VSTF: Vision and Sequence Models for Time-Frequency Time Series Forecasting

Junlin Liu [*1 2] **Yanting Yang** [*1 3] **Ren Wang** [1]

Abstract

Recent advancements in time series forecasting have shifted from purely sequential modeling toward incorporating multiple representations. Time series data exhibit two fundamental axes of variation: the interplay between time and frequency domains, and the ability to be represented as either numerical sequences or visual structures. Existing methods typically exploit only one or two of these axes, which leads to systematic blind spots in capturing complex patterns. To formalize a complete modeling basis, we propose the Vision and Sequence Time-Frequency (VSTF) framework, a quad-representation design that integrates Sequence-Time, Vision-Time, Sequence-Frequency, and Vision-Frequency views. Because these distinct modalities capture complementary features, simply ensembling them is sub-optimal. Instead, VSTF employs a cross-modal contrastive learning objective, treating the diverse views as natural augmentations to enforce agreement and extract a unified, noise-resistant latent space. Extensive experiments demonstrate that mapping this complete 2x2 representational basis to a unified latent space consistently outperforms uni-modal and bi-modal baselines in long-term forecasting, while also exhibiting strong generalization to anomaly detection and imputation tasks.

1. Introduction

Our work contributes to the long-term time series forecasting. Despite recent advances in transformer-based models (Nie et al., 2023; Liu et al., 2024) achieving strong performance, most forecasting methods rely on a single input representation, treating the time series as a one-dimensional numerical sequence.

Two principal axes underlie time series modeling: ❶ the domain axis, spanning the time and frequency domains, and ❷ the representation axis, spanning numerical sequences and visual structures. Relying on only one or a limited subset of these perspectives inevitably restricts a model's ability to capture the full complexity of the underlying data-generating process.

As illustrated in **Figure 1**, bi-modal combinations exhibit distinct "blind spots." Specific combinations of modalities often fail to capture morphological edges, lose fine-grained resolution, or fail to track abrupt frequency shifts. There is currently a lack of a framework that leverages the complementary strengths of these axes to close these gaps.

To formalize a complete modeling basis, we propose the Vision and Sequence Time-Frequency (VSTF) framework, a quad-modality architecture that integrates Sequence-Time, Vision-Time, Sequence-Frequency, and Vision-Frequency views. Rather than simply ensembling these views, VSTF aligns them using a cross-modal contrastive learning objective. While the branches capture different patterns, aligning them ensures they map to the same underlying data-generating process, filtering out modality-specific noise.

Extensive experiments demonstrate that VSTF consistently outperforms uni-modal and bi-modal baselines in long-term forecasting. Moreover, the learned representations generalize effectively to downstream tasks such as anomaly detection and imputation.

Our contributions are summarized as follows:

- **Quad-Modality Framework:** We propose VSTF, a unified framework that jointly spans representation and domain axes by integrating sequence and vision models across time and frequency domains.
- **Cross-Modal Contrastive Alignment:** We design a self-supervised fusion strategy that aligns the quad-modality representations via a contrastive learning objective, forming a unified and robust latent embedding.
- **SOTA Performance & Generalization:** Extensive evaluations show that VSTF achieves state-of-the-art (SOTA) results in long-term forecasting and generalizes well to anomaly detection and imputation tasks.

[*]Equal contribution [1]Illinois Institute of Technology [2]Georgetown University [3]CUNY - Baruch College. Correspondence to: Ren Wang <rwang74@illinoistech.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

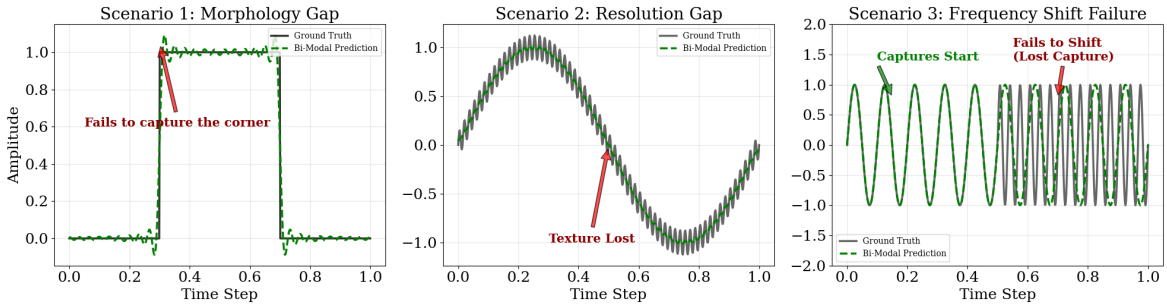

Figure 1. **Bi-Modal Failure Scenarios.** We visualize three limitations where specific bi-modal pairs (dashed green) fail to reconstruct the ground truth (gray). **(Left) Morphology Gap:** *Seq. Time + Seq. Freq.* fails to capture edgy shapes (e.g., sharp corners). **(Middle) Resolution Gap:** *Seq. Time + Vision. Time* loses fine-grained texture due to image downsampling blur. **(Right) Frequency Shift:** Both combinations fail to adapt when periodic patterns shift abruptly, leading to tracking loss.

2. Related Work

2.1. Vision Models for Time Series

Adapting computer vision to time series is a growing trend, comprehensively surveyed by (Ni et al.). Early works like GAF-CNN (Wang & Oates, 2015) encoded series for classification. Modern approaches reshape 1D data into 2D tensors to capture multi-scale dynamics: TimesNet (Wu et al., 2023) models temporal variations via 2D convolutions, while Times2D (Nematirad et al., 2025) extends this with multi-period decomposition and derivative mapping. Similarly, Timemixer++ (Wang et al., 2025) establishes a general pattern machine with time image decomposition. Beyond architectures, VisionTS (Chen et al., 2025) leverages pre-trained Visual MAEs as zero-shot forecasters.

2.2. Frequency Domain in Forecasting

Frequency-domain analysis often outperforms time-domain modeling in handling global seasonality and noise. FEDformer (Zhou et al., 2022b) performed attention via Fourier and Wavelet transforms. Building on this, FilterNet (Yi et al., 2024) proposed learnable frequency filters to selectively pass informative components. FiLM (Zhou et al., 2022a) and ETSformer (Woo et al., 2022) further advanced the field by applying Legendre polynomial projections and frequency-domain exponential smoothing, respectively. Distinct from these approaches which typically operate on global spectral features, we extend the patch-based methodology of PatchTST (Nie et al., 2023) into the frequency domain, enabling the model to capture fine-grained local spectral dependencies that global transforms may overlook.

2.3. Multi-Modality in Time Series

Recent forecasting models have shifted from uni-modal to bi-modal architectures that typically operate along a single interactive axis. Examples include Text-Sequence alignments (Time-LLM (Jin et al., 2024), LLM4TS (Chang et al., 2025)), Time-Frequency dynamics (TFPS (Sun et al., 2025)), and Vision-Sequence combinations (Multi-Modal View (Shen et al., 2025)). While emerging methods like Time-VLM (Zhong et al., 2025) extend this by integrating temporal, visual, and textual modalities, existing approaches still largely isolate different data sources (e.g., text/sequence) or domains (e.g., time/frequency). To unify these perspectives, VSTF introduces the first quad-modal architecture bridging both axes via four distinct views: Sequence-Time, Sequence-Frequency, Vision-Time, and Vision-Frequency.

3. Methodology

By crossing the representation axis, we establish a quad-modality framework that eliminates the "blind spots" of bi-modal approaches. Furthermore, by enforcing a cross-modal contrastive alignment among the disparate representations, we regularize the latent space to reduce noise and enhance the robustness of the learned features. This section details the architecture of the Vision and Sequence Time-Frequency (VSTF) model. **Figure 2** presents the overall architecture.

3.1. Problem Statement and Notations

Let $X_r = \{x_1, x_2, \ldots, x_L\}$ denote a multivariate time series instance, where each $x_t \in \mathbb{R}^C$ represents a vector of C variables at time step t. The look-back window length is L. The goal of the forecasting task is to predict the future sequence $Y = \{x_{L+1}, x_{L+2}, \ldots, x_{L+H}\}$, where H is the prediction horizon. We denote the raw input tensor as $\mathbf{X_r} \in \mathbb{R}^{L \times C}$, and the final forecast as $\hat{\mathbf{Y}} \in \mathbb{R}^{H \times C}$.

Following recent SOTA approaches, we adopt a Channel-Independent strategy. To pre-process the time series, we apply Reversible Instance Normalization (RevIn) to the raw input $\mathbf{X_r}$, yielding the normalized input $\mathbf{X}$.

3.2. Quad-Modality Architecture Overview

The VSTF architecture processes the decomposed components through four specialized parallel branches. These

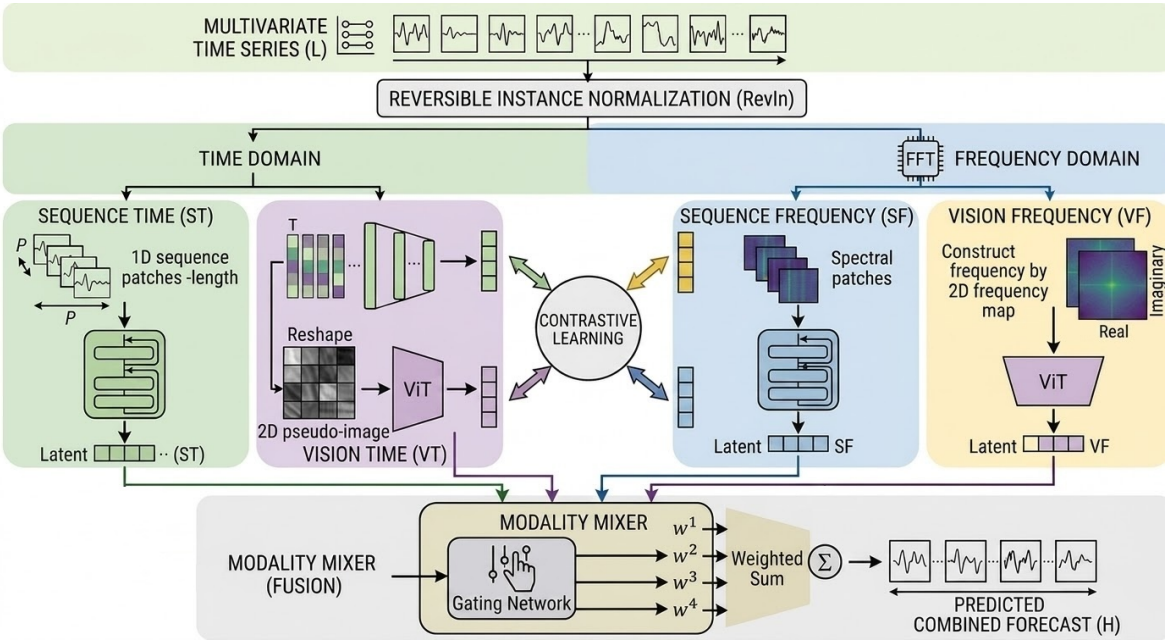

Figure 2. Overall Architecture of VSTF

branches effectively treat the input as either a 1D sequence or a 2D image, operating across both time and frequency domains to capture complementary signal characteristics:

1. **Sequence Time (ST):** Applies a time series encoder to time-domain patches to capture sequential dependencies and local trend patterns.
2. **Sequence Frequency (SF):** Applies a time series encoder to spectral patches, isolating localized, sequential periodicities before reconstruction.
3. **Vision Time (VT):** Reshapes time-domain patches into 2D representations and processes them with a vision encoder to capture global structures and shapelets.
4. **Vision Frequency (VF):** Treats frequency-domain patches (Real and Imaginary parts) as 2-channel images processed by a vision encoder to model global spectral dependencies.

Detailed technical implementations of each modality-specific branch are provided in the Appendix A.

3.3. Cross-Modal Contrastive Alignment

A naive ensembling of the quad-modality branches risks feature collapse and computational redundancy. To align these modalities and extract a robust, unified representation, we implement a cross-modal contrastive learning framework. Inspired by visual representation learning (Chen et al., 2020; Wang et al., 2021; 2022), we treat the SF, VT, and VF branches as representation-specific augmentations of the foundational Sequence-Time (ST) anchor.

Let $\mathbf{z}_{ST}, \mathbf{z}_{SF}, \mathbf{z}_{VT}, \mathbf{z}_{VF}$ denote the L_2-normalized latent representations output by their respective encoder branches. We utilize the normalized temperature-scaled cross-entropy loss (NT-Xent) to maximize agreement between the ST anchor and the other modalities. For a batch of size B, let $s_{i,j} = \mathbf{z}_{ST,i}^{\top}\mathbf{z}_{SF,j}$ denote the cosine similarity between the i-th ST anchor and the j-th SF augmentation. The contrastive loss for the i-th instance is formulated as:

$$\ell_i^{(ST,SF)} = -\log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{B} \mathbb{1}_{[j\neq i]} \exp(s_{i,j}/\tau)}, \qquad (1)$$

where τ is a temperature hyperparameter. The total alignment loss averages the mutual information across all three augmentation pairs $\mathcal{L}_{align} = \frac{1}{3}\sum_{i=1}^{B}\left(\ell_i^{(ST,SF)} + \ell_i^{(ST,VT)} + \ell_i^{(ST,VF)}\right)$.

3.4. Modality Mixer and Final Forecast

To adaptively fuse the representations, a two-layer MLP $\mathcal{G}$ computes dynamic mixing weights $[w^1, w^2, w^3, w^4] = \text{Softmax}\big(\mathcal{G}(\mathbf{X})\big)$. The combined forecast $\mathbf{Y}$ is then constructed as the weighted sum of the four modality outputs $\mathbf{Y} = \sum_{m\in\{ST,SF,VT,VF\}} w^m \mathbf{Y}_m$.

After applying RevIn de-normalization ($\hat{\mathbf{Y}} = \mathbf{Y}\cdot\sigma + \mu$), the model is trained end-to-end using a joint objective that balances task-specific reconstruction with the self-supervised alignment $\mathcal{L}_{total} = \text{MSE}(\hat{\mathbf{Y}}, \mathbf{Y}_{true}) + \alpha\mathcal{L}_{align}$, where α is a scaling coefficient.

Table 1. **Long-term Time Series Multi-Variate Forecasting (LTSF) average performance comparison. Red** values indicate the best MSE and MAE per row.

View		New		Multi-Modal				Visual		Language				Numerical									
Model		VSTF		Time-VLM		Moirai		VisionTS		GPT4TS		Time-LLM		PatchTST		TimesNet		DLinear		FEDformer		FreTS	
Dataset		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	Avg.	**0.392**	0.422	0.405	0.420	0.432	**0.404**	0.407	0.419	0.418	0.421	0.418	0.432	0.427	0.437	0.458	0.450	0.423	0.430	0.440	0.460	0.454	0.468
ETTh2	Avg.	**0.304**	**0.362**	0.341	0.391	0.342	0.372	0.351	0.386	0.354	0.389	0.361	0.396	0.328	0.379	0.414	0.427	0.431	0.447	0.437	0.449	0.451	0.465
ETTm1	Avg.	**0.342**	0.379	0.351	0.376	0.412	0.402	0.344	**0.373**	0.363	0.378	0.356	0.377	0.364	0.391	0.400	0.406	0.357	0.379	0.448	0.452	0.390	0.411
ETTm2	Avg.	**0.248**	**0.310**	0.248	0.311	0.271	0.320	0.267	0.327	0.254	0.311	0.261	0.316	0.284	0.338	0.291	0.333	0.267	0.332	0.305	0.349	0.287	0.339
Electricity	Avg.	**0.157**	0.253	0.172	0.272	0.163	0.255	0.159	**0.250**	0.170	0.263	0.165	0.259	0.168	0.262	0.193	0.295	0.166	0.264	0.214	0.327	0.187	0.296
Weather	Avg.	**0.220**	0.262	0.224	0.263	0.223	0.258	0.225	0.258	0.227	**0.255**	0.244	0.270	0.235	0.271	0.259	0.287	0.249	0.300	0.309	0.360	0.245	0.283
Traffic	Avg.	**0.385**	0.263	0.419	0.304	0.431	0.297	0.386	**0.256**	0.421	0.274	0.422	0.281	0.404	0.280	0.620	0.336	0.434	0.295	0.610	0.376	0.403	0.269
Overall		**0.293**	**0.322**	0.309	0.334	0.325	0.330	0.306	0.324	0.315	0.327	0.318	0.333	0.316	0.337	0.376	0.362	0.332	0.350	0.395	0.396	0.345	0.362

4. Results

4.1. Experimental Setup

We evaluate our method across four time series analysis tasks: multivariate and univariate forecasting, anomaly detection, and imputation. For forecasting, we benchmark using seven standard benchmarks: ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, and Traffic. For forecasting, we adopt a fixed lookback window of $L = 336$ for prediction horizons $T \in \{96, 192, 336, 720\}$. All datasets are split chronologically, utilizing a 60/20/20 partition for ETT datasets and 70/10/20 for the others. We report Mean Squared Error (MSE) and Mean Absolute Error (MAE) as the primary metrics. To ensure the reliability of our findings, our results are averaged over five independent runs.

4.2. SOTA Performance

We evaluate VSTF against diverse SOTA baselines across the seven aforementioned benchmarks. Our comparative analysis encompasses numerical-based methods (e.g., PatchTST, DLinear), language-based multi-modal models (e.g., Time-LLM, GPT4TS), the recent vision-centric model VisionTS, the visual-based multi-modal model Time-VLM, and the large foundational model Moirai.

As shown in **Table 1**, VSTF significantly outperforms competing methods. It achieves superior performance even when compared to VisionTS and Moirai, two highly competitive baselines that benefit from large-scale pretraining. Full comparative results are detailed in Appendix C.1.

4.3. Modality Ablation

To empirically validate the hypothesis that multi-modal integration enhances the representation learning capabilities of the original PatchTST backbone, we conduct ablation studies on the Vision Time (VT), Sequence Frequency (SF), and Vision Frequency (VF) modalities across all four tasks (detailed in Appendices C.2–C.4).

While primarily designed for long-term forecasting, VSTF's quad-modality representation naturally generalizes to other tasks. Because the framework optimizes for robust, general-purpose representation learning rather than purely task-specific objectives, downstream adaptation requires only linear adjustments to the final projection layer. The consistent performance improvements observed across these varied tasks substantiate that integrating auxiliary modalities into the PatchTST backbone establishes a highly generalized and robust architecture for time series analysis.

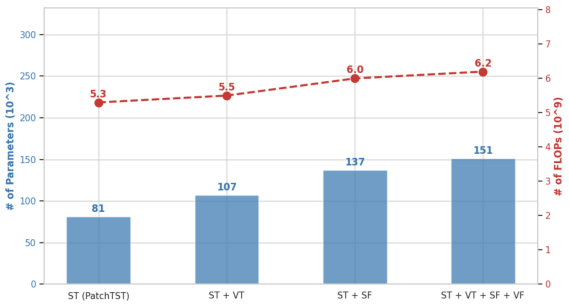

Figure 3. Model Complexity: Parameters vs. FLOPs

Furthermore, an ablation study on the contrastive learning component is provided in Appendix C.5. The results demonstrate that contrastive learning significantly boosts forecasting performance.

4.4. Parameters and FLOPs

As shown in **Figure 3**, VSTF achieves superior results with moderate increases in computation, an 86% increase in trainable parameters and an 17% increase in inference FLOPs relative to the baseline.

5. Conclusion

In this work, we propose VSTF, a novel framework that unifies the representation axis (Sequence vs. Vision) with the domain axis (Time vs. Frequency) through four distinct views: Sequence Time, Sequence Frequency, Vision Time, and Vision Frequency. By aligning these modalities via a contrastive learning objective, it extracts an unified representation. Extensive evaluations across tasks demonstrate that VSTF consistently achieves SOTA performance, and suggest that a complete modeling basis is critical for developing robust, general purpose time series models.

Reproducibility Statement

Our architectures and implementation details can be found in Section 3. The code for reproducing experiments is available at https://github.com/TIML-Group/VSTF-Time-Series-Forecasting.

Impact Statement

This paper contributes to the development of more reliable AI system for temporal data. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

This work was supported in part by the National Science Foundation under grants IIS-2246157, FMitF-2319243, and the Department of Energy under grant DE-CR0000042. The project was also supported by computational resources provided by the NSF ACCESS and Argonne National Lab.

References

Chang, C., Wang, W.-Y., Peng, W.-C., and Chen, T.-F. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–20, 2025.

Chen, M., Shen, L., Li, Z., Wang, X. J., Sun, J., and Liu, C. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. Time-llm: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Nematirad, R., Pahwa, A., and Natarajan, B. Times2d: Multi-period decomposition and derivative mapping for general time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19651–19658, 2025.

Ni, J., Zhao, Z., Shen, C., Tong, H., Song, D., Cheng, W., Luo, D., and Chen, H. Harnessing Vision Models for Time Series Analysis: A Survey.

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

Shen, C., Yu, W., Zhao, Z., Song, D., Cheng, W., Chen, H., and Ni, J. Multi-modal view enhanced large vision models for long-term time series forecasting. 2025.

Sun, Y., Xie, Z., Eldele, E., Chen, D., Hu, Q., and Wu, M. Learning pattern-specific experts for time series forecasting under patch-level distribution shift. 2025.

Wang, R., Xu, K., Liu, S., Chen, P.-Y., Weng, T.-W., Gan, C., and Wang, M. On fast adversarial robustness adaptation in model-agnostic meta-learning. In *International Conference on Learning Representations*, 2021.

Wang, R., Chen, T., Yao, P., Liu, S., Rajapakse, I., and Hero, A. O. Ask: Adversarial soft k-nearest neighbor attack and defense. *IEEE Access*, 10:103074–103088, 2022.

Wang, S., Li, J., Shi, X., Ye, Z., Mo, B., Lin, W., Shengtong, J., Chu, Z., and Jin, M. Timemixer++: A general time series pattern machine for universal predictive analysis. In *ICLR 2025: The Thirteenth International Conference on Learning Representations*, 2025.

Wang, Z. and Oates, T. Spatially encoding temporal correlations to classify temporal data using convolutional neural networks. *arXiv preprint arXiv:1509.07481*, 2015.

Woo, G., Liu, C., Sahoo, D., Kumar, A., and Hoi, S. C. H. Etsformer: Exponential smoothing transformers for time-series forecasting. 2022. URL https://arxiv.org/abs/2202.01381.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023.

Yi, K., Fei, J., Zhang, Q., He, H., Hao, S., Lian, D., and Fan, W. Filternet: Harnessing frequency filters for time series forecasting. *Advances in Neural Information Processing Systems*, 37:55115–55140, 2024.

Zhong, S., Ruan, W., Jin, M., Li, H., Wen, Q., and Liang, Y. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025.

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

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022b.

A. Technical Explanation of Components

In this section, we detail the technical implementations of each modality-specific branch.

Patching is a critical technique to form each branch, as it tokenizes local information, working well for different modalities. Specifically, to capture fine-grained local dependencies beyond raw point-wise observations, we adopt patching mechanism (Nie et al., 2023) across both the time and frequency domains.

Time Domain The normalized time domain input $\mathbf{X} \in \mathbb{R}^{C \times L}$ is patched into N overlapping segments of length P with stride S, resulting in a tensor $\mathbf{X}^p \in \mathbb{R}^{C \times N \times P}$, where $N = \lfloor \frac{L-P}{S} \rfloor + 1$.

Frequency Domain We apply the Fast Fourier Transform (FFT) to generate the frequency domain representation. The frequency dimension F relative to the input length L is $F = \lfloor L/2 \rfloor + 1$ for one-sided spectra. To preserve phase information, we stack the Real and Imaginary parts as below:

$$\mathbf{X}_f = \text{Concat}(\text{Real}(\text{FFT}(\mathbf{X}')), \text{Imag}(\text{FFT}(\mathbf{X}'))), \tag{2}$$

where $\mathbf{X}_f \in \mathbb{R}^{2C \times F}$. The frequency representation is then patched into $\mathbf{X}_f^p \in \mathbb{R}^{2C \times N_f \times P_f}$. Specifically, we apply a frequency patch size of P_f and a stride of S_f. The resulting number of frequency patches is $N_f = \left\lfloor \frac{F-P_f}{S_f} \right\rfloor + 1$.

Branch 1: Sequence Time (ST) The ST branch focuses on temporal modeling. It employs the Time Series Inverted Encoder(TSi Encoder). The patches are fed into the encoder, followed by a Head_{ST} to project the latent features to the prediction horizon:

$$\mathbf{Y}_{ST} = \text{Head}_{ST}(\text{TSi}_{ST}(\mathbf{X}^p)) \in \mathbb{R}^{B \times C \times H} \tag{3}$$

Branch 2: Sequence Frequency (SF) The SF branch models seasonality by leveraging sequential spectral dependencies. We introduce two primary innovations that bridge the gap between traditional signal processing and modern sequence modeling:

- **Frequency-Domain Sequence Patching:** While Sequence Patching is widely utilized in the time domain to capture local temporal patterns, we are the first to adapt this concept to the frequency domain. In traditional spectral analysis, frequencies are often treated as isolated components. However, real-world seasonality consists of clusters of related oscillations. By patching similar frequencies together, the model functions as a frequency detector, allowing it to learn "spectral textures" and relationships between adjacent frequency bins rather than treating them as independent spikes.

- **Invertible Complex-to-Real Transformation:** As defined in (2), To process spectral data within standard neural architectures without losing critical information, we designed a structure to convert the complex-valued $\text{FFT}(X')$ into a real-valued tensor $\mathbf{X}_f$. Rather than discarding phase information (by taking only the magnitude) or increasing the memory footprint, we stack the Real and Imaginary components. This ensures the transformation is completely invertible and memory-efficient, preserving the phase data essential for accurately reconstructing the timing of seasonal peaks.

The frequency patches are processed by a TSi encoder to capture latent spectral patterns. To generate the time-domain forecast of horizon H, we first project the encoder outputs to the required frequency resolution H_f corresponding to H (where $H_f = \lfloor H/2 \rfloor + 1$):

$$\mathbf{Z}_{SF} = \text{Head}_{SF}(\text{TSi}_{SF}(\mathbf{X}_f^p)) \tag{4}$$

where $\mathbf{Z}_{SF} \in \mathbb{R}^{2C \times H_f}$. Here, the projection head maps the latent patch representations back to the full frequency spectrum dimension H_f. The output $\mathbf{Z}_{SF}$ is then split into Real and Imaginary components to form a complex tensor $\mathcal{C}_{SF} \in \mathbb{C}^{C \times H_f}$. Finally, we apply the Inverse FFT (iFFT) to reconstruct the seasonal forecast in the time domain:

$$\mathbf{Y}_{SF} = \text{iFFT}(\mathcal{C}_{SF}) \in \mathbb{R}^{C \times H} \tag{5}$$

Branch 3: Vision Time (VT) The VT branch leverages the Vision Transformer(ViT Encoder). We reshape the time patches $\mathbf{X}^p$ into a 2D image structure suitable for ViT, treating the variable dimension C number of images for univariate processing. Specifically, we treat the time series input as C different grey-scale 1-channel images $\mathbf{X}^I \in \mathbb{R}^{C\times 1\times N\times P}$:

The ViT extracts a global class token, which is projected to the output horizon:

$$\mathbf{Y}_{VT} = \text{Project}(\text{ViT}_{VT}(\mathbf{X}^I)) \in \mathbb{R}^{C\times H} \tag{6}$$

Branch 4: Vision Frequency (VF) The VF branch applies ViT encoder to the spectral domain. Crucially, the Real and Imaginary components of the frequency patches are treated as two distinct image channels $\mathbf{X}_f^I \in \mathbb{R}^{C\times 2\times N_f\times P_f}$.

The ViT processes these dual-channel "patched spectral images." Similar to the SF branch, the output is projected to match the target frequency dimension H_f required for the horizon H. The head outputs $\mathbf{Z}_{VF} \in \mathbb{R}^{2C\times H_f}$, which represents the predicted real and imaginary coefficients. These are combined into complex form and transformed via iFFT to obtain the time-domain forecast $\mathbf{Y}_{VF} \in \mathbb{R}^{C\times H}$.

The details of TSi Encoder and ViT Encoder can be found in below section.

B. Encoder Architectures

To effectively process decomposed signal representations, VSTF leverages two specialized encoder backbones: the Time Series Inverted (TSi) Encoder for sequential dependency modeling and the Vision Transformer (ViT) for high-level structural pattern recognition.

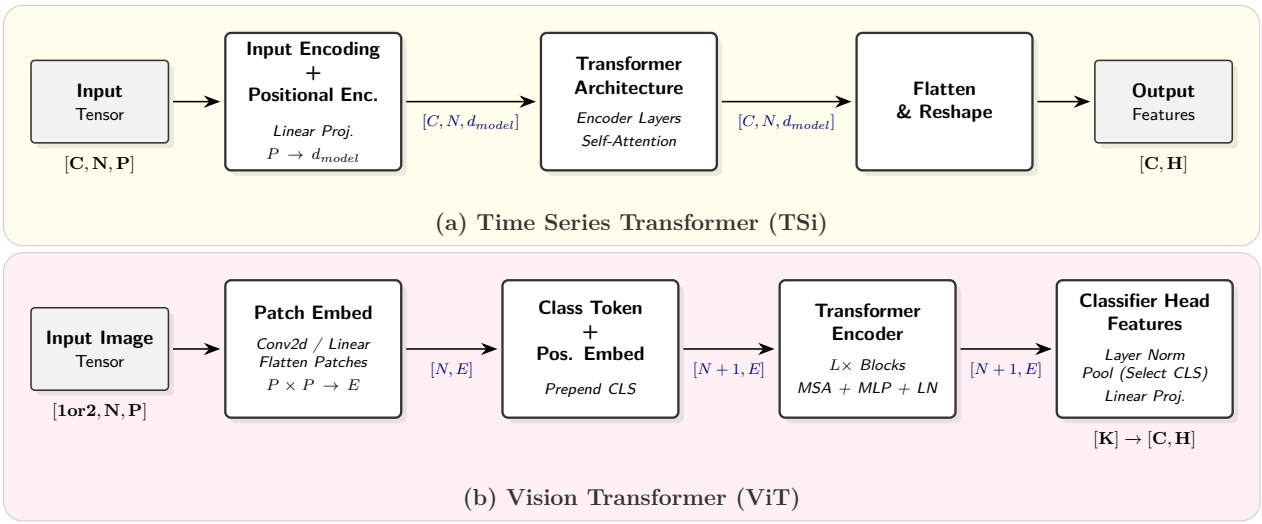

Figure A1. **Encoder Architectures. (a)** The Time Series Transformer (TSi) used for the Sequence branches (ST and SF). It processes patched inputs [C,N,P] using channel-independent modeling to capture sequential dependencies. **(b)** The Vision Transformer (ViT) used for the Vision branches (VT and VF). It treats the input as 2D images—single-channel for the time domain (1) and dual-channel for the frequency domain (2)—utilizing a CLS token to capture global structural patterns.

Time Series Inverted (TSi) Encoder serves as the backbone for the Sequence branches (ST and SF).

- **Input Processing:** The encoder accepts a patch tensor $\mathbf{X} \in \mathbb{R}^{C\times N\times P}$ (where N is the number of patches and P is patch length).
- **Tokenization & Independence:** A linear projection layer $W_P : \mathbb{R}^P \to \mathbb{R}^{d_{model}}$ maps each patch to a latent vector. The tensor is reshaped to $[\mathbf{C}, \mathbf{N}, d_{model}]$, merging the batch and channel dimensions.
- **Backbone:** The core architecture consists of a stack of encoder layers utilizing Multi-Head Self-Attention (MHSA) and Feed-Forward Networks (FFN) with GELU activation. A learnable positional encoding $W_{pos} \in \mathbb{R}^{N\times d_{model}}$ is added to the projected tokens to preserve sequential order.

- **Output:** The output is reshaped back to $[\mathbf{C}, \mathbf{N}, d_{model}]$ before being passed to the FlattenHead, which projects the flattened features directly to the prediction horizon.

Vision Transformer (ViT) Encoder utilizes a standard Vision Transformer to capture global shapelets and frequency patterns by treating time-series patches as 2D images.

- **Image Construction:** Unlike the TSi encoder, the ViT branches reshape the patched input into a "pseudo-image" where the height and width correspond to the number of patches (N) and patch length (P). The input tensor is reshaped to $[\mathbf{C}, C_{in}, \mathbf{N}, \mathbf{P}]$, effectively stacking all variates into the batch dimension.
- **Modality-Specific Channels:**
  - **Time Domain (VT):** The input is treated as a single-channel grayscale image ($C_{in} = 1$). The model utilizes a patch size of (S, S) (where S is the stride) to embed this 2D structure into latent tokens.
  - **Frequency Domain (VF):** The input is treated as a dual-channel image ($C_{in} = 2$), corresponding to the Real ($\Re$) and Imaginary ($\Im$) components of the RFFT spectrum. This allows the ViT to learn complex spectral correlations directly from the frequency map.
- **Classification Token:** A learnable CLS token is prepended to the sequence of image embeddings. After passing through the Transformer encoder blocks, only the CLS state is extracted ($\mathbf{z}_{cls} \in \mathbb{R}^{C \times d_{model}}$) and passed to the projection head to generate the forecast (for VF, predicting both Real and Imaginary parts).

C. Full Experiments

C.1. Comparison with Other Methods

Table A1. **Long-term Time Series Multi-Variate Forecasting (LTSF) performance comparison.** **Red** values indicate the best MSE and MAE per row, and blue values indicate the second best.

View		New		Multi-Modal				Visual		Language				Numerical									
Model		VSTF		Time-VLM		Moirai		VisionTS		GPT4TS		Time-LLM		PatchTST		TimesNet		DLinear		FEDformer		FreTS	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.356	0.389	0.361	0.386	0.384	**0.376**	**0.355**	0.386	0.370	0.389	0.376	0.402	0.372	0.398	0.384	0.402	0.375	0.399	0.376	0.419	0.385	0.410
	192	**0.376**	0.402	0.397	0.415	0.418	**0.394**	0.395	0.407	0.412	0.413	0.407	0.421	0.409	0.418	0.436	0.429	0.405	0.416	0.420	0.448	0.431	0.446
	336	**0.395**	0.421	0.420	0.421	0.439	**0.406**	0.419	0.421	0.448	0.431	0.430	0.438	0.417	0.433	0.491	0.469	0.439	0.416	0.459	0.465	0.457	0.473
	720	**0.441**	0.475	**0.441**	0.458	0.488	**0.440**	0.458	0.460	**0.441**	0.449	0.457	0.468	0.510	0.498	0.521	0.500	0.472	0.490	0.506	0.507	0.541	0.543
	Avg.	**0.392**	0.422	0.405	0.420	0.432	**0.404**	0.407	0.419	0.418	0.421	0.418	0.432	0.427	0.437	0.458	0.450	0.423	0.430	0.440	0.460	0.454	0.468
ETTh2	96	**0.229**	**0.306**	0.267	0.335	0.243	0.307	0.288	0.334	0.280	0.335	0.286	0.346	0.269	0.333	0.340	0.374	0.289	0.353	0.358	0.397	0.267	0.354
	192	**0.281**	**0.341**	0.326	0.373	0.305	0.348	0.349	0.380	0.348	0.380	0.361	0.391	0.324	0.368	0.402	0.414	0.383	0.418	0.429	0.439	0.364	0.420
	336	**0.320**	**0.374**	0.357	0.406	0.347	0.381	0.364	0.398	0.380	0.405	0.390	0.414	0.335	0.387	0.452	0.452	0.448	0.465	0.496	0.487	0.438	0.468
	720	**0.384**	**0.426**	0.412	0.449	0.472	0.453	0.403	0.431	0.406	0.436	0.405	0.434	**0.384**	**0.426**	0.462	0.468	0.605	0.551	0.463	0.474	0.733	0.618
	Avg.	**0.304**	**0.362**	0.341	0.391	0.342	0.372	0.351	0.386	0.354	0.389	0.361	0.396	0.328	0.379	0.414	0.427	0.431	0.447	0.437	0.449	0.451	0.465
ETTm1	96	0.286	0.346	0.304	0.346	0.351	0.370	**0.284**	**0.332**	0.300	0.340	0.291	0.341	0.298	0.348	0.338	0.375	0.299	0.343	0.379	0.419	0.320	0.368
	192	**0.324**	0.368	0.332	0.366	0.388	0.388	0.327	**0.362**	0.343	0.368	0.341	0.369	0.342	0.376	0.374	0.387	0.335	0.365	0.426	0.441	0.368	0.395
	336	**0.354**	0.386	0.364	0.383	0.418	0.405	**0.354**	0.382	0.376	0.386	0.359	**0.379**	0.387	0.408	0.410	0.411	0.369	0.386	0.445	0.459	0.416	0.428
	720	**0.402**	0.414	**0.402**	**0.410**	0.491	0.443	0.411	0.415	0.431	0.416	0.433	0.419	0.430	0.433	0.478	0.450	0.425	0.421	0.543	0.490	0.457	0.452
	Avg.	**0.342**	0.379	0.351	0.376	0.412	0.402	0.344	**0.373**	0.363	0.378	0.356	0.377	0.364	0.391	0.400	0.406	0.357	0.379	0.448	0.452	0.390	0.411
ETTm2	96	**0.158**	0.250	0.160	0.250	0.170	0.252	0.174	0.262	0.163	0.249	0.162	**0.248**	0.179	0.266	0.187	0.267	0.167	0.260	0.203	0.287	0.174	0.264
	192	**0.215**	**0.290**	**0.215**	0.291	0.229	0.294	0.228	0.297	0.222	0.291	0.235	0.304	0.240	0.309	0.249	0.309	0.224	0.303	0.269	0.328	0.246	0.306
	336	**0.269**	**0.325**	0.270	**0.325**	0.288	0.334	0.281	0.337	0.273	0.327	0.280	0.329	0.304	0.352	0.321	0.351	0.281	0.342	0.325	0.366	0.305	0.345
	720	**0.348**	**0.375**	**0.348**	0.378	0.396	0.398	0.384	0.410	0.357	0.376	0.366	0.382	0.411	0.426	0.408	0.403	0.397	0.421	0.421	0.415	0.422	0.441
	Avg.	**0.248**	**0.310**	**0.248**	0.311	0.271	0.320	0.267	0.327	0.254	0.311	0.261	0.316	0.284	0.338	0.291	0.333	0.267	0.332	0.305	0.349	0.287	0.339
Electricity	96	0.128	0.222	0.142	0.245	0.134	0.226	**0.127**	**0.217**	0.141	0.239	0.137	0.233	0.137	0.232	0.168	0.272	0.140	0.237	0.193	0.308	0.168	0.280
	192	**0.142**	0.239	0.157	0.260	0.149	0.243	0.148	**0.237**	0.158	0.253	0.152	0.247	0.151	0.249	0.184	0.289	0.153	0.249	0.201	0.315	0.177	0.289
	336	**0.160**	0.259	0.174	0.276	0.165	0.259	0.163	**0.253**	0.172	0.266	0.169	0.267	0.172	0.268	0.198	0.300	0.169	0.267	0.214	0.329	0.188	0.298
	720	**0.199**	0.292	0.214	0.308	0.203	0.292	**0.199**	0.293	0.207	0.293	0.200	**0.290**	0.212	0.300	0.220	0.320	0.203	0.301	0.246	0.355	0.216	0.317
	Avg.	**0.157**	0.253	0.172	0.272	0.163	0.255	0.159	**0.250**	0.170	0.263	0.165	0.259	0.168	0.262	0.193	0.295	0.166	0.264	0.214	0.327	0.187	0.296
Weather	96	0.154	0.205	0.148	0.200	**0.142**	0.191	0.146	0.191	0.148	**0.188**	0.155	0.199	0.156	0.202	0.172	0.220	0.176	0.237	0.217	0.296	0.168	0.205
	192	0.199	0.245	0.193	0.240	0.193	0.239	0.194	0.238	**0.192**	**0.230**	0.223	0.261	0.203	0.248	0.219	0.261	0.220	0.282	0.276	0.336	0.209	0.270
	336	**0.228**	**0.273**	0.243	0.281	0.243	**0.273**	0.243	0.275	0.246	**0.273**	0.251	0.279	0.253	0.287	0.280	0.306	0.265	0.319	0.339	0.380	0.255	0.315
	720	**0.300**	**0.324**	0.312	0.332	0.312	0.328	0.318	0.328	0.320	0.328	0.345	0.342	0.327	0.345	0.365	0.359	0.333	0.362	0.403	0.428	0.346	0.340
	Avg.	**0.220**	0.262	0.224	0.263	0.223	0.258	0.225	0.258	0.227	**0.255**	0.244	0.270	0.235	0.271	0.259	0.287	0.249	0.300	0.309	0.360	0.245	0.283
Traffic	96	0.358	0.248	0.393	0.290	0.400	0.283	**0.346**	**0.232**	0.396	0.264	0.392	0.267	0.387	0.275	0.593	0.321	0.410	0.282	0.587	0.366	0.381	0.257
	192	**0.366**	0.253	0.405	0.296	0.420	0.289	0.376	**0.245**	0.412	0.268	0.409	0.271	0.381	0.262	0.617	0.336	0.423	0.287	0.604	0.373	0.396	0.266
	336	**0.384**	0.262	0.420	0.305	0.435	0.299	0.389	**0.252**	0.421	0.273	0.434	0.296	0.405	0.282	0.629	0.336	0.436	0.296	0.621	0.383	0.405	0.271
	720	**0.430**	0.289	0.459	0.323	0.470	0.316	0.432	0.293	0.455	0.291	0.451	0.291	0.444	0.301	0.640	0.350	0.466	0.315	0.626	0.382	**0.430**	**0.283**
	Avg.	**0.385**	0.263	0.419	0.304	0.431	0.297	0.386	**0.256**	0.421	0.274	0.422	0.281	0.404	0.280	0.620	0.336	0.434	0.295	0.610	0.376	0.403	0.269
Overall		**0.293**	**0.322**	0.309	0.334	0.325	0.330	0.306	0.324	0.315	0.327	0.318	0.333	0.316	0.337	0.376	0.362	0.332	0.350	0.395	0.396	0.345	0.362

C.2. Ablation Study on Multivariate Long-Term Forecasting

Table A2. **Ablation Study on Multivariate Long-Term Forecasting.** Comparison of Sequence Time (PatchTST) against bi-modal and quad-modal variants. Red indicates the best result (lowest error) and Blue indicates the second best.

Dataset	Horizon	ST(PatchTST) MSE	ST(PatchTST) MAE	ST + VT MSE	ST + VT MAE	ST + SF MSE	ST + SF MAE	ST + VT + SF + VF MSE	ST + VT + SF + VF MAE
ETTh1	96	0.372	0.398	0.361	0.392	0.362	0.390	0.358	0.388
	192	0.409	0.418	0.401	0.414	0.400	0.412	0.399	0.412
	336	0.417	0.433	0.409	0.429	0.403	0.427	0.407	0.426
	720	0.510	0.498	0.486	0.483	0.469	0.477	0.457	0.465
	Avg.	0.427	0.437	0.414	0.430	0.408	0.426	0.405	0.423
ETTh2	96	0.269	0.333	0.279	0.339	0.266	0.332	0.280	0.338
	192	0.324	0.368	0.333	0.377	0.326	0.371	0.340	0.380
	336	0.335	0.387	0.327	0.382	0.326	0.381	0.333	0.386
	720	0.384	0.426	0.395	0.429	0.390	0.429	0.395	0.430
	Avg.	0.328	0.378	0.334	0.382	0.327	0.378	0.337	0.384
ETTm1	96	0.298	0.348	0.287	0.342	0.295	0.346	0.290	0.345
	192	0.342	0.376	0.335	0.373	0.334	0.373	0.331	0.371
	336	0.387	0.408	0.365	0.394	0.368	0.394	0.361	0.391
	720	0.430	0.433	0.414	0.423	0.421	0.427	0.416	0.425
	Avg.	0.364	0.391	0.350	0.383	0.354	0.385	0.350	0.383
ETTm2	96	0.179	0.266	0.174	0.261	0.171	0.259	0.177	0.262
	192	0.240	0.309	0.237	0.304	0.231	0.301	0.241	0.305
	336	0.304	0.352	0.295	0.344	0.291	0.345	0.292	0.342
	720	0.411	0.426	0.370	0.393	0.376	0.399	0.367	0.393
	Avg.	0.284	0.338	0.269	0.326	0.267	0.326	0.269	0.326
Weather	96	0.155	0.204	0.152	0.200	0.154	0.204	0.153	0.202
	192	0.205	0.251	0.197	0.244	0.197	0.244	0.194	0.241
	336	0.254	0.287	0.251	0.284	0.250	0.285	0.250	0.284
	720	0.325	0.341	0.328	0.338	0.322	0.336	0.326	0.336
	Avg.	0.235	0.271	0.232	0.266	0.231	0.267	0.231	0.266
Electricity	96	0.137	0.232	0.131	0.225	0.135	0.230	0.129	0.223
	192	0.151	0.249	0.144	0.240	0.149	0.246	0.144	0.239
	336	0.172	0.268	0.164	0.259	0.167	0.263	0.164	0.259
	720	0.212	0.300	0.203	0.293	0.209	0.298	0.202	0.293
	Avg.	0.168	0.262	0.160	0.254	0.165	0.259	0.160	0.254
Traffic	96	0.404	0.293	0.379	0.263	0.396	0.284	0.359	0.250
	192	0.387	0.267	0.379	0.259	0.397	0.280	0.370	0.253
	336	0.405	0.282	0.402	0.275	0.401	0.278	0.385	0.263
	720	0.444	0.301	0.440	0.296	0.439	0.297	0.431	0.289
	Avg.	0.410	0.286	0.400	0.273	0.408	0.285	0.386	0.264
Overall		0.316	0.338	0.309	0.331	0.309	0.332	0.305	0.328

C.3. Ablation Study on Anomaly Detection

Table A3. **Anomaly detection results (P, R, and F1 %) on five benchmark datasets**. Red indicates the best result for each dataset.

Datasets	SMD			MSL			SMAP			SWaT			PSM			Avg F1
Metrics	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	(%)
ST(PatchTST)	87.0	78.5	82.5	86.5	64.3	73.7	90.1	54.1	67.6	90.9	79.7	85.0	99.0	93.1	95.9	80.94
ST + VT	87.0	80.9	83.8	89.2	74.3	81.0	90.0	53.6	67.2	92.2	87.7	89.9	98.8	93.8	96.2	83.62
ST + SF	87.0	80.6	83.7	89.1	74.1	80.9	90.5	53.4	67.2	91.7	85.3	88.4	98.2	92.2	95.1	83.06
ST + VT + SF + VF	87.6	82.7	85.1	89.4	74.8	81.4	90.7	55.1	68.5	92.3	88.5	90.3	98.6	92.9	95.6	84.18

C.4. Ablation Study on Imputation

Table A4. We compare model performance under different mask ratios. **Red** indicates the best performance.

Models		ST(PatchTST)		ST + VT		ST + SF		ST + VT + SF + VF	
Dataset	Mask Ratio	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.125	0.093	0.203	0.082	0.190	0.083	0.192	**0.077**	**0.185**
	0.25	0.106	0.219	0.093	0.201	0.093	0.204	**0.089**	**0.198**
	0.375	0.118	0.230	0.110	0.218	**0.106**	0.218	0.107	**0.217**
	0.5	0.138	0.248	0.134	0.241	**0.126**	**0.237**	0.133	0.241
ETTm1	0.125	0.044	0.136	0.034	0.116	0.036	0.118	**0.033**	**0.114**
	0.25	0.048	0.143	0.038	0.122	0.039	0.123	**0.036**	**0.120**
	0.375	0.052	0.148	0.043	0.131	0.043	0.131	**0.042**	**0.129**
	0.5	0.059	0.157	**0.047**	**0.138**	0.050	0.141	0.050	0.143
ECL	0.125	0.054	0.160	0.044	0.136	0.044	**0.133**	**0.043**	**0.133**
	0.25	0.060	0.169	0.054	0.151	**0.052**	**0.144**	**0.052**	0.148
	0.375	0.069	0.180	0.063	0.163	**0.061**	**0.158**	0.063	0.164
	0.5	0.081	0.194	0.076	0.181	0.074	**0.175**	**0.073**	0.176
Weather	0.125	0.030	0.051	0.028	**0.046**	0.029	0.049	**0.027**	0.052
	0.25	0.032	0.055	0.030	**0.048**	0.031	0.051	**0.029**	0.055
	0.375	0.034	0.058	**0.033**	**0.051**	**0.033**	0.055	**0.033**	0.061
	0.5	0.038	0.062	**0.036**	**0.058**	0.037	0.059	**0.036**	0.064
Avg.	-	0.059	0.150	0.053	0.138	0.054	0.139	**0.052**	**0.137**

C.5. Ablation on Contrastive Learning

Table A5. **Forecasting results comparison.** **Red** values indicate the best MSE and MAE per row.

Model		VSTF w/o Contrast		VSTF w/ Contrast	
Dataset	**Length**	**MSE**	**MAE**	**MSE**	**MAE**
ETTh1	96	0.358	**0.388**	**0.356**	0.389
	192	0.399	0.412	**0.376**	**0.402**
	336	0.407	0.426	**0.395**	**0.421**
	720	**0.457**	**0.465**	0.462	0.475
ETTh2	96	0.280	0.338	**0.229**	**0.306**
	192	0.340	0.380	**0.281**	**0.341**
	336	0.333	0.386	**0.320**	**0.374**
	720	**0.395**	0.430	0.396	**0.429**
ETTm1	96	0.290	**0.345**	**0.288**	0.346
	192	0.331	0.371	**0.325**	**0.368**
	336	0.361	0.391	**0.359**	**0.386**
	720	**0.416**	0.425	0.419	**0.417**
ETTm2	96	0.177	0.262	**0.158**	**0.250**
	192	0.241	0.305	**0.217**	**0.290**
	336	0.292	0.342	**0.270**	**0.325**
	720	0.367	0.393	**0.349**	**0.375**
Weather	96	**0.153**	**0.202**	0.154	0.205
	192	**0.194**	**0.241**	0.199	0.246
	336	0.250	0.284	**0.228**	**0.274**
	720	0.326	0.336	**0.300**	**0.324**
Electricity	96	**0.129**	0.223	**0.129**	**0.222**
	192	0.144	**0.239**	**0.143**	**0.239**
	336	0.164	**0.259**	**0.163**	**0.259**
	720	0.202	0.293	**0.201**	**0.292**
Traffic	96	0.359	0.250	**0.358**	**0.249**
	192	**0.370**	**0.253**	**0.370**	**0.253**
	336	**0.385**	**0.263**	**0.385**	**0.263**
	720	**0.431**	**0.289**	**0.431**	**0.289**