Decoupling Trends and Shocks: A Two-Stage Framework for Robust Multimodal Time Series Forecasting

Tsz-Yeung Lau
Department of Computer and Information Systems
University of Aizu
Aizuwakamatsu 965-8580, Japan
m5292005@u-aizu.ac.jp

Konstantin Markov*
Department of Computer and Information Systems
University of Aizu
Aizuwakamatsu 965-8580, Japan
markov@u-aizu.ac.jp

***Abstract*—Recently, time series forecasting has witnessed a paradigm shift towards multimodal approaches that integrate numerical data and text. While recent studies often claim that adding text yields universal performance gains across all fields, our empirical evaluation reveals significant limitations to this assumption. Conventional joint training of numerical and textual data often causes cross-modal interference, leading to substantial prediction errors. In fact, in domains governed by strong periodicities, such as agriculture and the economy, text frequently acts as disruptive noise. To address this, we propose a Two-Stage Orthogonal Optimization framework. In the first stage, a pure numerical model (Mixture of Linear Experts) is trained to capture foundational temporal trends. In the second stage, the numerical backbone is frozen, and text representations are strictly used to predict sudden, unexpected residual shocks. Evaluations across nine diverse datasets demonstrate that our method effectively eliminates the catastrophic errors found in previous baseline models. Ultimately, we establish that textual data is not a universal solution, but rather a specialized modality highly effective for predicting sudden anomalies in event-driven domains like public health and energy.**

***Index Terms*—Time Series Forecasting, Multimodal Fusion, Orthogonal Decoupling, Feature Extraction**

I. Introduction

Time series forecasting (TSF) is a critical component in various real-world applications, spanning from economic planning [1], [2] to traffic management [3], [4]. Recently, the field has witnessed a paradigm shift towards multimodal forecasting, primarily driven by the integration of textual data alongside numerical sequences and the repurposing of Large Language Models (LLMs) for temporal data analysis [5], [6], [7]. Recent benchmarks, such as Time-MMD [8], have demonstrated the potential of this multimodal approach, asserting that fusing text with numerical data yields universal performance improvements across diverse domains.

However, a critical re-examination of this "universal gain" hypothesis reveals unexplored methodological limitations in how multimodal fusion is currently evaluated. Previous studies often introduce evaluation biases by benchmarking multimodal models against either overly simplified numerical baselines or highly parameterized transformer architectures, both of which struggle to properly model event-induced dynamics [8]. On one hand, simple linear baselines like DLinear [9] lack the representational capacity to capture sudden, event-based shifts. Because of this, when a multimodal model outperforms a simple baseline, the real reason remains unclear: is the combination of text and numerical data genuinely effective, or is the text simply making up for a weak numerical model? On the other hand, high-capacity models such as PatchTST [10] and iTransformer [11] are prone to overfitting localized anomalies in a forceful attempt to memorize sudden event shocks [9]. This tendency not only compromises their ability to learn robust normal periodicities but also artificially shrinks the residual error space, thereby obscuring the true orthogonal value that textual data provides during extreme events.

To rigorously isolate and evaluate the true marginal utility of the textual modality, we conduct a fine-grained, domain-specific empirical study across nine diverse datasets. Rather than evaluating black-box, state-of-the-art architectures, we adopt MoLE (Mixture of Linear Experts) [12] as a robust and transparent purely numerical baseline. By strictly utilizing masked linear projections across multiple historical windows, MoLE maximizes the capacity of pure numerical trend extraction while remaining mathematically constrained from overfitting non-linear event shocks. Furthermore, to eliminate the representation noise often introduced by generative models like GPT-2 [13]—a limitation well-documented in representation learning research such as Sentence-BERT [14]—we employ a state-of-the-art bidirectional text encoder, BGE-m3 [15] from BAAI, ensuring the highest quality of text feature alignment. Crucially, this rigorous setup allows us to reveal that current multimodal failures stem not from architectural deficits, but from two distinct bottlenecks: *representation interference in the feature space* and *optimization conflicts during joint training*.

Our extensive evaluation uncovers a stark and counter-intuitive reality: multimodal fusion is not a universally applicable solution, and in several scenarios, it actively degrades forecasting performance. Specifically, our empirical study yields

*Corresponding author.

© 2026, copyrighted by International Institute of Applied Informatics.

DOI 10.23919/IIAI-AAICPS00095.2026.00106
Authorized licensed use limited to: Aizu University. Downloaded on September 11,2026 at 04:23:29 UTC from IEEE Xplore. Restrictions apply.

three core contributions:

- **Demystifying the Domain Boundary of Textual Utility:** We shatter the illusion that multimodal fusion yields universal gains. Our domain-specific evaluation clarifies the role of textual features. In domains governed by strong intrinsic periodicities or high-frequency noise (e.g., Economy, Traffic, Agriculture), introducing text acts as a disruptive signal, actively degrading the performance of pure numerical baselines, a limitation similarly observed in financial NLP where unstructured and noisy text collections can actively dilute predictive signals [16]. Conversely, in macro-shock susceptible domains (e.g., Climate, Energy, Public Health), where external events dictate sudden trajectory shifts, text features serve as an indispensable "shock predictor," corroborating previous domain-specific studies on event-driven forecasting, such as influenza surveillance using social media [17], [18] and crude oil price forecasting incorporating news text [19].
- **Identifying the Bottleneck in Multimodal Optimization:** We reveal that the huge forecasting errors seen in existing models—such as Time-MMD's failure on specific datasets—are mainly caused by models joint training. Joint training forces the model to focus too much on the text, which hinders the numerical model from capturing time-series patterns. This phenomenon aligns with known modality competition issues in general multimodal learning, where joint optimization often leads to suboptimal convergence for individual modalities [20], [21]. Combined with the noisy text features from text encoder like GPT-2, this approach causes the text and numerical data to interfere with each other, ultimately lowering performance.
- **A Universal Rescuing Framework via Orthogonal Decoupling:** To overcome these bottlenecks, we propose a paradigm shift from entangled feature fusion to Orthogonal Decoupling via a novel Two-Stage Optimization framework. By first isolating the numerical backbone to learn pure periodicities, and subsequently restricting high-fidelity bidirectional text representations to exclusively fit the unpredictable residual shocks, we eliminate modality interference. Strikingly, applying our framework to existing architectures not only eradicates their catastrophic errors but also increase the performance across all macro-shock susceptible domains.

II. Methods

Given a historical multivariate time series $\mathbf{X}_{ts} \in \mathbb{R}^{L\times C}$ with a look-back window of length L and C channels (where $C = 1$ in our specific case), alongside the corresponding textual information $\mathbf{X}_{text}$ within the same period, our goal is to predict the future values $\mathbf{Y} \in \mathbb{R}^{H\times C}$ for the prediction horizon H. Formally, this multimodal time series forecasting task can be formulated as mapping the aligned text and numerical data to the future sequence:

$$\hat{\mathbf{Y}} = \mathcal{F}(\mathbf{X}_{ts}, \mathbf{X}_{text}; \Theta) \tag{1}$$

where $\mathcal{F}$ denotes the proposed framework and Θ represents the learnable parameters.

A. Numerical Backbone: Mixture of Linear Experts

To guarantee a robust extraction of pure numerical trends without the risk of overfitting sudden event shocks, we adopt the Mixture of Linear Experts (MoLE) [12] model that combine multiple DLinear [9] models with different look back window as our numerical backbone.

Following the standard DLinear design, the historical sequence $\mathbf{X}_{ts}$ is first decomposed into a trend component and a seasonal component using a moving average block. To capture features across different receptive fields, MoLE utilizes a parallel expert mechanism. Specifically, multiple DLinear models with distinct look-back windows are employed as independent experts. We configure $M = 4$ as the total number of experts, and let $\hat{\mathbf{Y}}^{(m)} \in \mathbb{R}^{H\times C}$ represent the forecasting output from the m-th expert. The representations from these parallel experts are stacked into a multi-scale tensor $\mathbf{T} \in \mathbb{R}^{M\times H\times C}$:

$$\mathbf{T} = \text{Stack}(\hat{\mathbf{Y}}^{(1)}, \hat{\mathbf{Y}}^{(2)}, \ldots, \hat{\mathbf{Y}}^{(M)}) \tag{2}$$

Subsequently, treating the feature channels as the depth dimension, a 2D convolution layer with a kernel of size $(M, 1)$ is applied to aggregate the multi-scale predictions across the expert dimension. This interpretable aggregation process yields a highly robust base numerical forecast $\hat{\mathbf{Y}}_{base} \in \mathbb{R}^{H\times C}$:

$$\hat{\mathbf{Y}}_{base} = \text{Squeeze}(\text{BatchNorm}(\text{GELU}(\text{Conv2D}(\mathbf{T})))) \tag{3}$$

B. Semantic Extraction and Orthogonal Projection

Unlike previous frameworks [8] that rely on autoregressive decoders (e.g., GPT-2 [13]) which inherently introduce unidirectional representation noise, we prioritize high-fidelity semantic extraction to prevent representation entanglement. To this end, we employ a state-of-the-art bidirectional encoder, BGE-m3 [15] (denoted as Enc), with frozen parameters to extract pure semantic embeddings from the text input $\mathbf{X}_{text}$:

$$\mathbf{E}_{text} = \text{Enc}(\mathbf{X}_{text}) \in \mathbb{R}^{d_{text}} \tag{4}$$

In our configuration, the semantic embedding dimension d_{text} is set to 1024.

To bridge the modality gap, we explore two distinct projection paradigms within our frozen-backbone framework.The first is Latent Residual Addition (employed in Time-MMD-BERT for direct baseline comparison), where the text semantic embedding is simply fused via element-wise addition:

$$\hat{\mathbf{Y}}_{text} = \text{MLP}_{proj}(\mathbf{E}_{text}) \tag{5}$$

$$\hat{\mathbf{Y}}_1 = \hat{\mathbf{Y}}_{base} + \hat{\mathbf{Y}}_{text} \tag{6}$$

Authorized licensed use limited to: Aizu University. Downloaded on September 11,2026 at 04:23:29 UTC from IEEE Xplore. Restrictions apply.

The second is Scale-and-Shift Modulation (employed in MoLE-BERT and DLinear-BERT), where text embeddings generate a scaling factor γ and a shifting factor β via 2 MLPs:

$$\hat{\mathbf{Y}}_2 = \hat{\mathbf{Y}}_{base} \odot \gamma + \beta \tag{7}$$

In both paradigms, the MLPs are designed to compress the 1024-dimensional semantic embedding into a 64-dimensional intermediate representation, utilizing a GELU activation function and a dropout rate of 0.5.

Crucially, unlike conventional end-to-end approaches where such direct modifications can corrupt the backbone's representations, our Two-Stage framework ensures that these scaling and shifting operations exclusively calibrate the residual space.

C. Two-Stage Orthogonal Optimization

To thoroughly eliminate modality interference and ensure stable multimodal alignment, we abandon the conventional end-to-end joint training paradigm. Instead, the proposed framework is trained using a novel Two-Stage Orthogonal Optimization strategy, as shown in Fig. 1.

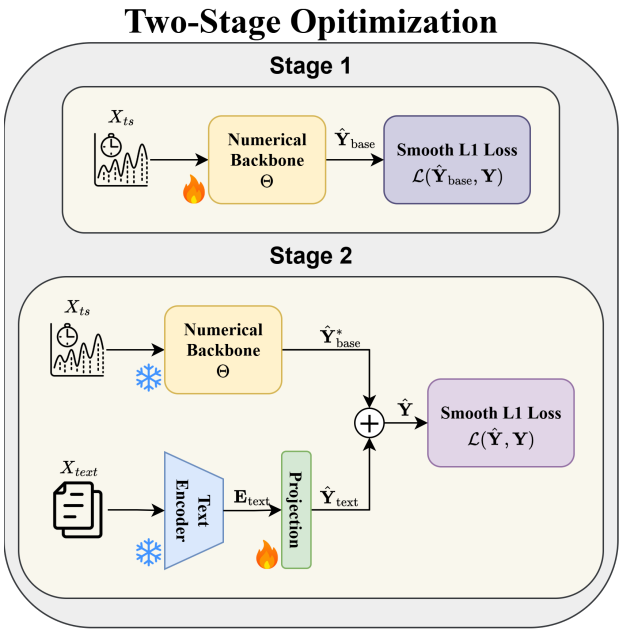


Fig. 1: Two-Stage Orthogonal Optimization.

To optimize the model, we employ the Smooth L1 Loss (Huber Loss) as the primary objective function. Compared to the standard Mean Squared Error (MSE), the Smooth L1 Loss is less sensitive to extreme outliers in time series data, providing more stable gradients. The loss function is averaged over all time steps and dimensions:

$$\mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{H \times C} \sum_{i=1}^{H} \sum_{j=1}^{C} \text{SmoothL1}(\hat{y}_{i,j}, y_{i,j}) \tag{8}$$

Our optimization strictly follows two decoupled phases to protect the integrity of temporal representations:

Stage 1: Base Trend Learning. In the first phase, the textual modality is entirely bypassed. We exclusively optimize the parameters of the numerical backbone (Θ_{num}) to establish a robust foundation in capturing pure temporal dynamics. The objective is:

$$\min_{\Theta_{num}} \mathcal{L}(\hat{\mathbf{Y}}_{base}, \mathbf{Y}) \tag{9}$$

This stage continues until convergence, forcing the backbone to fully exhaust the predictive limits of the numerical periodicities without any textual interference.

Stage 2: Semantic Residual Calibration. Once Stage 1 converges, the optimal deep numerical parameters Θ^*_{num} are strictly **frozen**. Consequently, the base representation $\hat{\mathbf{Y}}^*_{base}$ acts as a constant tensor. In this phase, we activate the textual projector (Θ_{proj}) to strictly optimize the latent residual addition. The objective shifts to:

$$\min_{\Theta_{proj}} \mathcal{L}(\hat{\mathbf{Y}}^*_{base} + \hat{\mathbf{Y}}_{text}, \mathbf{Y}) \tag{10}$$

III. Experiments

A. Datasets

To evaluate our Two-Stage Orthogonal Optimization framework, we tested it across nine different domains from the Time-MMD benchmark. Table I shows the overall forecasting performance, measured by Mean Squared Error (MSE). For comparison, we also include the results of the Transformer-based approach, PatchTST, and the LLM-based approach, Time-LLM, as reported in the Time-MMD paper [8].

Following the exact evaluation protocol established by the original Time-MMD benchmark, we report the Mean Squared Error (MSE). Furthermore, given that the primary objective of our textual modality is to capture sudden, macro-shock residuals rather than small routine fluctuations, MSE is intentionally selected as the sole metric, as its quadratic nature strictly penalizes large, catastrophic outlier errors associated with failed event predictions

B. Implementation Details

Throughout the experiments, the framework is trained for a maximum of 100 epochs with a batch size of 16, random seed 3407. Parameters are updated using the AdamW optimizer with a weight decay of 0.005 to prevent overfitting. To ensure stable convergence across phases, Stage 1 is trained with a base learning rate of 5×10^{-4}, while Stage 2 employs a halved learning rate of 2.5×10^{-4}. Furthermore, an early stopping mechanism with a patience of 10 epochs is utilized in both stages to halt the training process if the validation loss plateaus, thereby ensuring optimal generalization to unseen events.

As discussed in Section I, we deliberately exclude highly parameterized Transformer models from our baseline comparison. Previous studies have proven that these complex models tend to overfit sudden shocks and local anomalies [9]. This overfitting makes it difficult to measure the true benefit of adding text data. Therefore, we strictly rely on linear-based models (DLinear and MoLE) to provide a clean, transparent,

Authorized licensed use limited to: Aizu University. Downloaded on September 11,2026 at 04:23:29 UTC from IEEE Xplore. Restrictions apply.

TABLE I: MSE Performance: Semantic-Rich vs. Signal-Dominated Domains

Domain	Numerical Backbone		Baselines			With Text (Proposed Two-Stage)		
	DLinear	MoLE	Time-MMD	PatchTST	Time-LLM	Time-MMD-BERT	DLinear-BERT	MoLE-BERT
Semantic-Rich Domains (Text-Beneficial):								
Climate	0.4540	0.4727	1.0650	1.0250	1.0275	0.5462	0.4616	**0.4484**
Energy	0.1164	0.1115	0.3350	0.2525	0.2500	0.1155	0.1073	**0.1023**
Environment	0.2197	0.2236	0.4000	0.3125	0.2950	**0.2024**	0.2089	0.2051
Public-Health	0.5128	0.4949	1.6875	1.3350	1.4200	**0.4669**	0.4728	0.4776
Signal-Dominated (Text-Disruptive):								
Agriculture	0.3608	0.1181	0.1375	**0.0925**	**0.0925**	0.4083	0.6896	1.0003
Economy	0.0242	0.2640	0.0500	**0.0200**	**0.0200**	0.2943	0.8563	0.5820
SocialGood	0.2556	**0.2520**	1.0550	0.9650	1.0175	0.3090	0.3333	0.3450
Traffic	**0.0978**	0.0980	0.2850	0.2100	0.2150	0.0980	0.1134	0.1036
Security	5.3142	**5.2854**	107.3675	112.3225	114.1475	5.8506	5.5296	5.3559

*Best in **Red Bold**, Second Best in Blue Underline

and fair comparison for evaluating the exact impact of the textual modality.

C. Summary of Evaluated Models

To prevent confusion and clearly highlight the differences in our experiments, we summarize the specific configurations of the six models evaluated in this study:

- **DLinear:** A standard pure numerical baseline. It decomposes the time series into trend and seasonal parts and uses linear layers for forecasting. It does not use any text data.
- **MoLE:** A pure numerical backbone, Mixture of Linear Experts. It combine multiple DLinear models with different look-back windows using a 2D convolution layer. It does not use any text data.
- **Time-MMD (Baseline):** The original multimodal baseline from the Time-MMD benchmark. It processes both numerical and textual data simultaneously using standard end-to-end joint training, DLinear as numerical backbone and GPT-2 as the text encoder.
- **Time-MMD-BERT:** A modified version of the baseline to prove our training method. It keeps the exact same numerical architecture as the original Time-MMD, but using BGE-m3 as a text encoder and is trained using our Two-Stage Orthogonal Optimization.
- **DLinear-BERT:** Our proposed two-stage framework applied to a simple baseline. DLinear is trained and frozen in Stage 1, and the text embeddings from BGE-m3 are used to generate a scaling factor γ and a shifting factor β to perform multimodal fusion in Stage 2.
- **MoLE-BERT:** MoLE is trained and frozen in Stage 1, and the text embeddings from BGE-m3 are used to generate a scaling factor γ and a shifting factor β to perform multimodal fusion in Stage 2.

IV. Results

A. When Does Text Actually Help?

Our results show that combining text and numerical data does not always improve forecasting. Instead, the usefulness of text depends entirely on the nature of the domain.

In domains with strong, regular patterns (like Agriculture, Economy, SocialGood, and Traffic), adding text actually hurts performance. For example, in the Economy domain, the pure numerical DLinear model achieves a very low error of 0.0242. However, when we force the model to use text features (DLinear-BERT), the error jumps to 0.8563. We see a similar drop in Agriculture: the pure MoLE model scores 0.1181, but adding text (MoLE-BERT) worsens the error to 1.0003. In these cases, text acts as noisy interference that messes up the clear trends found by the numerical models.

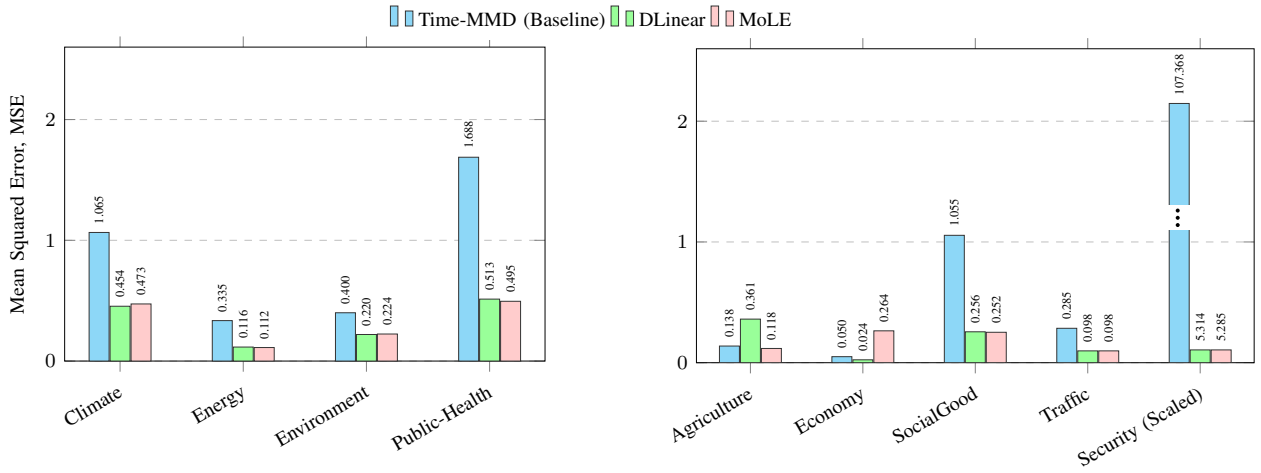
On the other hand, text is highly valuable in event-driven domains where sudden external changes happen frequently, such as Climate, Energy, Environment, and Public Health. In the Energy dataset, our MoLE-BERT model reduces the error from 0.1115 (pure MoLE) to 0.1023. In the Climate domain, MoLE-BERT (0.4484) consistently beats its pure numerical version (0.4727). This proves that text features are excellent "shock predictors," but only when the domain is sensitive to sudden news or events.

B. Fixing Catastrophic Errors

As illustrated in Fig. 2, standard joint training (Time-MMD) leads to massive performance degradation and catastrophic errors in specific domains. Most notably, it yields an extreme MSE of 107.368 in the Security dataset, compared to the stable performance of pure numerical backbones (around 5.3). It also fails in Public Health with an error of 1.688.

As shown in Fig. 3, the proposed two-stage training (Time-MMD-BERT) framework dramatically stabilizes model performance. Except for the Agriculture and Economy domains, our approach yields consistent error reductions across the board. The impact is most striking where end-to-end fusion fails catastrophically: the extreme MSE in the Security domain drops sharply from 107.368 to 5.851, and the error in Public Health falls from 1.688 to 0.467. Ultimately, this demonstrates that our framework retains the robustness of pure numerical models without discarding valuable textual features.

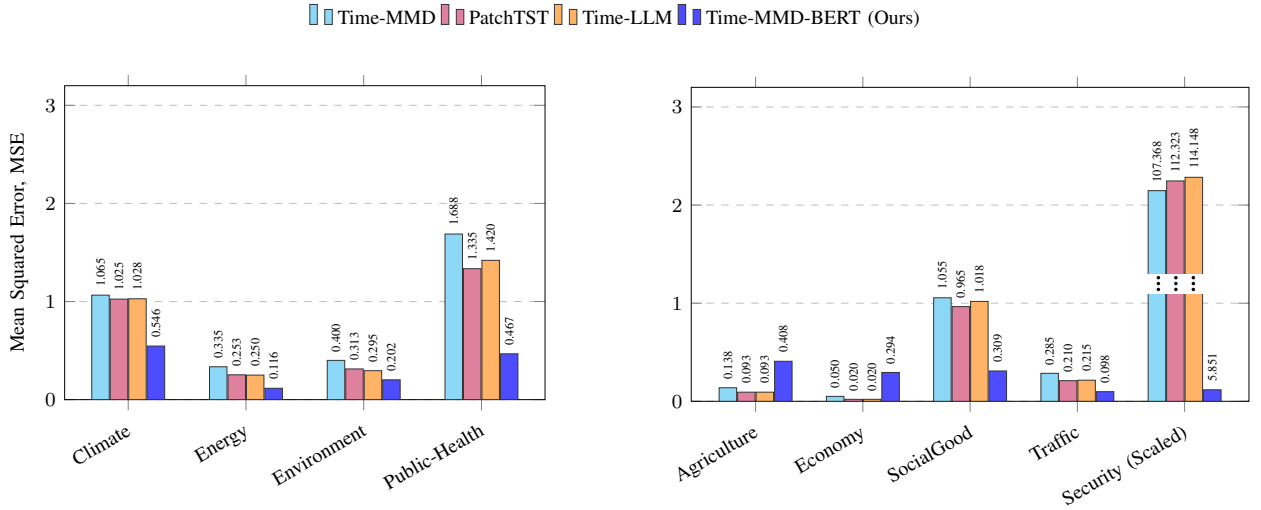
Authorized licensed use limited to: Aizu University. Downloaded on September 11,2026 at 04:23:29 UTC from IEEE Xplore. Restrictions apply.



(a) Event-Driven Domains (Text-Beneficial)

(b) Periodic Domains (Text-Disruptive)

Fig. 2: Performance comparison of base numerical models (DLinear, MoLE) against the end-to-end multimodal baseline (Time-MMD) grouped by domain characteristics. Note: Security domain is scaled down by a factor of 50.



(a) Event-Driven Domains (Text-Beneficial)

(b) Periodic Domains (Text-Disruptive)

Fig. 3: Performance comparison between the baselines (Time-MMD, PatchTST, Time-LLM) and our proposed Two-Stage Orthogonal Optimization method (Time-MMD-BERT) across grouped domains.

V. Discussion

A. The Danger of End-to-End Multimodal Training

The catastrophic errors observed in the original Time-MMD baseline demonstrate the critical necessity of training a robust purely numerical backbone first. We argue that standard end-to-end training suffers from "optimization entanglement." Processing numerical and textual data simultaneously forces the model to disproportionately focus on noisy textual features, severely disrupting its capacity to capture underlying temporal dynamics.

B. Why Two-Stage Decoupling Works

The significant stabilization achieved by Time-MMD-BERT validates our hypothesis that the training paradigm itself—specifically modality interference—is the primary bottleneck in conventional methods. By strictly decoupling numerical trend extraction from semantic text adjustments, our method effectively mitigates cross-modal interference.

In our two-stage framework, the numerical backbone is trained first and then frozen, making it fully responsible for learning the inherent periodic patterns. Subsequently, in

Authorized licensed use limited to: Aizu University. Downloaded on September 11,2026 at 04:23:29 UTC from IEEE Xplore. Restrictions apply.

the second stage, the text projection layer only learns to predict the ”residuals” (e.g., unexpected shocks or sudden anomalies). This orthogonal decoupling ensures that textual features provide extra value without corrupting the temporal trends learned by the numerical backbone. Ultimately, this empirical evidence establishes that orthogonal decoupling is an essential prerequisite for robust multimodal forecasting, prioritizing inherent stability over simply chasing an absolute state-of-the-art across all scenarios.

C. Rethinking Multimodal Forecasting

Our empirical study suggests a necessary shift in how the field approaches multimodal time series forecasting. The assumption that ”multimodal is always better” is incorrect. Text should not be treated as a universal solution for all forecasting tasks. Instead, it should be viewed as a specialized tool for predicting sudden anomalies in event-driven domains.

VI. Conclusion and Future Work

In this paper, we presented a study on domain-dependent multimodal time series forecasting. Our results clearly show that adding text does not always improve forecasting. Instead, we found a clear boundary: text features are very helpful for predicting sudden changes in event-driven domains (e.g., Public Health, Security), but they act as harmful noise in domains with regular, repeating patterns (e.g., Agriculture, Economy). Furthermore, we demonstrated that standard joint-training methods fail because the text and numerical data interfere with each other. By introducing a Two-Stage Optimization framework, we successfully separated the mathematical trend learning from the text-based adjustments. This approach eliminated the catastrophic failures seen in previous baseline models and achieved highly reliable performance.

Despite these promising results, our study has some limitations. The current framework uses fixed, pre-trained text features (BGE-m3) and does not explore the benefits of fine-tuning large language models during the second training stage. Also, the decision of whether to use text features still relies on human knowledge about the specific domain.

For future work, we plan to explore adaptive methods that can automatically evaluate the usefulness of text in real-time, allowing the model to dynamically adjust how much it relies on text data. Furthermore, we plan to explore using cross-attention for multimodal fusion, strictly built on top of a robust numerical backbone model.

References

[1] 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,” 2021. [Online]. Available: https://arxiv.org/abs/2012.07436

[2] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, “Financial time series forecasting with deep learning : A systematic literature review: 2005-2019,” 2019. [Online]. Available: https://arxiv.org/abs/1911.13288

[3] B. Yu, H. Yin, and Z. Zhu, “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, ser. IJCAI-2018. International Joint Conferences on Artificial Intelligence Organization, Jul. 2018, p. 3634–3640. [Online]. Available: http://dx.doi.org/10.24963/ijcai.2018/505

[4] Y. Li, R. Yu, C. Shahabi, and Y. Liu, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” 2018. [Online]. Available: https://arxiv.org/abs/1707.01926

[5] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen, “Time-llm: Time series forecasting by reprogramming large language models,” 2024. [Online]. Available: https://arxiv.org/abs/2310.01728

[6] H. Xue and F. D. Salim, “Promptcast: A new prompt-based learning paradigm for time series forecasting,” 2023. [Online]. Available: https://arxiv.org/abs/2210.08964

[7] M. Jin, Q. Wen, Y. Liang, C. Zhang, S. Xue, X. Wang, J. Zhang, Y. Wang, H. Chen, X. Li, S. Pan, V. S. Tseng, Y. Zheng, L. Chen, and H. Xiong, “Large models for time series and spatio-temporal data: A survey and outlook,” 2023. [Online]. Available: https://arxiv.org/abs/2310.10196

[8] H. Liu, S. Xu, Z. Zhao, L. Kong, H. Kamarthi, A. B. Sasanur, M. Sharma, J. Cui, Q. Wen, C. Zhang, and B. A. Prakash, “Time-mmd: Multi-domain multimodal dataset for time series analysis,” 2025. [Online]. Available: https://arxiv.org/abs/2406.08627

[9] A. Zeng, M. Chen, L. Zhang, and Q. Xu, “Are transformers effective for time series forecasting?” 2022. [Online]. Available: https://arxiv.org/abs/2205.13504

[10] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, “A time series is worth 64 words: Long-term forecasting with transformers,” 2023. [Online]. Available: https://arxiv.org/abs/2211.14730

[11] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, “itransformer: Inverted transformers are effective for time series forecasting,” 2024. [Online]. Available: https://arxiv.org/abs/2310.06625

[12] L. Zhang, L. Shen, Y. Zheng, S. Piao, Z. Li, and F. Tsung, “Lemole: Llm-enhanced mixture of linear experts for time series forecasting,” 2024. [Online]. Available: https://arxiv.org/abs/2412.00053

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI*, 2019, accessed: 2024-11-15. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[14] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[15] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation,” 2025. [Online]. Available: https://arxiv.org/abs/2402.03216

[16] F. Z. Xing, E. Cambria, and R. E. Welsch, “Natural language based financial forecasting: a survey,” *Artif. Intell. Rev.*, vol. 50, no. 1, p. 49–73, Jun. 2018. [Online]. Available: https://doi.org/10.1007/s10462-017-9588-9

[17] E. Aramaki, S. Maskawa, and M. Morita, “Twitter catches the flu: Detecting influenza epidemics using Twitter,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, R. Barzilay and M. Johnson, Eds. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 1568–1576. [Online]. Available: https://aclanthology.org/D11-1145/

[18] M. J. Paul, M. Dredze, and D. Broniatowski, “Twitter improves influenza forecasting,” *PLoS currents*, vol. 6, 2014.

[19] Y. Bai, X. Li, H. Yu, and S. Jia, “Crude oil price forecasting incorporating news text,” *International Journal of Forecasting*, vol. 38, no. 1, pp. 367–383, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207021001060

[20] W. Wang, D. Tran, and M. Feiszli, “What makes training multi-modal classification networks hard?” 2020. [Online]. Available: https://arxiv.org/abs/1905.12681

[21] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” 2022. [Online]. Available: https://arxiv.org/abs/2203.02053

Authorized licensed use limited to: Aizu University. Downloaded on September 11,2026 at 04:23:29 UTC from IEEE Xplore. Restrictions apply.