OccamVTS: Distilling Vision Models to 1% Parameters for Time Series Forecasting

Sisuo Lyu[1], Siru Zhong[1], Weilin Ruan[1], Qingxiang Liu[1], Qingsong Wen[2], Hui Xiong[1], Yuxuan Liang[1]*

[1]The Hong Kong University of Science and Technology (Guangzhou), China
[2]Squirrel Ai Learning, USA
{sisuolyu, siruzhong, yuxliang}@outlook.com; xionghui@ust.hk;
{rwlinno, qingxiangliu737, qingsongedu}@gmail.com

Abstract

Time series forecasting is fundamental to diverse applications, with recent approaches leveraging large vision models (LVMs) to capture temporal patterns through visual representations. We reveal that while vision models enhance forecasting performance, 99% of their parameters are unnecessary for time series tasks. Through cross-modal analysis, we find that time series align with low-level textural features but not high-level semantics, which can impair forecasting accuracy. We propose OccamVTS, a knowledge distillation framework that extracts only the essential 1% of predictive information from LVMs into lightweight networks. Using pre-trained LVMs as privileged teachers, OccamVTS employs pyramid-style feature alignment combined with correlation and feature distillation to transfer beneficial patterns while filtering out semantic noise. Counterintuitively, this aggressive parameter reduction improves accuracy by eliminating overfitting to irrelevant visual features while preserving essential temporal patterns. Extensive experiments across multiple benchmark datasets demonstrate that OccamVTS consistently achieves state-of-the-art performance with only 1% of the original parameters, particularly excelling in few-shot and zero-shot scenarios.

Code — https://github.com/sisuolv/OccamVTS
Extended version — https://arxiv.org/abs/2508.01727

1 Introduction

Time series forecasting (TSF) is a fundamental task in machine learning, underpinning a wide range of critical applications, including energy demand prediction, financial market analysis, meteorological modeling, and traffic flow optimization (Idrees, Alam, and Agarwal 2019; Kiyasseh, Zhu, and Clifton 2021; Xu et al. 2021; Bi et al. 2023). The objective of TSF is to anticipate future values based on historical observations of one or more temporally evolving variables. Despite its broad utility, TSF presents significant challenges due to inherent characteristics such as non-stationarity, long-range dependencies, stochastic noise, and the simultaneous presence of localized patterns and global trends across multiple temporal scales.

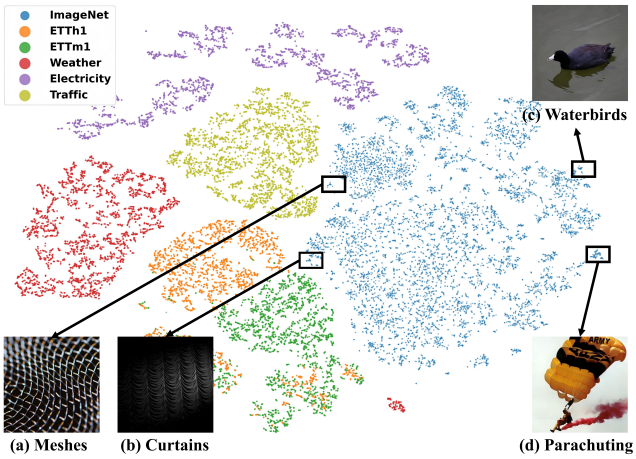


Figure 1: Modality visualization of images (ImageNet) and time series (ECL, Weather, Electricity, ETT) via the MAE encoder. (a)-(d): Original image samples extracted from the corresponding boxes in the t-SNE plot.

Deep learning has emerged as the dominant paradigm for TSF, with state-of-the-art models consistently achieving strong performance across diverse benchmarks. These methods operate directly on numerical sequences or their frequency-domain counterparts, enabling rich representations of complex temporal dynamics. Convolutional neural networks (CNNs) and Transformers, for example, are particularly adept at capturing both short-term fluctuations and long-term dependencies (Ismail Fawaz et al. 2019); advanced models such as Autoformer (Wu et al. 2021) and FEDformer (Zhou et al. 2022) explicitly decompose sequences into trend and seasonal components (Bai, Kolter, and Koltun 2018; Hatami, Gavet, and Debayle 2018; Li et al. 2019; Liu et al. 2022a; Wen et al. 2023). Distinct from other modalities that rely on abstract semantics, these architectures effectively model time series as structured numerical signals, ultimately grounding predictions in precise and interpretable temporal behaviors.

In parallel, vision models have recently gained significant traction as an alternative and intriguing approach to TSF, driven by the observation that humans can often see

*Corresponding author

meaningful patterns in time series plots. For instance, a seasoned bond trader may visually recognize a shift in trend or momentum in a price curve and form predictive judgments based on that perception. This intuitive alignment between visual understanding and temporal inference has inspired researchers to transform time series data into image-like representations, such as spectrograms, recurrence plots, or frequency-enhanced textures (Ni et al. 2025), and directly leverage pre-trained vision models for downstream time series analysis tasks. Representative examples include TimesNet (Wu et al. 2023a), which recasts sequences into two-dimensional formats for CNN processing; VisionTS (Chen et al. 2024), fine-tuning Masked Autoencoders (MAE) on visualized time series (He et al. 2022); and TimeVLM (Zhong et al. 2025), leveraging vision-language models for multimodal forecasting. These approaches repurpose vision models' capability to detect edges, gradients, and frequency textures for identifying temporal structures and dynamics.

However, repurposing large vision models (LVMs) for TSF introduces substantial redundancy and misalignment. These models are originally architected for image data rich in semantic content – a modality that fundamentally diverges from the purely numerical signals of time series. To investigate this misalignment and understand which visual features benefit time series forecasting, we conduct t-SNE visualization of features extracted by MAE from ImageNet (Deng et al. 2009) and four time-series benchmarks (ECL, Weather, Electricity, ETT), as shown in Figure 1. The results reveal that while some visual features exhibit overlap and similarity with time series data, others show significant distributional differences between the two modalities. Fig.1 (a) and (b) lie at the interface of temporal and visual features, exhibiting similar texture patterns, while Fig.1 (c) and (d) appear in distant image clusters, demonstrating richer semantic content. Further analysis uncovers a distinct binary differentiation pattern: time-series features align closely with texture-rich images such as metal meshes and stage curtains, yet diverge markedly from semantically complex scenes like waterbirds and parachuting activities. This pattern suggests that vision models capture low-level textural features relevant to time series, while their high-level semantic representations not only add unnecessary complexity but also impair forecasting by introducing overfitting to irrelevant visual features. In limited-data scenarios, this redundancy becomes particularly detrimental as models latch onto spurious visual patterns instead of essential temporal dynamics.

Two fundamental modality mismatches underlie this problem. The first is the positional sensitivity paradox, which arises since natural images reward translation invariance, whereas time series signals critically depend on absolute temporal position. The second is the semantic representation mismatch, which stems from vision backbones that pursue high-level object semantics absent from purely numerical temporal data. These intrinsic differences manifest as three-dimensional redundancy when deploying vision models for time series forecasting: (1) Computational inefficiency from architectures over-optimized for image resolution; (2) Representational redundancy where most parameters focus on high-level semantic distinction rather than essential temporal patterns; (3) Objective misalignment where classification features directly conflict with regression tasks. This redundancy not only wastes computational resources but also risks negative transfer through semantic noise, interfering with trend prediction. These findings prompt a pressing central question: *How can we retain the useful inductive biases of vision models while eliminating components that are redundant or even detrimental to TSF?*

We answer in the affirmative with **OccamVTS**: a novel cross-modal knowledge-distillation framework that transfers only the most salient **1%** of predictive information from off-the-shelf vision models into compact forecasting networks. Drawing inspiration from Occam's razor, OccamVTS systematically prunes unnecessary complexity while preserving essential core temporal cues. Unlike existing efforts that rely on direct fine-tuning (as in VisionTS) or architectural adaptation (as in TimeVLM), our approach employs native pre-trained LVMs as privileged teachers, guiding lightweight student models via carefully designed pyramid-style feature alignment and selective distillation. This strategy differs in three critical ways: (1) It avoids inheriting architectural constraints by directly distilling from unmodified vision backbones; (2) It introduces a hierarchical pyramid-style alignment mechanism to precisely map spatial features to temporal representations; (3) It explicitly targets redundancy, with empirical results showing that removing up to 99% of vision model parameters not only retains performance, but can actually improve accuracy by systematically mitigating overfitting to irrelevant features. Through this principled simplification, OccamVTS effectively captures frequency textures and gradient-like structures while eliminating semantic noise, ultimately achieving more with less.

2 Related Work

Transforming time series data into visual representations has emerged as a popular paradigm in forecasting research. However, existing approaches have yet to systematically investigate which visual features genuinely benefit forecasting performance and which introduce harmful noise. Early efforts converted sequences into images and trained convolutional neural networks (CNNs) from scratch to perform forecasting tasks (Wang and Oates 2015; Sezer and Ozbayoglu 2018; Li, Kang, and Li 2020; Sood et al. 2021; Semenoglou, Spiliotis, and Assimakopoulos 2023).

As this line of research has progressed, structural-level and model-level innovations have both played key roles. On the structural side, TimesNet decomposes time series into multi-periodic components and rearranges them into 2D tensors (Wu et al. 2023b), while TimeMixer and its successor TimeMixer++ apply mixing operations across time and feature dimensions to capture patterns at multiple scales (Wang et al. 2024a,b). Meanwhile, model-level adaptations harness the power of vision backbones pre-trained on natural images: BEiT was first repurposed for forecasting (Zhou et al. 2023), VisionTS demonstrated the effectiveness of Masked Autoencoders (MAE) on grayscale time-series images (Chen et al. 2024), TimeVLM introduced multimodal forecasting with vision-language models (Zhong

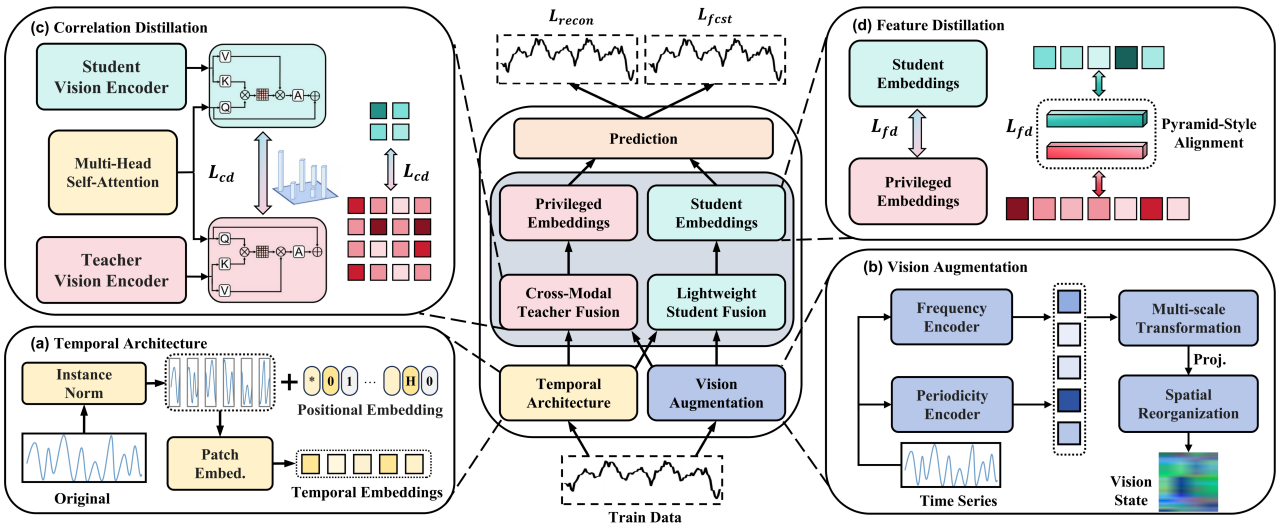


Figure 2: Overview of the OccamVTS framework.

et al. 2025), and LDM4TS leveraged multi-view diffusion models for enhanced prediction (Ruan et al. 2025).

However, off-the-shelf vision backbones remain laden with parameters irrelevant to time-series data: the helpful low-level textures are often diluted by object-centric semantic features. The key challenge is to preserve these predictive cues while simultaneously shedding harmful redundancy. OccamVTS confronts this issue by selectively distilling the crucial 1% of visual knowledge and pruning the remaining 99%, thereby consistently boosting forecasting accuracy.

3 Methodology

To address the severe parameter redundancy when applying vision models to time series forecasting, we propose OccamVTS, a knowledge distillation framework that selectively transfers essential visual knowledge to lightweight temporal models. As illustrated in Figure 2, the framework comprises three core components:

- **Cross-Modal Representation Module.** Extracts temporal features through transformer-based patch embeddings while transforming time series into visual augmentations via multi-scale convolutions and frequency encoding. This dual representation reveals complementary patterns: temporal dependencies in 1D sequences and texture-like patterns in 2D space that enhance forecasting.
- **Teacher-Student Model Module.** Implements an asymmetric design where a frozen pretrained LVM serves as teacher with only its fusion and prediction heads trained to produce privileged supervisory signals. The lightweight student encoder is jointly trained through forecasting and distillation losses to learn cross-modal patterns from the teacher. After training, only the efficient student model is retained for deployment.
- **Knowledge Distillation Module.** Transfers predictive knowledge through pyramid-style feature alignment and selective attention distillation, ensuring the student replicates only forecasting-relevant behaviors. Correlation distillation aligns temporal attention while feature distillation employs a composite objective to match fused representations. Adaptive loss weighting dynamically balances knowledge transfer with prediction accuracy.

3.1 Cross-Modal Representation Module

This module extracts complementary temporal and visual representations through dual parallel pathways. See Appendix K for complete algorithmic details.

Temporal Feature Extraction. Given an input time series $x = [x_1, \ldots, x_T]$, where $x \in \mathbb{R}^{B\times T\times C}$ denotes a batch of multivariate sequences, B is batch size, T sequence length, and C the number of variables. We employ a patch-based transformer architecture. The sequence is segmented into overlapping patches of length L with stride s, each projected into a d-dimensional embedding space with positional encodings. These embeddings are processed through L_{enc} transformer layers:

$$\mathbf{h}^{(\ell)} = \text{TransformerBlock}(\mathbf{h}^{(\ell-1)}), \tag{1}$$

where $\ell = 1, \ldots, L_{\text{enc}}$, $\mathbf{h}^{(0)}$ is the patch embedding sequence, and L_{enc} is the number of encoder blocks. We then pool the T' patch tokens to obtain a sequence-level embedding $\mathbf{h}_{\text{pool}} \in \mathbb{R}^{B\times d_{\text{model}}}$, where T' is the number of temporal patches. This pooled embedding serves as the temporal query in cross-modal fusion.

Visual Augmentation for Time Series. Our analysis reveals that time series data align with texture-rich visual patterns rather than high-level semantic content. This motivates a transformation pipeline that emphasizes low-level visual features to exploit pre-trained vision models:

Pattern Enhancement. We augment the time series with frequency and periodicity features:

$$\mathbf{X}_{\text{aug}} = \text{concat}[\mathbf{x}, \text{FFT}(\mathbf{x}), \text{PE}(\mathbf{x})], \tag{2}$$

where $\mathbf{X}_{\text{aug}} \in \mathbb{R}^{B\times L\times C\times 3}$ stacks the raw window, frequency, and periodic channels for each length-L window, and we instantiate the two operators as

$$\begin{aligned} \text{FFT}(x_{\text{enc}}) &= \left| \sum_{t=0}^{L-1} x_{\text{enc}}(t) \cdot e^{-2\pi i k t/L} \right|, \\ \text{PE}(t) &= \left[\sin\left(\frac{2\pi t}{P}\right), \cos\left(\frac{2\pi t}{P}\right) \right]. \end{aligned} \tag{3}$$

Here $\mathbf{x}_{\text{enc}}$ denotes the normalized length-L window used for spectral analysis, t is the time index within a window, k the frequency index, L the transform length, and P the dataset-specific periodicity hyperparameter (Appendix A). In practice, we use the normalized magnitude of $\text{FFT}(\cdot)$.

Multi-Scale Transformation. The augmented features undergo hierarchical convolutions. First, lightweight depthwise–pointwise 1D blocks to capture short-range local dependencies, then shallow 2D layers on the (channel×time) grid to create spatial patterns. This yields intermediate tensors $\mathbf{F}_{\text{multi-scale}} \in \mathbb{R}^{B\times C\times h\times w}$ before resizing, enabling local variations and global trends to interact as visual textures that pre-trained vision encoders are biased to detect.

Spatial Reorganization. Features are transformed into 2D images through bilinear interpolation and normalized to $[0, 255]$, resulting in $I_{\text{visual}} \in \mathbb{R}^{B\times C\times H_{img}\times W_{img}}$:

$$\mathbf{I}_{\text{visual}} = \text{Normalize}(\text{Interpolate}(\mathbf{F}_{\text{multi-scale}})), \tag{4}$$

where H_{img}, W_{img} denote image height/width (in pixels) after resizing and should not be confused with the forecasting horizon H. Concretely, for a target pixel (x, y) we use bilinear interpolation

$$I(x, y) = \sum_{i=1}^{2}\sum_{j=1}^{2} I(x_i, y_j)\, w_{ij}, \tag{5}$$

where (x_i, y_j) are the four nearest neighbors and w_{ij} are distance-based weights; min-max normalization then scales intensities to the vision backbone's range

$$I_{\text{norm}} = 255 \cdot \frac{I_{\text{raw}} - \min(I_{\text{raw}})}{\max(I_{\text{raw}}) - \min(I_{\text{raw}}) + \varepsilon}. \tag{6}$$

This interpolation-normalization pipeline aligns pseudo-images with pre-trained vision encoders, revealing temporal patterns invisible in 1D sequences while leveraging minimal vision capabilities necessary for time series analysis.

3.2 Teacher-Student Model Design

Cross-Modal Teacher Model. The teacher model combines the temporal features $\mathbf{h}_T$ and visual augmentation I_{visual} from the Cross-Modal Representation Module to produce privileged embeddings for forecasting.

Visual Feature Extraction. The teacher employs frozen large pre-trained vision backbones $\mathcal{V}(\cdot)$ to extract visual features:

$$\mathbf{F}_{\text{vis}}^{T} = \text{GlobalAvgPool}(\mathcal{V}(I_{\text{visual}})) \cdot \mathbf{W}_{\text{proj}}^{v}, \tag{7}$$

where $\text{GlobalAvgPool}(\cdot)$ removes spatial dimensions, $\mathbf{W}_{\text{proj}}^{v}$ projects the backbone output to the fusion dimension, $\mathbf{F}_{\text{vis}}^{T} \in \mathbb{R}^{B\times d_{\text{fus}}}$ and d_{fus} is the fusion dimension for aligning different modalities. We deliberately employ global pooling to prevent noisy spatial alignment between synthetic images and natural-image priors, extracting only aggregate visual features relevant to temporal patterns. Despite containing predominantly redundant parameters for time series tasks, these large models enable the teacher to explore which visual patterns genuinely benefit forecasting.

Cross-Modal Fusion and Privileged Representation Learning. To combine modalities, we apply cross-attention where temporal features query visual representations:

$$\begin{aligned} &\mathbf{Q} = \mathbf{h}_T\,\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{F}_{\text{vis}}^{T}\,\mathbf{W}_K, \quad \mathbf{V} = \mathbf{F}_{\text{vis}}^{T}\,\mathbf{W}_V, \\ &\mathbf{A} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\Big(\frac{\mathbf{Q}\,\mathbf{K}^{\top}}{\sqrt{d_k}}\Big)\,\mathbf{V}. \end{aligned} \tag{8}$$

Here $\mathbf{W}_Q \in \mathbb{R}^{d_{\text{model}}\times d_k}$ and $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{fus}}\times d_k}$ are learned projections, and d_k is the key/query width. The fused representation combines attention output with temporal features:

$$\mathbf{F}_{\text{fus}} = \text{LayerNorm}\big(\mathbf{W}_O\,\mathbf{A} + \mathbf{h}_T\big) \in \mathbb{R}^{B\times d_{\text{fus}}}. \tag{9}$$

Teacher Model Prediction and Supervisory Signal Generation. The teacher produces forecasts through a prediction head:

$$\hat{\mathbf{Y}}_T = \mathbf{W}_{\text{pred}}\mathbf{F}_{\text{fus}} + \mathbf{b}_{\text{pred}}, \tag{10}$$

where $\hat{\mathbf{Y}}_T \in \mathbb{R}^{B\times H\times D}$, H is the forecasting horizon, and D is the number of predicted variables. To ensure high-quality supervisory signals, we optimize:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{SmoothL1}}(\hat{\mathbf{Y}}_T, \mathbf{Y}) \tag{11}$$

where $\mathbf{Y} \in \mathbb{R}^{B\times H\times D}$ denotes ground-truth targets. Beyond predictions, the teacher provides attention matrices $\mathbf{P}_{\text{tea}}$ capturing temporal dependencies, fused representations $\mathbf{F}_{\text{fus}}$, and soft prediction targets for distillation.

Lightweight Student Model. The student processes identical inputs but uses compact vision encoders $\mathcal{V}_{\text{student}}$ achieving substantial parameter reduction:

$$\mathbf{F}_{\text{vis}}^{S} = \mathcal{V}_{\text{student}}(I_{\text{visual}}) \in \mathbb{R}^{B\times d_{\text{vis}}^{S}}, \tag{12}$$

where d_{vis}^{S} is the student's reduced visual feature dimension.

The student employs the same cross-attention fusion mechanism as the teacher but with reduced dimensions. The key innovation is that the student's compact encoder naturally filters out semantic noise, focusing on the minimal set of visual features essential for forecasting.

During training, the student simultaneously optimizes two objectives: forecasting accuracy and knowledge acquisition from the teacher. This design validates that strategic redundancy elimination through knowledge distillation enhances rather than compromises forecasting performance.

3.3 Knowledge Distillation Module

This module precisely orchestrates the selective transfer of predictive knowledge from the teacher to the student, ensuring that only beneficial visual patterns are retained while redundant semantic features are effectively filtered out.

Pyramid-Style Feature Alignment. The dimensional and representational differences between teacher and student models necessitate sophisticated alignment strategies. We introduce a pyramid-style feature alignment mechanism that projects student features through multiple pathways:

$$\mathbf{F}^S_{\text{aligned}} = \sum_{i=0}^{N_s} w_i \, \phi_i\big(\mathbf{F}^S_{\text{fus}}\big), \tag{13}$$

where ϕ_i represents projection functions operating at scale i, N_s is the number of scales, and w_i are learnable weights normalized through softmax. This multi-scale approach enables the student to accurately match teacher representations across multiple different levels of abstraction, from fine-grained local patterns to global temporal trends.

Selective Knowledge Transfer. Our distillation framework employs two complementary mechanisms to transfer essential knowledge while filtering redundancy:

1) *Correlation Distillation.* This component encourages the student to replicate the teacher's temporal dependency patterns. Let $P^{(i)}_{\text{tea}}, P^{(i)}_{\text{stu}} \in \mathbb{R}^{T' \times T'}$ denote the attention matrices for the i-th sample, where T' is the number of temporal patches. We align these matrices via temperature-scaled KL divergence:

$$\mathcal{L}_{\text{cd}} = \frac{\tau^2}{B} \sum_{i=1}^{B} \mathrm{D}_{KL}\left(\sigma\left(\frac{P^{(i)}_{\text{tea}}}{\tau}\right) \middle\| \sigma\left(\frac{P^{(i)}_{\text{stu}}}{\tau}\right)\right), \tag{14}$$

where $\sigma(\cdot)$ denotes the softmax operator and τ is an adaptive temperature parameter that controls the smoothness of attention distributions, flexibly allowing the student to learn both sharp and smoothly distributed attention patterns.

2) *Feature Distillation.* This component aligns the student's fused representations with the teacher's privileged embeddings. We employ a composite loss combining multiple perspectives:

$$\mathcal{L}_{\text{fd}} = \lambda_{\text{MSE}} \cdot \mathcal{L}_{\text{MSE}} + \lambda_{\text{cos}} \cdot \mathcal{L}_{\text{cosine}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}}, \tag{15}$$

where $\mathcal{L}_{\text{MSE}}$ measures direct feature similarity, $\mathcal{L}_{\text{cosine}}$ captures semantic relationships through:

$$\mathcal{L}_{\text{cosine}} = \mathbf{1} - \frac{\mathbf{F}^T_{\text{fus}} \cdot \mathbf{F}^S_{\text{fus}}}{||\mathbf{F}^T_{\text{fus}}|| \cdot ||\mathbf{F}^S_{\text{fus}}||}, \tag{16}$$

and $\mathcal{L}_{\text{KL}}$ effectively aligns output distributions through adaptive temperature scaling while preserving sharpness.

Training Objectives and Optimization. The complete distillation objective combines correlation and feature components:

$$\mathcal{L}_{\text{distill}} = \lambda_{cd}\mathcal{L}_{\text{cd}} + \lambda_{fd}\mathcal{L}_{\text{fd}}, \tag{17}$$

where λ_{cd} and λ_{fd} are implemented as learnable parameters $\boldsymbol{\lambda} = \exp(\boldsymbol{\theta}_\lambda)$, with $\boldsymbol{\theta}_\lambda$ being neural network parameters optimized through gradient descent. This adaptive weighting eliminates manual hyperparameter tuning.

Student's Total Objective. During training, the student minimizes:

$$\mathcal{L}_{\text{student}} = \underbrace{\mathcal{L}_{\text{fcst}}}_{\text{forecasting loss}} + \underbrace{\lambda_{\text{distill}}\,\mathcal{L}_{\text{distill}}}_{\text{distillation alignment loss}}, \tag{18}$$

where $\mathcal{L}_{\text{fcst}} = \mathcal{L}_{\text{SmoothL1}}(\hat{Y}_S, Y)$ is the student's forecasting loss, and λ_{distill} adaptively balances between independent forecasting and cross-modal knowledge acquisition.

Through this carefully designed distillation framework, the student learns to replicate the teacher's beneficial behaviors while discarding redundant features. The adaptive weighting and temperature mechanisms ensure effective knowledge transfer throughout training, enabling the student to achieve comparable or superior performance with dramatically fewer parameters. See Appendix L for details.

3.4 Contributions at a glance

Our contributions are threefold: first, a texture-oriented cross-modal pipeline that converts time series into pseudo-images using frequency and periodicity cues and a lightweight 1D→2D transform, then applies spatial reorganization and global pooling to exploit low-level priors in pre-trained vision backbones; second, an asymmetric teacher–student design that freezes the vision teacher and fuses modalities via cross-attention, letting temporal queries attend to aggregated visual keys and values; third, a selective, scalable knowledge transfer scheme that couples pyramid-style alignment with correlation and representation distillation, so the student preserves forecasting-relevant textures while suppressing semantic redundancy in applications.

4 Experiments

4.1 Experimental Settings

Datasets & Evaluation Metrics. We evaluate OccamVTS on eight benchmark datasets: ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, Traffic (Zhou et al. 2021; Lai et al. 2018), and M4 (Makridakis, Spiliotis, and Assimakopoulos 2018). Performance is measured using MAE and MSE for the first seven datasets, while M4 uses SMAPE, MASE, and OWA following competition protocols (Oreshkin et al. 2019). Dataset details and metric specifications are in Appendix A and B.

Baselines. We extensively compare OccamVTS with state-of-the-art time series models and ablation variants (teacher-only and student-only configurations), including recent vision-augmented methods like TimeVLM (Zhong et al. 2025), LDM4TS (Ruan et al. 2025), TimeMixer++ (Wang et al. 2024b), TimeMixer (Wang et al. 2024a), and TimesNet (Wu et al. 2023a); transformer-based architectures like iTransformer (Liu et al. 2023), PatchTST (Nie et al. 2022), FEDformer (Zhou et al. 2022), ETSformer (Woo et al. 2022), Non-Stationary Transformer (Liu et al. 2022b), Autoformer (Wu et al. 2021), and Informer (Zhou et al. 2021); and highly competitive linear models like DLinear (Zeng et al. 2023) and LightTS (Campos et al. 2023).

Implementation Details. We use a unified benchmarking pipeline (Wu et al. 2023a) with a knowledge distillation framework. From the teacher pool (MAE variants, CLIP, EfficientNet-B3, ResNet-101) and the student pool (EfficientNet-B0, MobileNet-V3, Tiny-ViT), we instantiate one teacher and one student per experiment; unless otherwise noted, we adopt MAE-Large as the teacher and Tiny-ViT as the student, which corresponds to ≈**1%** of the

Methods	Ours		Only Teacher		Only Student		TimeVLM		TimeMixer++		TimeMixer		LDM4TS		TimesNet		iTransformer		DLinear		PatchTST	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTh1$	**0.403**	0.421	0.416	0.433	0.434	0.444	0.405	**0.420**	0.419	0.432	0.447	0.440	0.443	0.454	0.458	0.450	0.454	0.447	0.422	0.437	0.450	0.449
$ETTh2$	**0.336**	0.383	0.338	0.388	0.342	0.394	0.341	0.391	0.339	**0.380**	0.365	0.395	0.387	0.427	0.414	0.427	0.383	0.407	0.431	0.446	0.382	0.411
$ETTm1$	**0.347**	**0.373**	0.354	0.377	0.355	0.377	0.347	0.377	0.369	0.378	0.381	0.396	0.352	0.387	0.400	0.406	0.407	0.410	0.357	0.378	0.388	0.402
$ETTm2$	**0.245**	**0.307**	0.252	0.313	0.258	0.317	0.248	0.311	0.269	0.320	0.275	0.323	0.333	0.380	0.291	0.333	0.288	0.332	0.267	0.333	0.293	0.336
$Weather$	**0.224**	**0.259**	0.229	0.268	0.230	0.269	0.224	0.263	0.226	0.262	0.240	0.272	0.229	0.277	0.259	0.287	0.258	0.278	0.248	0.300	0.258	0.280
ECL	**0.162**	0.259	0.168	0.267	0.170	0.270	0.172	0.272	0.165	**0.253**	0.182	0.273	0.199	0.299	0.192	0.304	0.178	0.270	0.166	0.263	0.204	0.294
$Traffic$	**0.407**	0.279	0.415	0.292	0.419	0.297	0.419	0.298	0.416	**0.264**	0.485	0.298	0.550	0.321	0.620	0.336	0.428	0.282	0.433	0.295	0.482	0.308

Table 1: Long-term forecasting results. Results are averaged over forecasting horizons $H \in \{96, 192, 336, 720\}$. Lower values indicate better performance.**Red**: best, Blue: second best. Full results see Appendix D

Methods	Ours		Only Teacher		Only Student		TimeVLM		TimeMixer++		TimeMixer		LDM4TS		TimesNet		iTransformer		DLinear		PatchTST	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTh1$	**0.422**	**0.439**	0.443	0.456	0.446	0.458	0.431	0.442	0.517	0.512	0.613	0.520	0.471	0.468	0.869	0.628	0.518	0.488	0.691	0.600	0.633	0.542
$ETTh2$	**0.344**	**0.390**	0.356	0.402	0.357	0.403	0.356	0.402	0.379	0.391	0.402	0.433	0.452	0.460	0.479	0.465	0.428	0.438	0.605	0.538	0.415	0.431
$ETTm1$	**0.356**	**0.379**	0.364	0.387	0.365	0.387	0.360	0.382	0.398	0.431	0.487	0.461	0.371	0.393	0.677	0.537	0.447	0.432	0.411	0.429	0.501	0.466
$ETTm2$	**0.253**	**0.313**	0.261	0.321	0.262	0.322	0.263	0.323	0.291	0.351	0.311	0.367	0.336	0.373	0.320	0.353	0.295	0.338	0.316	0.368	0.296	0.343
$Weather$	**0.227**	**0.262**	0.230	0.268	0.231	0.269	0.233	0.274	0.241	0.271	0.242	0.281	0.229	0.276	0.279	0.301	0.272	0.290	0.241	0.283	0.242	0.279
ECL	0.181	0.283	0.206	0.310	0.209	0.312	0.188	0.291	**0.168**	**0.271**	0.187	0.277	0.172	0.275	0.323	0.392	0.202	0.288	0.180	0.280	0.180	0.273
$Traffic$	0.460	0.332	0.531	0.385	0.536	0.390	0.484	0.357	0.483	0.315	0.536	0.349	0.621	0.357	0.951	0.535	0.470	0.318	0.447	0.313	**0.430**	**0.305**

Table 2: Few-shot learning on 10% training data. We use the same protocol in Table 1. Full results see Appendix E

teacher's total parameters. Models are trained using Adam optimizer with learning rate 10^{-3} on NVIDIA RTX A6000 GPU (48GB). See Appendix C for details.

4.2 Long-term Forecasting

We evaluate the long-term forecasting capabilities of our model across seven benchmark datasets and compare against a wide range of state-of-the-art baselines. As shown in Table 1, our approach consistently outperforms state-of-the-art baseline methods across all datasets. On the ETTh2 dataset, OccamVTS achieves a 12.0% MSE reduction compared to PatchTST, demonstrating significant improvements in capturing long-term temporal dependencies. The advantages become more pronounced in high-dimensional scenarios, where we achieve 1.8% improvement over TimeMixer++ on the Electricity dataset and 4.9% improvement over iTransformer on the Traffic dataset. Crucially, the knowledge distillation variant consistently outperforms its non-distilled counterpart across every benchmark, with improvements ranging from 2.3% to 8.1%, demonstrating that knowledge distillation(KD) is a key driver of performance gains. Even without KD, our vision-enhanced architecture achieves second-best performance on multiple datasets, validating the effectiveness of cross-modal temporal modeling.

4.3 Few-shot Forecasting

To evaluate the data efficiency and generalization capability of our proposed method, we conducted comprehensive few-shot learning experiments using only 10% of the training data. As shown in Table 2, our method achieved the best performance on five out of seven datasets (ETTh1, ETTh2, ETTm1, ETTm2, and Weather) in terms of both MSE and MAE metrics. On the ETTh1 dataset, we obtain 2.1% improvement over TimeVLM, while on ETTm2, our approach achieves significant 4% MSE reduction compared to the second-best performer. The ablation experiments in the few-shot setting further validate the importance of the knowledge distillation mechanism, where the complete method consistently outperforms the variant without knowledge distillation by 3.0-4.6% across all datasets. This indicates that the knowledge distillation component is particularly valuable when training data is limited, as it enables more effective cross-modal knowledge transfer from pre-trained vision models to lightweight temporal architectures.

4.4 Zero-shot Forecasting

We conduct zero-shot transfer experiments across ETT datasets without any fine-tuning to evaluate cross-domain generalization capabilities. As shown in Table 3, OccamVTS achieves the best performance in 5 MSE and 7 MAE metrics out of 8 scenarios, demonstrating strong cross-domain transferability. For challenging transfer tasks like `ETTh2->ETTh1` and `ETTm2->ETTm1`, our approach achieves 13.5% and 6.7% MSE improvements over TimeVLM respectively, significantly outperforming other baseline methods. The ablation study confirms that knowl-

Methods	Ours		Only Teacher		Only Student		TimeVLM		TimeMixer++		TimeMixer		LDM4TS		TimesNet		iTransformer		DLinear		PatchTST	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTh1 \rightarrow ETTh2$	0.342	**0.385**	0.351	0.396	0.350	0.395	**0.338**	0.385	0.367	0.391	0.427	0.424	0.458	0.452	0.421	0.431	0.384	0.404	0.493	0.488	0.380	0.405
$ETTh1 \rightarrow ETTm2$	0.295	**0.350**	0.300	0.355	0.301	0.356	**0.293**	0.350	0.301	0.357	0.361	0.397	0.369	0.400	0.327	0.361	0.337	0.374	0.415	0.452	0.314	0.360
$ETTh2 \rightarrow ETTh1$	**0.429**	**0.446**	0.453	0.466	0.532	0.508	0.496	0.480	0.511	0.498	0.679	0.577	0.723	0.577	0.865	0.621	0.657	0.563	0.703	0.574	0.565	0.513
$ETTh2 \rightarrow ETTm2$	**0.285**	**0.343**	0.288	0.346	0.295	0.352	0.297	0.353	0.329	0.370	0.342	0.378	0.432	0.444	0.342	0.376	0.336	0.374	0.328	0.386	0.325	0.365
$ETTm1 \rightarrow ETTh2$	0.357	0.398	0.359	0.400	0.359	0.399	**0.354**	**0.397**	0.417	0.422	0.452	0.441	0.452	0.434	0.457	0.454	0.443	0.443	0.464	0.475	0.439	0.438
$ETTm1 \rightarrow ETTm2$	**0.259**	**0.315**	0.262	0.319	0.263	0.319	0.264	0.319	0.291	0.331	0.329	0.357	0.354	0.367	0.322	0.354	0.301	0.337	0.335	0.389	0.296	0.334
$ETTm2 \rightarrow ETTh2$	**0.357**	**0.394**	0.366	0.402	0.364	0.403	0.359	0.399	0.432	0.443	0.413	0.427	0.494	0.474	0.435	0.443	0.457	0.456	0.455	0.471	0.409	0.425
$ETTm2 \rightarrow ETTm1$	**0.403**	**0.410**	0.451	0.442	0.437	0.434	0.432	0.426	0.427	0.448	0.554	0.478	0.588	0.487	0.769	0.567	0.719	0.546	0.649	0.537	0.568	0.492

Table 3: Zero-shot learning results. We use the same protocol in Table 1. Full results see Appendix F

Methods	Ours	Only Teacher	Only Student	TimeVLM	Timemixer++	Timemixer	TimesNet	iTransformer	DLinear	PatchTST	ETSformer	LightTS	FEDformer	Informer
SMAPE	12.050	12.205	12.222	**11.894**	11.905	11.947	12.880	12.684	13.639	12.059	14.718	13.525	13.160	14.086
MASE	1.611	1.642	1.643	**1.592**	1.611	1.614	1.836	1.764	2.095	1.623	2.408	2.111	1.775	2.718
OWA	0.866	0.879	0.880	**0.855**	0.860	0.862	0.955	0.929	1.051	0.869	1.172	1.051	0.949	1.230

Table 4: Short-term time series forecasting results (Average). The forecasting horizons are in [6, 48] and the results are weighted averaged from all datasets under different sampling intervals. Full results see Appendix G

edge distillation consistently improves zero-shot performance across all transfer scenarios, with particularly notable improvements of 5.4% on `ETTh2->ETTh1` and 10.7% on `ETTm2->ETTm1` when comparing KD versus non-KD variants. While most baseline methods show significant performance degradation (20-40%) in cross-dataset transfers, our method maintains consistent performance with only ±8% variation across different transfer pairs, suggesting robust generalization capabilities through effective distillation of universal temporal patterns from vision models.

4.5 Short-term Forecasting

To evaluate performance on short-term prediction tasks, we conduct experiments with forecasting horizons ranging from 6 to 48 time steps across multiple datasets. As shown in Table 4, our method consistently demonstrates 1.3% improvement in SMAPE compared to traditional approaches and maintains significant consistent advantages over the variant without knowledge distillation, with improvements of 1.4-2% across all metrics. This validates the effectiveness of our knowledge distillation mechanism even in short-term forecasting scenarios, where the distilled visual features help capture subtle fine-grained temporal patterns that might be missed by pure temporal models. The results suggest that OccamVTS maintains remarkably robust temporal modeling capabilities across different prediction horizons, from short-term operational forecasting to long-term strategic planning.

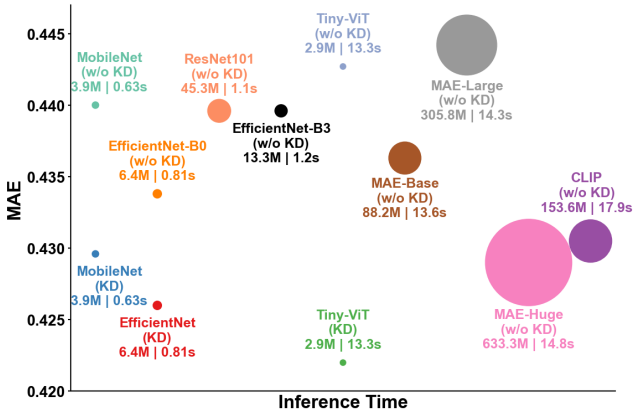


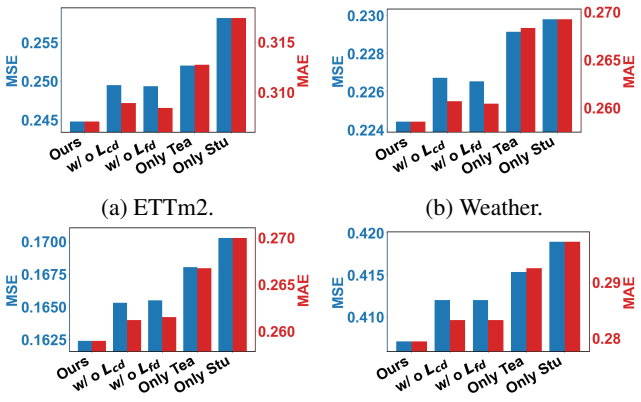
Figure 3: Model Efficiency Comparison, MAE vs Inference Time vs Parameters.

	Full		w/o Vision		w/o Temporal	
Horizon	MSE	MAE	MSE	MAE	MSE	MAE
96	0.150	0.201	0.172	0.225	0.269	0.314
192	0.197	0.245	0.210	0.255	0.293	0.330
336	0.248	0.285	0.252	0.287	0.322	0.348
720	0.323	0.343	0.323	0.338	0.367	0.378
Avg	0.229	0.268	0.239	0.276	0.313	0.343

Table 5: Ablation Study on Multimodal Components on the Weather Dataset, Reporting MSE and MAE.

4.6 Model Analysis

Computational Efficiency Analysis. Figure 3 presents the average MAE and inference time across four forecasting horizons on the ETTh1 dataset, demonstrating the effectiveness of OccamVTS distillation. Student models with knowledge distillation (KD) consistently outperform their non-distilled counterparts (w/o KD) across all architectures. Remarkably, these lightweight student models, using less than 1-2% of teacher model parameters, achieve substantially superior performance compared to massive teacher models.



(a) ETTm2. (b) Weather.

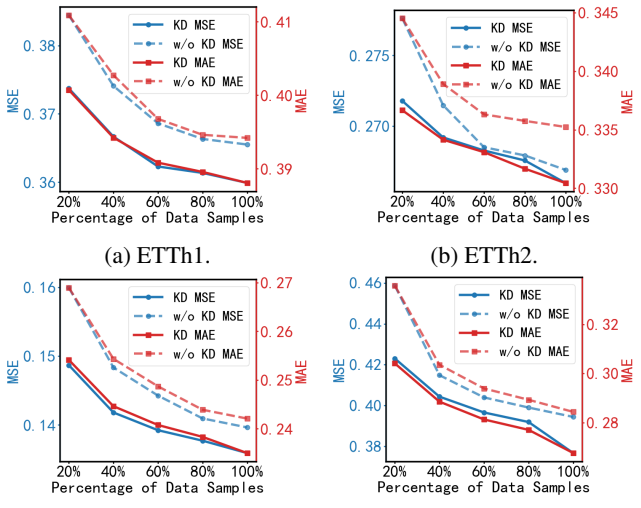
(c) ECL. (d) Traffic.

Figure 4: Ablation Experiment on Four Datasets.

This 99% parameter reduction is achieved while improving accuracy, as OccamVTS effectively extracts essential temporal patterns while filtering out redundant visual features. The results empirically validate our hypothesis that vision models contain substantial redundancy for time series tasks, and OccamVTS can achieve superior performance with minimal computational resources. For comprehensive results across all teacher-student pairings, see Appendix H.

Ablation Study. To validate the effectiveness of each component, we first conduct ablation experiments on the undistilled teacher model. As shown in Table 5, removing the vision component causes a 4.4% MSE degradation on the Weather dataset, confirming that visual features meaningfully enhance temporal modeling. The more severe 36.7% MSE increase when removing temporal components validates that both modalities are essential, with visual augmentation complementing core temporal representations.

Figure 4 analyzes our knowledge distillation framework across four datasets, comparing our complete method against variants without correlation distillation(w/o L_{cd}), without feature distillation(w/o L_{fd}), teacher-only(Only Tea), and student-only(Only Stu) configurations. The results clearly show that our full framework consistently achieves the best performance. Both distillation components contribute meaningfully, while the teacher-only model suffers from interference of redundant parameters in pre-trained vision models, and the student-only model lacks essential cross-modal guidance. These ablation experiments confirm that our architecture benefits from the synergistic combination of all proposed components.



(a) ETTh1. (b) ETTh2.

(c) ECL. (d) Traffic.

Figure 5: Effect of Different Training Data on Four Datasets.

Scalability Study. The scalability analysis on ETTh1, ETTh2, ECL, and Traffic datasets clearly demonstrates OccamVTS's significant advantages under varying data availability. As shown in Figure 5, we evaluate performance with forecasting horizon $H = 96$ as training data increases from 20% to 100%, where the KD version consistently outperforms the non-KD variant across all datasets.

The performance gap is particularly most pronounced under data scarcity (20%-40% training data), where knowledge distillation effectively transfers cross-modal knowledge despite limited samples. On ETTh1 and ETTh2, the MSE/-MAE gap reaches 10-15% in low-data scenarios. ECL and Traffic also show similar patterns, though with higher overall errors. As data availability increases, the performance gap narrows but persists, confirming that knowledge distillation provides substantial benefits even with sufficient data. These results validate OccamVTS's effectiveness, particularly in data-scarce scenarios common in practical applications.

5 Conclusion

We present OccamVTS, a novel knowledge distillation framework that eliminates visual model redundancy in time series forecasting. By distilling only the essential 1% of predictive knowledge from vision models into lightweight student networks, OccamVTS achieves state-of-the-art performance while dramatically reducing computational overhead. Our approach challenges the paradigm of directly deploying massive pre-trained models, demonstrating that strategic knowledge distillation preserves cross-modal benefits while eliminating harmful redundancy. This proves especially valuable in data-scarce scenarios where traditional methods suffer from overfitting. By embodying Occam's razor principle, OccamVTS establishes a new direction for efficient cross-modal time series forecasting.

Future work may explore distilling from foundation models, multi-expert ensembles, other modalities beyond vision, and domain-specific compression ratios. For comprehensive discussion of these directions, see Appendix M.

Acknowledgments

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2025A1515011994, 2023B1515120057). It was also supported by the National Key R&D Program of China (Grant No. 2023YFF0725001); the National Natural Science Foundation of China (Grant Nos. 62402414, 92370204); the

Guangzhou Municipal Science and Technology Project (Grant No. 2023A03J0011); the Guangzhou Industrial Information and Intelligent Key Laboratory Project (Grant No. 2024A03J0628); a grant from the State Key Laboratory of Resources and Environmental Information System; the Guangdong Provincial Key Laboratory of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (Grant No. 2023B1212010007); the Key-Area Special Project of Guangdong Provincial Ordinary Universities (Grant No. 2024ZDZX1007); and the Education Bureau of Guangzhou, which we gratefully acknowledge.

References

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*.

Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; and Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970): 533–538.

Campos, D.; Zhang, M.; Yang, B.; Kieu, T.; Guo, C.; and Jensen, C. S. 2023. LightTS: Lightweight time series classification with adaptive ensemble distillation. *Proceedings of the ACM on Management of Data*, 1(2): 1–27.

Chen, M.; Shen, L.; Li, Z.; Wang, X. J.; Sun, J.; and Liu, C. 2024. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Hatami, N.; Gavet, Y.; and Debayle, J. 2018. Classification of time-series images using deep convolutional neural networks. In *ICMV*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *CVPR*.

Idrees, S. M.; Alam, M. A.; and Agarwal, P. 2019. A prediction approach for stock market volatility based on time series data. *IEEE Access*, 7: 17287–17298.

Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; and Muller, P.-A. 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4): 917–963.

Kiyasseh, D.; Zhu, T.; and Clifton, D. A. 2021. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, 5606–5615. PMLR.

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 95–104.

Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.-X.; and Yan, X. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in neural information processing systems*.

Li, X.; Kang, Y.; and Li, F. 2020. Forecasting with time series imaging. *Expert Systems with Applications*, 160: 113680.

Liu, S.; Yu, H.; Liao, C.; Li, J.; Lin, W.; Liu, A. X.; and Dustdar, S. 2022a. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*.

Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022b. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35: 9881–9893.

Makridakis, S.; Spiliotis, E.; and Assimakopoulos, V. 2018. The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4): 802–808.

Ni, J.; Zhao, Z.; Shen, C.; Tong, H.; Song, D.; Cheng, W.; Luo, D.; and Chen, H. 2025. Harnessing vision models for time series analysis: A survey. *arXiv preprint arXiv:2502.08869*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Oreshkin, B. N.; Carpov, D.; Chapados, N.; and Bengio, Y. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*.

Ruan, W.; Zhong, S.; Wen, H.; and Liang, Y. 2025. Vision-Enhanced Time Series Forecasting via Latent Diffusion Models. *arXiv preprint arXiv:2502.14887*.

Semenoglou, A.-A.; Spiliotis, E.; and Assimakopoulos, V. 2023. Image-based time series forecasting: A deep convolutional neural network approach. *Neural Networks*, 157: 39–53.

Sezer, O. B.; and Ozbayoglu, A. M. 2018. Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. *Applied Soft Computing*, 70: 525–538.

Sood, S.; Zeng, Z.; Cohen, N.; Balch, T.; and Veloso, M. 2021. Visual time series forecasting: an image-driven approach. In *Proceedings of the Second ACM International Conference on AI in Finance*, 1–9.

Wang, S.; Li, J.; Shi, X.; Ye, Z.; Mo, B.; Lin, W.; Ju, S.; Chu, Z.; and Jin, M. 2024b. Timemixer++: A general time series pattern machine for universal predictive analysis. *arXiv preprint arXiv:2410.16032*.

Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and Zhou, J. 2024a. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*.

Wang, Z.; and Oates, T. 2015. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In *AAAI Workshop*.

Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; and Sun, L. 2023. Transformers in time series: a survey. In *IJCAI*.

Woo, G.; Liu, C.; Sahoo, D.; Kumar, A.; and Hoi, S. 2022. Etsformer: Exponential smoothing transformers for time-series forecasting. *arXiv preprint arXiv:2202.01381*.

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

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; et al. 2023b. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *ICLR*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34: 22419–22430.

Xu, J.; Wu, H.; Wang, J.; and Long, M. 2021. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11121–11128.

Zhong, S.; Ruan, W.; Jin, M.; Li, H.; Wen, Q.; and Liang, Y. 2025. Time-VLM: Exploring Multimodal Vision-Language Models for Augmented Time Series Forecasting. *arXiv preprint arXiv:2502.04395*.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on Artificial Intelligence*, 11106–11115.

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

Zhou, T.; Niu, P.; Sun, L.; Jin, R.; et al. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36: 43322–43355.