# EEVEE: Ensemble Expertise from Visual-Temporal-Symbolic Embeddings Enhancing Financial Time Series Forecasting


1st Yue Yuan
*Key Laboratory of Interdisciplinary Research of Computation and Economics*
*Shanghai University of Finance and Economics*
*Advanced Micro Devices, Inc.*
Shanghai, China
yueyuan@stu.sufe.edu.cn

2nd Yahao He
*Advanced Micro Devices, Inc.*
Beijing, China
hyh15@tsinghua.org.cn

3rd Songqiao Han†
*Key Laboratory of Interdisciplinary Research of Computation and Economics*
*Shanghai University of Finance and Economics*
Shanghai, China
han.songqiao@shufe.edu.cn



***Abstract*—While human traders have decoded market movements through candlestick patterns for centuries, modern neural networks often discard this visual wisdom, processing markets purely as numbers. We propose EEVEE, which unifies three complementary perspectives: visual knowledge distilled from commercial LLMs into a vision-language model for K-line patterns; temporal knowledge from synergistically combining DMMV-A and DMMV-S decompositions; and finance-specific knowledge from Kronos with hierarchical coarse–fine token processing. Cross-modal attention fusion yields 64.9 percent directional accuracy and 0.88 percent MAPE on our primary pred20 stock benchmark against eight strong baselines, with additional experiments on stocks and cryptocurrencies.**



***Index Terms*—financial time series forecasting, multi-perspective knowledge fusion, ensemble expertise, visual-temporal-symbolic embeddings, vision-language models, knowledge distillation, cross-modal attention**


## I. Introduction

A curious gap exists between how humans and machines analyze financial markets. Historically, traders have relied on K-line charts—visual representations that encode complex price dynamics into intuitive patterns like "head and shoulders" or "support levels" [1]. This visual approach has proven remarkably robust to noise and scale variations. Yet modern deep learning methods [2]–[5] process market data purely as numerical sequences, discarding these rich visual semantics.

This disconnect is particularly problematic for financial time series, which present unique challenges: (1) *high noise-to-signal ratio* from market microstructure effects, (2) *non-stationarity* with regime changes and structural breaks, (3) *asymmetric patterns* between bull and bear markets, and (4) *multi-scale dependencies* spanning intraday to long-term horizons. These characteristics demand approaches that can leverage multiple complementary views of the data.


†Corresponding author.


Our key insight is that visual and numerical representations capture different but complementary information: numerical data provides precision but is noise-sensitive; visual patterns provide robustness and scale invariance. Can we bridge this gap by integrating multiple visual perspectives with numerical forecasting?

To answer this question, we leverage three distinct knowledge perspectives: **(1) General Visual Knowledge** from vision-language models that understand professional financial chart patterns through knowledge distillation, enabling domain-specific semantic understanding while preserving universal visual reasoning capabilities; **(2) General Time Series Knowledge** from decomposition-based methods that capture fundamental temporal structures applicable across domains, providing robust trend-seasonal separation and multi-scale pattern extraction; and **(3) Finance-Specific Knowledge** from financial foundation models trained on K-line sequences, encoding market-specific dynamics and trading patterns unique to financial data.

We propose **EEVEE** (**E**nsemble **E**xpertise from **V**isual-temporal-symbolic **E**mbeddings **E**nhancing financial forecasting), a multi-perspective knowledge fusion framework that unifies three complementary expertise sources. EEVEE integrates visual knowledge from LLaVA, temporal knowledge from DMMV-AS, and symbolic knowledge from Kronos through cross-modal attention fusion:

1) **General Visual Knowledge via LLaVA Fine-tuning**: We fine-tune LLaVA [6] on K-line charts using knowledge distillation from commercial LLMs, achieving professional-level technical analysis while enabling on-premise deployment for enterprise applications requiring data privacy.
2) **General Time Series Knowledge via DMMV Synergy**: We make a surprising empirical discovery: DMMV-A

and DMMV-S—proposed as *separate model variants* in DMMV [16]—produce strong synergy when combined on financial data, reaching **0.89%** MAPE on pred20 in our stock benchmark (Table V), with over 90% error reduction compared to DMMV-A alone.

3) **Finance-Specific Knowledge via Kronos Integration**: We integrate Kronos [22] with separate processing of its hierarchical coarse and fine tokens from K-line sequences, capturing market-specific dynamics and multi-scale trading patterns.
4) **Cross-Modal Attention Fusion**: We unify these three knowledge perspectives through cross-modal attention, achieving state-of-the-art performance with competitive model size.

**Open Science.** Training code, checkpoints, and experiment configurations will be released under an open license with URL and terms per IEEE and institutional disclosure rules in the camera-ready version.

## II. Related Work

**Time Series Forecasting.** Traditional methods like LSTM [8] have long been used for sequence modeling. Transformer-based models [7] have further advanced time series forecasting: Informer [2] introduces sparse attention, Autoformer [3] proposes auto-correlation, PatchTST [4] applies patching strategies, and TimesNet [5] captures multi-periodicity. However, these methods operate exclusively on numerical sequences, ignoring the rich visual patterns that human traders rely on. **Vision-Language Models.** Vision Transformers [14] have revolutionized image understanding. LLaVA [6] pioneered visual instruction tuning by connecting a CLIP visual encoder to a language model, enabling multimodal reasoning. We fine-tune LLaVA-v1.5-7B on K-line charts via knowledge distillation from Kimi-K2, extracting its visual features as a domain-specific encoder for financial chart analysis. **Multi-Modal Decomposition.** DMMV [16] proposes a decomposition-based framework that generates image representations of time series. It introduces DMMV-A (MAE-based adaptive decomposition [19]) and DMMV-S (spatial moving average decomposition) as separate model variants. Our key discovery is that combining both produces strong synergy in financial data, reaching **0.89%** MAPE on pred20 (Table V) with over 90% error reduction compared to DMMV-A alone. VisionTS [18] demonstrates that visual masked autoencoders can serve as zero-shot forecasters, while Time-VLM [20] explores vision-language models for time series tasks. However, these methods lack domain-specific expertise and multi-perspective knowledge integration. **Financial Foundation Models.** Kronos [22] is a foundation model for K-line sequences with Binary Spherical Quantization (BSQ), converting OHLCV data into hierarchical discrete tokens. We integrate Kronos with hierarchical embedding for multi-scale pattern capture. **Visual Financial Analysis.** Technical analysis [1] documents chart patterns widely used in discretionary trading. Chen and Tsai [21] encode K-line charts for CNN-based classification. Our work bridges visual pattern recognition with numerical forecasting through cross-modal attention fusion.

**Cross-Modal Learning.** Cross-modal meta-learning in other domains (e.g., person retrieval with CAMeL [24]) and decomposition-style forecasting (DUET [25], WPMixer [26]) motivate our fusion under finance noise and non-stationarity; orthogonal-subspace learning with multimodal LLMs [27] parallels limited supervised windows (Sec. IV).

## III. Method

### A. Problem Formulation

Given a historical time series $\mathbf{X} = \{x_1, x_2, \ldots, x_T\} \in \mathbb{R}^{T \times d}$ of length $T$ with $d$ features (e.g., OHLCV data), and a corresponding K-line chart image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, our goal is to predict the future price sequence $\mathbf{Y} = \{y_1, y_2, \ldots, y_P\} \in \mathbb{R}^P$ for the next $P$ time steps.

### B. Overall Architecture

As illustrated in Figure 1, EEVEE consists of four main components:

**(1) General Visual Knowledge (LLaVA)**: Fine-tuned on K-line charts via knowledge distillation from commercial LLMs, providing universal visual understanding of chart patterns applicable across financial instruments.

**(2) General Time Series Knowledge (DMMV-AS)**: Decomposition-based approach capturing universal temporal structures (trends, seasonality, residuals) applicable across time series domains, with DMMV-A and DMMV-S synergistically combined.

**(3) Finance-Specific Knowledge (Kronos)**: K-line symbolic tokenization via Binary Spherical Quantization, encoding market-specific trading dynamics, support/resistance levels, and regime transitions unique to financial markets.

**(4) Cross-Modal Attention Fusion**: Integrates the three knowledge perspectives through multi-head attention mechanism, enabling complementary information exchange for final prediction.

### C. General Visual Knowledge: LLaVA Fine-tuning

While general-purpose vision-language models possess broad visual understanding, they lack knowledge of financial chart patterns. We leverage knowledge distillation [11] from commercial LLMs to create LLaVA-Fin-Kline-v1.5-7B, a domain-specific visual encoder for professional chart analysis.

**1) Instruction Dataset Construction.** We construct our instruction-tuning dataset by sampling 4,421 K-line chart images from the FinMultiTime dataset [23] (810 from HS300, 3,611 from S&P500, spanning 2000-2025), resized to 336×336 pixels. Each chart covers a 6-month semi-annual trading period. Unlike FinMultiTime's original use of 1-5 trend labels, we repurpose these charts for professional technical analysis via knowledge distillation. Each instruction consists of a K-line chart image paired with a structured 5-dimension analysis query covering: (1) Trend Judgment, (2) Pattern Recognition, (3) Volume-Price Analysis, (4) Key Price Levels, and (5) Trading Strategy.

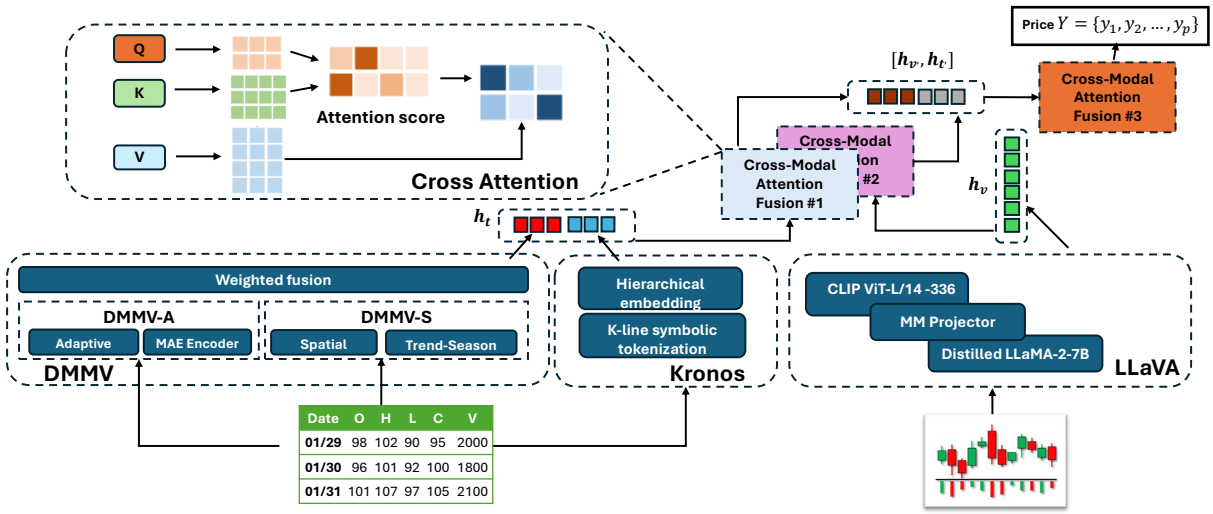


Fig. 1. EEVEE framework overview integrating three knowledge perspectives through cross-modal attention. **Upper**: Overall architecture showing DMMV-AS, Kronos, and LLaVA feeding into bidirectional cross-modal attention for prediction. **Lower**: LLaVA fine-tuning pipeline achieving 92.47/100 quality via knowledge distillation from Kimi-K2.

**2) Multi-Model Response Generation.** To ensure high-quality training data, we compare multiple LLMs for instruction generation:

TABLE I
LLM COMPARISON FOR INSTRUCTION GENERATION (PROF.: PROFESSIONALISM; PREC.: PRECISION; FMT.: FORMAT). BOLD: BEST.

| Model | Prof./25 | Prec./25 | Logic/25 | Fmt./25 | Total/100 |
|---|---|---|---|---|---|
| Kimi-K2 | **24** | **23** | **23** | 21 | **91** |
| GPT-OSS-120B | 21 | 22 | 21 | **22** | 86 |
| Llama-4-Scout | 19 | 17 | 20 | **22** | 78 |

As shown in Table I, Kimi-K2 achieves the highest score (91) due to its superior coverage of technical indicators and professional pattern recognition. We select Kimi-K2 to generate all 4,421 instruction-response pairs. We leverage Kimi-K2 as the teacher to generate training data, then distill this expertise into LLaVA-v1.5-7B through instruction fine-tuning, enabling on-premise deployment for enterprise applications requiring data privacy and cost efficiency.

**3) Quality Verification.** To avoid self-evaluation bias, we employ a different model (Llama-4-Scout 17B) to score the fine-tuned LLaVA outputs on 443 held-out test samples:

**4) Fine-Tuning Configuration.** We fine-tune LLaVA-v1.5-7B using DeepSpeed ZeRO-2: 3,534 training samples, 444 validation samples, effective batch size 128, learning rate 2e-5, 5 epochs, bfloat16 precision, cosine scheduler with 3% warmup. The Vision Tower (CLIP ViT-L/14-336, 304M) is frozen; only the LLM and projector are trained.

**5) Visual Feature Extraction.** We leverage LLaVA [6] to extract semantic features. For image $\mathbf{I}$, we encode it using CLIP's visual encoder [12], pass through LLaVA's projection layer, and apply adaptive pooling to obtain $\mathbf{h}_v \in \mathbb{R}^d$.

TABLE II
FINE-TUNED LLAVA QUALITY (443 TEST SAMPLES; SCORED BY LLAMA-4-SCOUT).

| Dimension | Score/20 | Full Mark Rate |
|---|---|---|
| Completeness | 18.00 | 90.0% |
| Accuracy | 18.89 | 94.5% |
| Professionalism | 19.95 | 99.8% |
| Logic | 18.32 | 91.6% |
| Practicality | 17.24 | 86.2% |
| **Overall** | **92.47/100** | **92.5%** |

### *D. General Time Series Knowledge: DMMV-AS Decomposition*

We build upon the DMMV framework [16], which proposed DMMV-A (backcast-residual based adaptive decomposition) and DMMV-S (spatial moving average decomposition) as **separate model variants** for time series forecasting. Our contribution is demonstrating that **combining** these two mechanisms produces a strong synergistic effect in the financial forecasting domain.

Financial time series contain patterns at multiple scales: long-term trends, medium-term cycles, and short-term fluctuations. The original DMMV work [16] proposed two decomposition mechanisms as separate model alternatives. We describe them briefly and then present our combined approach.

**1) DMMV-A: Backcast-Residual Based Adaptive Decomposition [16].** DMMV-A uses a Masked Autoencoder (MAE) [19] to perform backcast-residual based adaptive decomposition: $\mathbf{H}_A = \text{MAE}_{\text{enc}}(\mathbf{X})$. MAE learns to identify salient patterns by reconstructing masked portions of the time series through backcast-residual learning, enabling it to capture complex temporal dependencies. This data-driven approach is flexible but can be sensitive to noise in financial data.

**2) DMMV-S: Spatial Moving Average Decomposition [16].** DMMV-S applies a period-based moving average filter to extract trend and seasonal components (kernel size $2\lfloor P/2 \rfloor + 1$): $\mathbf{X}_{\text{trend}} = \text{MovingAverage}(\mathbf{X})$ and $\mathbf{X}_{\text{seasonal}} = \mathbf{X} - \mathbf{X}_{\text{trend}}$. We then project the trend component: $\mathbf{H}_S = \text{MLP}(\mathbf{X}_{\text{trend}})$.

**3) Synergistic Combination (Our Discovery).** While the original DMMV work [16] evaluated DMMV-A and DMMV-S as separate model variants, we find that combining them produces a synergistic effect in financial time series. The key insight is that they capture complementary information:

- DMMV-A: Backcast-residual mechanism excels at capturing irregular patterns and long-range dependencies
- DMMV-S: Spatial moving average excels at capturing smooth trends and periodic components

We combine them through learnable weighted fusion:

$$w = \sigma(\alpha), \quad \mathbf{H}_{\text{DMMV}} = w \cdot \mathbf{H}_A + (1 - w) \cdot \mathbf{H}_S \tag{1}$$

where $\alpha$ is a learnable scalar parameter initialized to 0.5.

Our experiments reveal a strong synergistic effect—DMMV-A alone fails catastrophically, DMMV-S alone achieves moderate performance, but their combination reaches **0.89%** MAPE on pred20 (Table V), reducing error by over 90% compared to DMMV-A alone.

### *E. Finance-Specific Knowledge: Kronos Tokenization*

We integrate Kronos [22], a foundation model for financial K-line sequences. Kronos uses **Binary Spherical Quantization (BSQ)** to convert continuous multi-dimensional OHLCV data into hierarchical discrete tokens.

**1) BSQ Tokenization Process.** Given an input K-line sequence $\mathbf{X} \in \mathbb{R}^{T \times 5}$ (OHLCV), Kronos encodes it through a Transformer encoder and quantizes into binary representations via BSQ: $\mathbf{Q} = \text{BSQ}(\text{TransformerEnc}(\mathbf{X})) \in \{-1, +1\}^{T \times 16}$, where we use $s_1 = s_2 = 8$ bits for coarse and fine granularity.

**2) Hierarchical Embedding.** While Kronos provides hierarchical quantized representations (coarse and fine subtokens), we propose to separately process them through dedicated MLPs to better preserve multi-scale information:

$$\mathbf{H}_K = \text{Fusion}([\text{MLP}_1(\mathbf{Q}_{:,:8}); \text{MLP}_2(\mathbf{Q}_{:,8:})]) \tag{2}$$

This preserves the multi-scale nature where coarse bits capture major movements and fine bits capture subtle variations.

### *F. Cross-Modal Attention Fusion*

To integrate the three knowledge perspectives, we employ a bidirectional cross-modal attention mechanism that enables mutual enhancement between visual and temporal modalities.

We first project all features to dimension $d$=256: $\mathbf{h}_{v'} = \text{MLP}_v(\mathbf{h}_v)$, and similarly for $\mathbf{H}_{A'}$, $\mathbf{H}_{S'}$, $\mathbf{H}_{K'}$. Let $\mathbf{h}_t = [\mathbf{H}_{A'}; \mathbf{H}_{S'}; \mathbf{H}_{K'}]$ be the concatenated temporal features. We perform bidirectional cross-attention:

$$\mathbf{h}_{t'} = \mathbf{h}_t + \text{Attn}(\mathbf{h}_t, \mathbf{h}_{v'}, \mathbf{h}_{v'}) \tag{3}$$

$$\mathbf{h}_{v''} = \mathbf{h}_{v'} + \text{Attn}(\mathbf{h}_{v'}, \mathbf{h}_{t'}, \mathbf{h}_{t'}) \tag{4}$$

where $\text{Attn}(Q, K, V)$ denotes multi-head attention with query $Q$, key $K$, and value $V$. Finally, we concatenate $[\mathbf{h}_{t'}; \mathbf{h}_{v''}]$ and apply self-attention via Transformer encoder, then generate predictions via an MLP head with residual from average DMMV outputs. This enables complementary information exchange among all perspectives.

### *G. Training Objective*

In financial applications, correctly predicting the direction of price movement is often more important than minimizing absolute error. We use a composite loss function:

$$\mathcal{L} = (1 - \alpha) \cdot \mathcal{L}_{\text{MAPE}} + \alpha \cdot \mathcal{L}_{\text{Dir}} \tag{5}$$

where the directional loss uses binary cross-entropy on price movement direction. Ablation studies (Table VI) show that $\alpha = 0$ (pure MAPE loss) achieves the best balance.

## IV. Experiments

### *A. Experimental Setup*

**1) Dataset.** We evaluate on Yahoo Finance data spanning 2018–2025. Cryptocurrency (BTC-USD, ETH-USD) differs from equities in volatility and directional bias, so our *primary* benchmark uses five liquid U.S. listings—AAPL, MSFT, GOOGL, AMZN, and SPY—covering large-cap technology names plus ETF SPY as a market baseline. We position this as a *representative* mega-cap slice (not exhaustive), consistent with FinMultiTime [23] chart coverage (HS300 and S&P500 constituents) used for LLaVA instruction data. The choice prioritizes liquid names for stable training signals and reproducibility; it does not claim full market coverage (e.g., small-caps, sector breadth, or non-U.S. listings), which we leave to future work. Each sample uses a 96-day window and K-line charts (336×336); chronological splits are training (1,165), validation (250), and test (490). Table III (pred20) additionally aggregates *Stock*, *Crypto*, and *Stock+Crypto* settings to report robustness under domain shift beyond the five-ticker core table.

**2) Prediction Horizons.** We use **pred5**, **pred20**, and **pred60** (5/20/60-day forecasts from a 96-day input).

**3) Evaluation Metrics.** We report MAPE (lower is better), direction accuracy (higher is better), and IC as the test-split correlation between predictions and targets (higher is better).

**4) Baselines.** Informer [2], Autoformer [3], DLinear [10], PatchTST [4], TimesNet [5], TSMixer [17], iTransformer [9], and DMMV [16].

**5) Implementation Details.** All experiments use PyTorch with Adam optimizer [13] on NVIDIA L40 and AMD MI300X platforms. For price prediction training, we use learning rate 1e-4, batch size 4, and train for 30 epochs with early stopping (patience=5). LLaVA-Fin-Kline-v1.5-7B fine-tuning uses DeepSpeed ZeRO-2 with learning rate 2e-5 and effective batch size 128. During price prediction, LLaVA weights are frozen to preserve learned technical analysis knowledge; Kronos tokenizer weights are also frozen. Only DMMV modules, hierarchical embedding, and fusion layers are trained (9.8M parameters).

TABLE III
AVERAGE PERFORMANCE BY MARKET TYPE (PRED20).

| Method | MAPE(%)↓ | Dir(%)↑ | IC↑ |
|---|---|---|---|
| Informer [2] | 5.18±2.61 | 48.7±11.8 | .969±.032 |
| Autoformer [3] | 4.77±2.36 | 53.8±6.6 | .971±.024 |
| DLinear [10] | 1.56±0.54 | 55.0±3.0 | .990±.009 |
| PatchTST [4] | 4.88±3.63 | 53.3±11.6 | .948±.038 |
| TimesNet [5] | 1.73±0.33 | 50.1±9.6 | .988±.013 |
| TSMixer [17] | 1.46±0.24 | 51.1±8.5 | .989±.014 |
| iTransformer [9] | 15.90±13.15 | 55.4±6.6 | .743±.266 |
| DMMV [16] | 1.16±0.31 | 49.3±0.7 | .992±.009 |
| **EEVEE (Ours)** | **0.99±0.14** | **56.4±7.9** | **.996±.004** |

TABLE IV
ABLATION OVER KNOWLEDGE COMPONENTS (PRED20).

| Configuration | MAPE(%)↓ | Dir(%)↑ | ΔMAPE | ΔDir |
|---|---|---|---|---|
| w/o LLaVA | **0.86** | 48.6 | +0.02 | -16.3pp |
| w/o Kronos | 1.08 | 45.4 | -0.20 | -19.5pp |
| w/o DMMV | 3.10 | 35.1 | -2.22 | -29.8pp |
| Full Model | 0.88 | **64.9** | 0 | 0 |

### B. Main Results

Our multi-view approach targets both low MAPE and high directional accuracy. We first summarize average performance by market type (pred20) in Table III. The full five-ticker benchmark across prediction horizons is given in Table VII at the end of Sec. IV.

Across Stock, Crypto, and Stock+Crypto (pred20), EEVEE retains the strongest MAPE/IC; PatchTST and iTransformer degrade most on mixed data.

### C. Ablation Study

Table IV quantifies each knowledge perspective's contribution ($\Delta\text{MAPE} = \text{MAPE}_{\text{full}} - \text{MAPE}_{\text{ablated}}$; ΔDir in percentage points). **General Time Series Knowledge (DMMV)**: Removing DMMV is most damaging—MAPE rises from 0.88% to 3.10% ($\Delta\text{MAPE} = -2.22$), and direction falls by 29.8pp. **Finance-Specific Knowledge (Kronos)**: Omitting Kronos lowers direction by 19.5pp, showing the value of K-line symbolic tokenization. **General Visual Knowledge (LLaVA)**: Omitting LLaVA lowers direction by 16.3pp while slightly improving MAPE (0.86% vs. 0.88%), a modest MAPE–direction trade-off. The full model reaches 0.88% MAPE and 64.9% direction.

### D. Further Analysis

TABLE V
DMMV-A VS. DMMV-S VS. FUSION (MAPE, %).

| | MAPE(%)↓ | | |
|---|---|---|---|
| **Config** | **pred5** | **pred20** | **pred60** |
| DMMV-A (Attention) | 17.51 | 34.66 | 18.60 |
| DMMV-S (Spatial) | 1.47 | 1.51 | 1.75 |
| **DMMV-AS (Fusion)** | **0.63** | **0.89** | **1.58** |

Table V reveals the second-level fusion gain within the General Time Series Knowledge perspective. DMMV-A (attention-based adaptive) and DMMV-S (spatial smoothing)—originally proposed as *separate variants* in DMMV [16]—achieve strong synergy when fused. DMMV-A alone fails on financial data (17.51–34.66% MAPE), DMMV-S achieves moderate performance (1.47–1.75% MAPE), but their fusion (DMMV-AS) reduces error by over 90% consistently. This demonstrates that combining data-driven adaptive patterns with stable spatial smoothing creates robust temporal representations for the overall framework.

TABLE VI
SENSITIVITY TO DIRECTION-LOSS WEIGHT $\alpha$ (SEE SEC. III-G).

| Direction Weight $\alpha$ | MAPE(%)↓ | Direction(%)↑ |
|---|---|---|
| 0.0 (Pure MAPE Loss) | **0.88** | 64.9 |
| 0.3 | 0.96 | **65.3** |
| 0.5 | 0.92 | 53.1 |
| 0.9 | 0.90 | 52.9 |

Table VI analyzes sensitivity to the direction loss weight $\alpha$. Pure MAPE loss ($\alpha = 0$) achieves the best balance between error (0.88%) and direction (64.9%). While $\alpha = 0.3$ slightly improves direction to 65.3%, MAPE degrades to 0.96%. Higher values (0.5, 0.9) surprisingly degrade both metrics.

Table VII lists all methods on the five-ticker stock benchmark across horizons. EEVEE attains the best MAPE (0.88%) and tied-best direction (64.9%, with PatchTST) on pred20 (about one trading month). Direction on pred5/pred60 is weaker; we emphasize pred20 as the primary operating regime. Average MAPE across horizons is 0.94±0.38%.

**TABLE VII**
Stock benchmark: comparison across prediction horizons (five tickers).

| Method | pred5 | | | pred20 | | | pred60 | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAPE(%)↓ | Dir(%)↑ | IC↑ | MAPE(%)↓ | Dir(%)↑ | IC↑ | MAPE(%)↓ | Dir(%)↑ | IC↑ | MAPE(%)↓ | Dir(%)↑ | IC↑ |
| Informer [2] | 2.51 | 39.8 | .952 | 2.77 | 35.1 | .934 | 3.30 | **72.0** | .912 | 2.86±0.39 | 55.6±18.8 | .933±.020 |
| Autoformer [3] | 2.29 | **58.6** | .960 | 2.50 | 60.8 | .942 | 3.16 | 71.1 | .924 | 2.65±0.40 | **63.5±7.6** | .942±.018 |
| DLinear [10] | 1.15 | 41.8 | .989 | 1.32 | 58.2 | .981 | 1.65 | 48.3 | .973 | 1.37±0.25 | 49.4±9.6 | .981±.008 |
| PatchTST [4] | 2.55 | 44.9 | .948 | 2.65 | **64.9** | .930 | 2.90 | 58.2 | .912 | 2.70±0.17 | 56.0±14.1 | .930±.018 |
| TimesNet [5] | 1.68 | 35.1 | .974 | 1.68 | 39.2 | .974 | 2.91 | 38.8 | .938 | 2.09±0.50 | 37.7±3.4 | .962±.019 |
| TSMixer [17] | 1.46 | 41.0 | .984 | 1.61 | 58.4 | .973 | 2.70 | 46.4 | .952 | 1.92±0.53 | 48.6±9.6 | .970±.017 |
| iTransformer [9] | 5.79 | 29.6 | .897 | 7.17 | 63.3 | .449 | 8.57 | 38.2 | .304 | 6.84±1.25 | 43.7±18.0 | .550±.303 |
| DMMV [16] | 0.63 | 53.1 | .992 | 0.89 | 49.9 | .981 | 1.58 | 49.5 | .974 | 1.03±0.49 | 50.8±2.0 | .982±.009 |
| **EEVEE (Ours)** | **0.59** | 48.0 | **.996** | **0.88** | **64.9** | **.991** | **1.35** | 45.2 | **.986** | **0.94±0.38** | 52.7±10.8 | **.991±.005** |

## V. Conclusion

EEVEE bridges K-line visuals and numeric sequences by fusing LLaVA-based visual knowledge (distilled from Kimi-K2), DMMV-AS temporal decomposition, and Kronos symbolic tokens via cross-modal attention, reaching 0.88% MAPE and 64.9% direction on pred20 with 9.8M trainable parameters.

**Limitations.** LLaVA and Kronos are frozen at prediction time; direction is asymmetric across moves. Evaluation is a representative mega-cap slice (Sec. IV), not a full market cross-section.

**Future Work.** Longer horizons with deeper fusion; LLaVA-Plus-style agents [15] that route DMMV/Kronos by regime; adaptive modality weighting.


## Acknowledgment

The authors thank the anonymous reviewers for their constructive feedback.

This work was supported by the National Natural Science Foundation of China (Nos. 72271151, 72172085, 72442024, 72342009).


*AI disclosure (IEEE):* No AI-generated text was used to draft or substantially revise this article; authors are responsible for all technical content. Sections using AI-generated text would cite the system, per IEEE policy.

## References


[1] J. J. Murphy, "Technical analysis of the financial markets: A comprehensive guide to trading methods and applications," Penguin, 1999.

[2] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," in Proc. AAAI, 2021, pp. 11106–11115, doi: 10.1609/aaai.v35i12.17325.

[3] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting," in Proc. NeurIPS, 2021.

[4] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," in Proc. ICLR, 2023.

[5] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis," in Proc. ICLR, 2023.

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in Proc. NeurIPS, 2023.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," in Proc. NeurIPS, 2017.

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "iTransformer: Inverted Transformers Are Effective for Time Series Forecasting," in Proc. ICLR, 2024.

[10] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are Transformers Effective for Time Series Forecasting?" in Proc. AAAI, 2023, pp. 11121–11128, doi: 10.1609/aaai.v37i9.26317.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv preprint arXiv:1503.02531, 2015.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, PMLR, Jul. 2021, pp. 8748–8763.

[13] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, 2015.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.

[15] S. Liu, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu, L. Zhang, J. Gao, and C. Li, "LLaVA-Plus: Learning to Use Tools for Creating Multimodal Agents," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 126–142, doi: 10.1007/978-3-031-72970-6_8.

[16] C. Shen, W. Yu, Z. Zhao, D. Song, W. Cheng, H. Chen, and J. Ni, "Multi-Modal View Enhanced Large Vision Models for Long-Term Time Series Forecasting," arXiv preprint arXiv:2505.24003, 2025.

[17] S. A. Chen, C. L. Li, N. Yoder, S. O. Arik, and T. Pfister, "TSMixer: An All-MLP Architecture for Time Series Forecasting," arXiv preprint arXiv:2303.06053, 2023.

[18] M. Chen, L. Shen, Z. Li, X. J. Wang, J. Sun, and C. Liu, "VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters," in *Proc. 42nd Int. Conf. Mach. Learn. (ICML)*, vol. 267, PMLR, Jul. 2025, pp. 8979–9007.

[19] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in Proc. CVPR, pp. 16000–16009, 2022.

[20] S. Zhong, W. Ruan, M. Jin, H. Li, Q. Wen, and Y. Liang, "Time-VLM: Exploring Multimodal Vision-Language Models for Augmented Time Series Forecasting," in *Proc. 42nd Int. Conf. Mach. Learn. (ICML)*, vol. 267, PMLR, Jul. 2025, pp. 78478–78497.

[21] J.-H. Chen and Y.-C. Tsai, "Encoding candlesticks as images for pattern classification using convolutional neural networks," *Financial Innovation*, vol. 6, no. 1, Art. no. 26, Jun. 2020, doi: 10.1186/s40854-020-00187-0.

[22] Y. Shi, Z. Fu, S. Chen, B. Zhao, W. Xu, C. Zhang, and J. Li, "Kronos: A Foundation Model for the Language of Financial Markets," in Proc. AAAI, 2026, pp. 25366–25373, doi: 10.1609/aaai.v40i30.39730.

[23] W. Xu, D. Xiang, Y. Liu, X. Wang, Y. Ma, L. Zhang, S. Hu, C. Xu, and J. Zhang, "FinMultiTime: A Four-Modal Bilingual Dataset for Financial Time-Series Analysis," arXiv preprint arXiv:2506.05019, 2025.

[24] H. Yu, J. Wen, and Z. Zheng, "CAMeL: Cross-modality Adaptive Meta-Learning for Text-based Person Retrieval," arXiv preprint arXiv:2504.18782, 2025.

[25] X. Qiu, X. Wu, Y. Lin, C. Guo, J. Hu, and B. Yang, "DUET: Dual Clustering Enhanced Multivariate Time Series Forecasting," in *Proc. 31st ACM SIGKDD Conf. Knowl. Discovery Data Mining (KDD)*, 2025, pp. 1185–1196, doi: 10.1145/3690624.3709325.

[26] M. M. N. Murad, M. Aktukmak, and Y. Yilmaz, "WPMixer: Efficient Multi-Resolution Mixing for Long-Term Time Series Forecasting," in Proc. AAAI, 2025, pp. 19581–19588.

[27] X. Zhou, H. Yu, S. Yang, J. Huo, and P. Tian, "Learning from Orthogonal Space with Multimodal Large Models for Generalized Few-shot Segmentation," ACM Trans. Multimedia Comput. Commun. Appl., vol. 21, no. 11, pp. 309:1–309:22, Nov. 2025, doi: 10.1145/3712597.