency patterns inherent to time series from different domains. Moreover, flattening multivariate series into a univariate sequence may weaken the recognition of cross-channel patterns. Secondly, previous cross-modal fusion methods often relied on textual modalities to provide auxiliary temporal information [17, 21–24], yet underutilizing the information capture capabilities of large models from different modalities. Besides, the global and local

⋆ Equal contribution.
† Corresponding author.

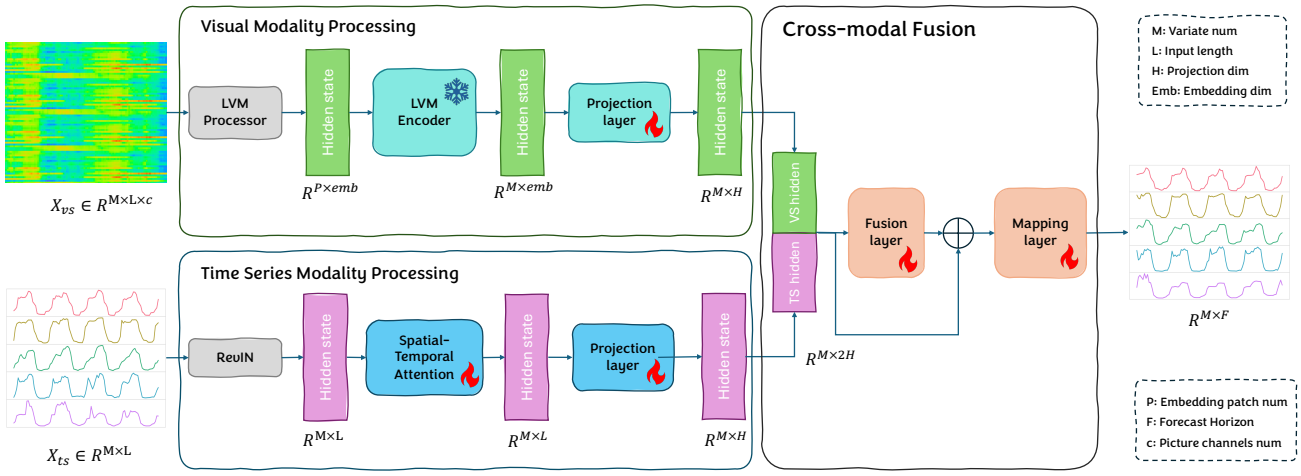


**Fig. 2**. The overall structure of VIFO, which simultaneously processes information across temporal and spatial dimensions from both image and time series modalities.

feature extraction capabilities of visual modalities have not been fully leveraged in spatiotemporal processing of time series data [25].

To address the aforementioned challenges, we develop a novel multi-modal solution, which not only considers the sequence of numerical data points, but also leverages the visual representational power of pre-trained visual models to capture spatiotemporal dependencies from raw multivariate time series plots. In particular, it employs the variable-size variants of pre-trained visual models to overcome input format issues caused by the imbalance between the number of variables and time series length in multivariate time series. These features captured by the visual model are then projected onto encoded representations in the temporal modality and further enhanced through cross-modal fusion (we call it the VIFO model). By this way, significant improvements in forecasting performance could be achieved as verified by extensive experiments. This work makes four main contributions:

- It offers a novel approach for leveraging large foundation models to tackle cross-variable dependency, demonstrating that pre-trained vision models can effectively capture complex cross-channel dependencies from simple 2D visualizations of time series data.
- It provide a method to apply pre-trained vision models to multivariate time series of arbitrary dimensions by rendering them as variable-sized images.
- During training, the model freezes the majority of its parameters, with only 7.45% of the total model parameters are trainable, and this efficient training method is verified to be effective for time series.
- Through cross-modal fusion and alignment, the VIFO model achieves impressive forecasting performance on multiple benchmark datasets.

## 2. METHODS

### 2.1. Problem Formulation

As shown in Figure 2, the input multivariate time series is denoted as $X_{ts} \in \mathbb{R}^{M\times L}$ with look-back window $L$ : $(x_1, x_2, \ldots, x_L)$ and $M$ variables, where $x_t$ represents the $M$-dimensional vector at time step $t$. Additionally, the visual input $X_{vs} \in \mathbb{R}^{M\times L\times c}$ is generated by rendering the $M \times L$ time series data as an $M \times L$ pixel image, where each row represents a variable and each column as time step, and $c$ represents the number of RGB channels of the image. The prediction target of the model is the future values of the next $F$ time steps, denoted as $(x_{L+1}, \ldots, x_{L+F})$.

### 2.2. Visual Analysis of Multivariate Time Series

Vision is the primary way for humans to obtain information from the external world. To investigate whether multivariate time-series graphs also contain patterns and regularities recognizable by the human eye, we visualized the Traffic and ECL datasets - commonly used in time-series forecasting tasks - as shown in Figure 1. As can be seen from the figure, the multivariate values of each dataset exhibit different color variations as they change over time. Among them, both the ECL and Traffic datasets show relatively obvious characters, such as daily periodic changes.

Specifically, in the upper subplot of Figure 1 displaying the ECL data, different variables exhibit distinct periodic characteristics. Some variables demonstrate brief yet pronounced peak patterns, while others display superimposed daily and weekly cyclical patterns. Moreover, certain variable pairs maintain relatively stable lead-lag relationships, which provide a foundation for capturing cross-variable dependencies in time series forecasting. For the visualization of the Traffic dataset in the lower part of Figure 1, it exhibits obvious multi-period superimposed patterns. Specifically,

from the zoomed-in section, based on the numerical changes represented by colors, distinct diurnal (day-night) traffic flow variations can be observed. Furthermore, there are significant differences in traffic flow between five consecutive weekdays and the following two weekend days: during the daytime on weekdays, relatively distinct morning and evening rush hours are visible, while this phenomenon is significantly alleviated on weekends. In addition, small holidays can also be identified from the multivariate time series graph, which exhibits a continuous alleviation of traffic rush hours lasting for more than two days (4 days in total for holidays).

All the aforementioned patterns and regularities are directly observable to the naked eye from the visualizations of multivariate time series - let alone more complex patterns that cannot be directly observed. These observations demonstrate the feasibility of visual analysis for multivariate time series, as well as the importance of integrating the temporal and spatial dimensions of time series data - an aspect that is difficult to achieve with channel-independent architectures.

### 2.3. Model Architecture

Our model, depicted in Figure 2, comprises three key stages: (1)Unimodal feature extraction via visual processor and time series processor, (2) Cross-Modal Fusion layer, and (3) a final Mapping layer for prediction. At first stage, the input data of each modality passes through respective modality-specific processing module, extracting features from a single modality and output the high-dimensional hidden states. During the cross-modal fusion in the second stage, the hidden states from each modality in the previous stage are merged. Then, the cross-modal attention module in the fusion layer processes the cross-modal information, and finally, the predicted time series is obtained after mapping.

### 2.4. Visual Modality Processing

In the visual modality processing stage, the input $X_{vs}$ first passes through the Processor of the vision large model, and the resulting output is $X_{vs}^{out} \in \mathbb{R}^{P \times emb}$, where $P$ denotes the number of embedding patches and $emb$ represents the embedding dimension. Subsequently, the output is fed into the encoder of the vision large model to obtain the encoded representation. Then, it goes through a mapping layer (a multi-layer MLP structure) to output the processed hidden state $hidden_{ts} \in \mathbb{R}^{M \times H}$, where $M$ is the number of variables and $H$ is the dimension of the projection representation. During the training process, the parameters of the vision large model are frozen, and only other parts of the network need to be fine-tuned. This approach not only leverages the capabilities of the vision model but also greatly reduces the number of trainable parameters.

### 2.5. Time Series Modality Processing

In the time series modality processing stage, the input $X_{ts}$ first undergoes normalization [26] along the temporal dimension, and then is fed into an attention module capable of capturing spatiotemporal dimension information. Here, We adopt a spatialtemporal attention structure [27], which converts multivariate time series into spatiotemporal segments and applies the self-attention mechanism to these segments to capture global and local information of cross-channel time series. Subsequently, the processed hidden state is output through the projection layer, which is an MLP structure.

**Table 1**. Details of parameter amounts in the VIFO Modules.

| | **Frozen Part** | **Trainable Part** | | | |
|---|---|---|---|---|---|
| | **Visual Model** | **Visual Projection** | **Temporal Network** | **Fusion Network** | **Total** |
| **Parameters** | 375M | 1.8M | 16M | 12M | 30M |
| **Proportion** | 92.55% | 0.45% | 4.00% | 3.00% | 7.45% |

### 2.6. Cross-Modal Fusion

In this stage, the hidden states of information from each modality obtained in the previous stage are merged first. Then, the cross-modal information is processed through the spatiotemporal attention module in the fusion layer, and finally, the predicted time series is obtained through mapping. The parameter counts of each part are presented in Table 1.

## 3. EXPERIMENTS

### 3.1. Setup

Since the number of variates in multivariate time series varies across different datasets, and the variable dimension and temporal dimension length of the input time series are often highly imbalanced, we adopt the encoder of SigLip2-base-Naflex [28] as the backbone of the vision large model, which allows arbitrary adjustment of image height and width. During training, all parameters of the visual encoder are frozen, and only the remaining parts of VIFO are trained, with trainable parameters accounting for only 7.45% of the total parameters. The input time series length $L = 512$, and the output length $F \in \{96, 192, 336, 720\}$.

We select multiple competitive models as baselines, including the Base and Large versions of Chronos [29] and Moirai [19], as well as UniTST [30] and GPT4TS [13]. Small models such as TimesNet [6] and PatchTST [10] were also included. Experiments were conducted on 7 datasets, including ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Weather, and Traffic. To ensure the robustness of results, each experiment was run 3 times with different random seeds. Chronos and Moirai did not report results on the Traffic dataset, as this dataset has an excessive number of variables.

### 3.2. Results

As shown in Table 2, the VIFO model achieved lower MSE and MAE loss values on 7 datasets, demonstrating its competitive predictive performance. Furthermore, the patterns

**Table 2**. Forecasting results comparison. MSE and MAE are evaluated on the benchmark dataset, with the prediction horizon $F \in \{96, 129, 336, 720\}$ and input length of 512. A lower value indicates better performance. Values in **bold** denote the best performance, while values with underline indicate the second-best performance.

| Models | | VIFO | | Chronos$_{Base}$ | | Chronos$_{Large}$ | | Moirai$_{Base}$ | | Moirai$_{Large}$ | | GPT4TS | | UniTST | | TimesNet | | PatchTST | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric | | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| ETTh1 | 96 | **0.348** | **0.383** | 0.440 | 0.393 | 0.441 | 0.390 | 0.376 | 0.392 | 0.381 | 0.388 | 0.376 | 0.397 | 0.383 | 0.398 | 0.452 | 0.463 | 0.404 | 0.413 |
| | 192 | **0.383** | **0.405** | 0.492 | 0.426 | 0.502 | 0.524 | 0.412 | 0.413 | 0.434 | 0.415 | 0.416 | 0.418 | 0.434 | 0.426 | 0.474 | 0.477 | 0.454 | 0.430 |
| | 336 | **0.398** | **0.414** | 0.550 | 0.462 | 0.576 | 0.467 | 0.433 | 0.428 | 0.485 | 0.445 | 0.442 | 0.433 | 0.471 | 0.445 | 0.493 | 0.489 | 0.497 | 0.462 |
| | 720 | **0.401** | **0.434** | 0.882 | 0.591 | 0.835 | 0.583 | 0.447 | 0.444 | 0.611 | 0.510 | 0.477 | 0.515 | 0.479 | 0.469 | 0.560 | 0.534 | 0.496 | 0.481 |
| | Avg | **0.383** | **0.409** | 0.591 | 0.468 | 0.589 | 0.491 | 0.417 | 0.419 | 0.478 | 0.440 | 0.428 | 0.441 | 0.442 | 0.435 | 0.495 | 0.491 | 0.463 | 0.447 |
| ETTh2 | 96 | **0.266** | **0.330** | 0.308 | 0.343 | 0.320 | 0.345 | 0.294 | **0.330** | 0.296 | **0.330** | 0.285 | 0.342 | 0.292 | 0.342 | 0.340 | 0.374 | 0.312 | 0.358 |
| | 192 | **0.328** | **0.369** | 0.384 | 0.392 | 0.406 | 0.399 | 0.365 | 0.375 | 0.361 | 0.371 | 0.354 | 0.389 | 0.370 | 0.390 | 0.402 | 0.414 | 0.397 | 0.408 |
| | 336 | **0.357** | **0.394** | 0.429 | 0.430 | 0.492 | 0.453 | 0.376 | 0.390 | 0.390 | 0.390 | 0.373 | 0.407 | 0.382 | 0.408 | 0.452 | 0.452 | 0.435 | 0.440 |
| | 720 | **0.372** | **0.413** | 0.501 | 0.477 | 0.603 | 0.511 | 0.416 | 0.433 | 0.423 | 0.418 | 0.406 | 0.441 | 0.409 | 0.431 | 0.462 | 0.468 | 0.436 | 0.449 |
| | Avg | **0.331** | **0.377** | 0.406 | 0.411 | 0.455 | 0.427 | 0.363 | 0.382 | 0.368 | 0.377 | 0.355 | 0.395 | 0.363 | 0.393 | 0.414 | 0.427 | 0.395 | 0.414 |
| ETTm1 | 96 | **0.291** | **0.343** | 0.454 | 0.408 | 0.457 | 0.403 | 0.363 | 0.356 | 0.380 | 0.361 | 0.292 | 0.346 | 0.313 | 0.352 | 0.338 | 0.375 | 0.344 | 0.373 |
| | 192 | **0.325** | **0.364** | 0.567 | 0.477 | 0.530 | 0.450 | 0.388 | 0.375 | 0.412 | 0.383 | 0.332 | 0.372 | 0.359 | 0.380 | 0.371 | 0.387 | 0.367 | 0.386 |
| | 336 | **0.355** | **0.381** | 0.662 | 0.525 | 0.577 | 0.481 | 0.416 | 0.392 | 0.436 | 0.400 | 0.366 | 0.394 | 0.395 | 0.404 | 0.410 | 0.411 | 0.392 | 0.407 |
| | 720 | **0.406** | **0.411** | 0.900 | 0.591 | 0.660 | 0.526 | 0.460 | 0.418 | 0.462 | 0.420 | 0.417 | 0.421 | 0.449 | 0.440 | 0.478 | 0.450 | 0.464 | 0.442 |
| | Avg | **0.344** | **0.375** | 0.646 | 0.500 | 0.556 | 0.465 | 0.407 | 0.385 | 0.423 | 0.391 | 0.352 | 0.383 | 0.379 | 0.394 | 0.399 | 0.406 | 0.392 | 0.402 |
| ETTm2 | 96 | **0.166** | **0.255** | 0.199 | 0.274 | 0.197 | 0.271 | 0.205 | 0.273 | 0.211 | 0.274 | 0.173 | 0.262 | 0.178 | 0.262 | 0.187 | 0.267 | 0.177 | 0.260 |
| | 192 | **0.220** | **0.292** | 0.261 | 0.322 | 0.254 | 0.314 | 0.275 | 0.316 | 0.281 | 0.318 | 0.229 | 0.301 | 0.243 | 0.304 | 0.249 | 0.309 | 0.246 | 0.305 |
| | 336 | **0.276** | **0.328** | 0.326 | 0.366 | 0.313 | 0.353 | 0.329 | 0.350 | 0.341 | 0.355 | 0.286 | 0.341 | 0.302 | 0.341 | 0.321 | 0.351 | 0.305 | 0.343 |
| | 720 | **0.368** | **0.385** | 0.455 | 0.439 | 0.416 | 0.415 | 0.437 | 0.411 | 0.485 | 0.428 | 0.378 | 0.401 | 0.398 | 0.395 | 0.497 | 0.403 | 0.410 | 0.405 |
| | Avg | **0.258** | **0.315** | 0.310 | 0.350 | 0.295 | 0.338 | 0.312 | 0.338 | 0.330 | 0.344 | 0.267 | 0.326 | 0.280 | 0.326 | 0.314 | 0.333 | 0.285 | 0.328 |
| ECL | 96 | **0.128** | **0.220** | 0.154 | 0.231 | 0.152 | 0.229 | 0.160 | 0.250 | 0.153 | 0.241 | 0.139 | 0.238 | 0.139 | 0.235 | 0.184 | 0.288 | 0.186 | 0.269 |
| | 192 | **0.147** | **0.238** | 0.179 | 0.254 | 0.172 | 0.250 | 0.175 | 0.263 | 0.169 | 0.255 | 0.153 | 0.251 | 0.155 | 0.250 | 0.192 | 0.295 | 0.190 | 0.273 |
| | 336 | **0.161** | **0.255** | 0.214 | 0.284 | 0.203 | 0.276 | 0.187 | 0.277 | 0.187 | 0.273 | 0.169 | 0.266 | 0.170 | 0.268 | 0.200 | 0.303 | 0.206 | 0.290 |
| | 720 | **0.193** | **0.288** | 0.311 | 0.346 | 0.289 | 0.337 | 0.228 | 0.309 | 0.237 | 0.313 | 0.206 | 0.297 | 0.198 | 0.293 | 0.228 | 0.325 | 0.247 | 0.322 |
| | Avg | **0.157** | **0.250** | 0.215 | 0.279 | 0.204 | 0.273 | 0.188 | 0.275 | 0.187 | 0.271 | 0.167 | 0.263 | 0.166 | 0.262 | 0.201 | 0.303 | 0.207 | 0.289 |
| Weather | 96 | **0.153** | **0.201** | 0.203 | 0.238 | 0.194 | 0.235 | 0.220 | 0.217 | 0.199 | 0.211 | 0.162 | 0.212 | 0.156 | 0.202 | 0.169 | 0.228 | 0.177 | 0.218 |
| | 192 | **0.201** | **0.248** | 0.256 | 0.290 | 0.249 | 0.285 | 0.271 | 0.259 | 0.246 | 0.251 | 0.204 | **0.248** | 0.207 | 0.250 | 0.222 | 0.269 | 0.222 | 0.259 |
| | 336 | **0.250** | 0.290 | 0.314 | 0.336 | 0.302 | 0.327 | 0.286 | 0.297 | 0.274 | 0.291 | 0.254 | **0.286** | 0.263 | 0.292 | 0.290 | 0.310 | 0.277 | 0.297 |
| | 720 | **0.319** | **0.337** | 0.397 | 0.396 | 0.372 | 0.378 | 0.373 | 0.354 | 0.337 | 0.340 | 0.326 | 0.337 | 0.340 | 0.341 | 0.376 | 0.364 | 0.352 | 0.347 |
| | Avg | **0.231** | **0.269** | 0.293 | 0.315 | 0.279 | 0.306 | 0.288 | 0.282 | 0.264 | 0.273 | 0.237 | 0.271 | 0.242 | 0.271 | 0.264 | 0.293 | 0.257 | 0.280 |
| Traffic | 96 | **0.362** | **0.251** | - | - | - | - | - | - | - | - | 0.388 | 0.282 | 0.402 | 0.255 | 0.593 | 0.315 | 0.462 | 0.295 |
| | 192 | **0.376** | **0.259** | - | - | - | - | - | - | - | - | 0.407 | 0.290 | 0.426 | 0.268 | 0.596 | 0.317 | 0.466 | 0.296 |
| | 336 | **0.398** | **0.269** | - | - | - | - | - | - | - | - | 0.412 | 0.294 | 0.449 | 0.275 | 0.600 | 0.319 | 0.482 | 0.304 |
| | 720 | **0.429** | **0.284** | - | - | - | - | - | - | - | - | 0.450 | 0.312 | 0.489 | 0.297 | 0.619 | 0.335 | 0.514 | 0.322 |
| | Avg | **0.391** | **0.266** | - | - | - | - | - | - | - | - | 0.414 | 0.295 | 0.442 | 0.274 | 0.602 | 0.322 | 0.481 | 0.304 |

**Table 3**. Ablation analysis on the ETTh1 and ETTh2 datasets

| Variant | ETTh1 | | | | ETTh2 | | | |
|---|---|---|---|---|---|---|---|---|
| | **96** | **192** | **336** | **720** | **96** | **192** | **336** | **720** |
| **w/o TS modal** | 0.358 | 0.389 | 0.402 | 0.425 | 0.276 | 0.332 | 0.359 | 0.382 |
| **w/o VS modal** | 0.355 | 0.389 | 0.412 | 0.441 | 0.273 | 0.336 | 0.359 | 0.386 |
| **w/o projection** | 0.353 | 0.385 | 0.405 | 0.416 | **0.266** | **0.328** | 0.360 | 0.376 |
| **w/ all** | **0.348** | **0.383** | **0.398** | **0.401** | **0.266** | **0.328** | **0.357** | **0.372** |

extracted through the visual modality enhance the ability to recognize long-term patterns in time series, which makes the model performance degrade more slowly in long-sequence forecasting. For example, the MSE loss with the forecasting length of 720 is only increased by 0.75% and 5% compared to the forecasting length of 360 for the ETTh1 and ETTh2 datasets, respectively, showing competitive long-term forecasting capability.

To further analyze the contribution of each module to the predictive performance of VIFO, we conducted ablation experiments as shown in Table 3, which examine the impact of with or without each of those two modality processing modules and the projection layer on performance. From the results, the absence of any module leads to a decrease in the predictive performance of the model, among which the visual module has a relatively greater impact on the forecasting of long-term time series. In addition, the added projection layer can better align information from the two modalities, facilitating subsequent cross-modal information fusion, and thus enhancing predictive performance.

## 4. CONCLUSIONS

This work investigates the application of visual representation of multivariate time series and cross-modal fusion in the field of time series forecasting. Specifically, it explores the discovery of structured spatiotemporal patterns from spatiotemporal visualization graphs. By leveraging the feature extraction capability of large vision model, we constructed VIFO to capture spatiotemporal stable features. The experiments show that VIFO exhibits competitive predictive performance on various time series datasets, while also providing a new feasible solution for the design of variable dependency structures in large time series models.

## 5. REFERENCES


[1] Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, et al., "End-to-end data-driven weather prediction," *Nature*, vol. 641, no. 8065, pp. 1172–1179, 2025.

[2] Tao Wu, Xiangyun Gao, Sufang An, and Siyao Liu, "Time-varying pattern causality inference in global stock markets," *International Review of Financial Analysis*, vol. 77, pp. 101806, 2021.

[3] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.

[4] Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu, "FourierGNN: Rethinking multivariate time series forecasting from a pure graph perspective," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[5] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 22419–22430, 2021.

[6] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," *arXiv preprint arXiv:2210.02186*, 2022.

[7] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *International conference on machine learning*. PMLR, 2022, pp. 27268–27286.

[8] 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.

[9] Luo donghao and wang xue, "ModernTCN: A modern pure convolution structure for general time series analysis," in *The Twelfth International Conference on Learning Representations*, 2024.

[10] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," 2023.

[11] Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao, "MICN: Multi-scale local and global context modeling for long-term series forecasting," in *The Eleventh International Conference on Learning Representations*, 2023.

[12] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen, "Time-LLM: Time series forecasting by reprogramming large language models," in *The Twelfth International Conference on Learning Representations*.

[13] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al., "One fits all: Power general time series analysis by pretrained lm," *Advances in neural information processing systems*, vol. 36, pp. 43322–43355, 2023.

[14] Luke Nicholas Darlow, Qiwen Deng, Ahmed Hassan, Martin Asenov, Rajkarn Singh, Artjom Joosen, Adam Barker, and Amos Storkey, "DAM: Towards a foundation model for forecasting," in *The Twelfth International Conference on Learning Representations*.

[15] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski, "MOMENT: A family of open time-series foundation models," in *Forty-first International Conference on Machine Learning*, 2024.

[16] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou, "A decoder-only foundation model for time-series forecasting," *arXiv preprint arXiv:2310.10688*, 2023.

[17] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann, "UniTime: A language-empowered unified model for cross-domain time series forecasting," in *Proceedings of the ACM on Web Conference 2024*, 2024, pp. 4095–4106.

[18] Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen, "Are language models actually useful for time series forecasting?," *Advances in Neural Information Processing Systems*, vol. 37, pp. 60162–60191, 2024.

[19] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo, "Unified training of universal time series forecasting transformers," *arXiv preprint arXiv:2402.02592*, 2024.

[20] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long, "Timer: Generative pre-trained transformers are large time series models," 2024.

[21] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu, "TEMPO: Prompt-based generative pre-trained transformer for time series forecasting," in *The Twelfth International Conference on Learning Representations*, 2024.

[22] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen, "Time-LLM: Time series forecasting by reprogramming large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[23] Chenxi Sun, Yaliang Li, Hongyan Li, and Shenda Hong, "TEST: Text prototype aligned embedding to activate llm's ability for time series," *arXiv preprint arXiv:2308.08241*, 2023.

[24] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao, "Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment," 2025.

[25] Donghao Luo and Xue Wang, "Moderntcn: A modern pure convolution structure for general time series analysis," in *The twelfth international conference on learning representations*, 2024, pp. 1–43.

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

[27] Yanlong Wang, Jian Xu, Fei Ma, Shao-Lun Huang, Danny Dongning Sun, and Xiao-Ping Zhang, "Psformer: Parameter-efficient transformer with segment attention for time series forecasting," *arXiv preprint arXiv:2411.01419*, 2025.

[28] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," 2025.

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

[30] Juncheng Liu, Chenghao Liu, Gerald Woo, Yiwei Wang, Bryan Hooi, Caiming Xiong, and Doyen Sahoo, "UniTST: Effectively modeling inter-series and intra-series dependencies for multivariate time series forecasting," *Transactions on Machine Learning Research*, 2025.