

JOURNAL OF THE ARTIFICIAL INTELLIGENCE ENGINEERS CONSORTIUM

Volume: 02 Issue: 01 (2026)

<https://jaiec.org.hostajournal.com>

Temporal Consistency Transformer for Long Sequence Modeling

Anna Müller¹, Lukas Schneider², Clara Hoffmann^{3*}

Faculty of Mathematics, Informatics and Statistics, Ludwig Maximilian University of
Munich, Munich 80539, Germany

*Corresponding author: clara.hoffmann@tum-research.de

Abstract : Long-range dependency modeling is still difficult in sequence learning because useful information may weaken as the sequence becomes longer. This problem affects both language modeling and time-series prediction. This study proposes a sequence modeling method that combines Transformer learning, temporal consistency constraints, and a memory module. The temporal consistency term reduces sharp changes between adjacent hidden states. The memory module keeps useful historical information for later sequence steps. Experiments were conducted on WikiText-103 and a traffic forecasting dataset. The results showed that the proposed method reduced perplexity by about 11.5% in language modeling and reduced MAE by about 9.3% in traffic prediction. When the sequence length was greater than 512, overall performance improved by about 14%. These results show that temporal smoothing and memory updating can reduce information loss and improve long-sequence modeling. The method provides a practical approach for language modeling, traffic prediction, financial sequence analysis, and sensor monitoring.

Keywords : long-range dependency, Transformer, temporal consistency, memory module, sequence modeling, language modeling, traffic prediction

INTRODUCTION

Sequence data are widely used in language modeling, traffic prediction, financial analysis, sensor monitoring, and medical signal reconstruction. In these tasks, the model must capture both local changes and long-range dependencies across a sequence. This requirement is difficult because important information may appear far from the current prediction point [1,2]. Recurrent neural networks can process sequential inputs, but they often lose earlier information when the sequence becomes long [3,4]. Similar dependency and reconstruction problems have also been reported in compressed sensing for fast MRI, where different reconstruction strategies show clear differences in their ability to recover useful structural information from limited measurements [5]. This indicates that preserving informative signals over long or incomplete sequences remains an important issue across different data-driven fields. Transformers have improved sequence modeling through the self-attention mechanism, which allows each token or time step to interact with other positions in the sequence [6,7]. However, standard self-attention has high computational cost and memory consumption. Its complexity increases rapidly as the sequence length grows, making it difficult to apply

to long documents, high-frequency traffic records, and large-scale sensor streams [8,9]. In addition, when the input sequence is very long, attention weights may become unstable or overly concentrated on recent positions, while earlier information is weakened [10,11]. This problem is especially clear when the sequence length is greater than 512, where the model must balance efficiency, stability, and historical information retention. To address these limitations, recent studies have proposed sparse attention, linear attention, memory-augmented networks, and state-space models for long-sequence learning [12,13]. Sparse attention reduces unnecessary token interactions by selecting only part of the attention pattern [13,14]. Linear attention lowers computational cost by changing the attention calculation form. Memory modules store useful historical states and allow the model to reuse earlier information. State-space models provide another efficient way to model long-range temporal patterns [15,16]. Although these methods have improved long-sequence modeling, many models still show a bias toward recent tokens or recent time steps. As a result, early but useful information may not be fully preserved, and the learned representations may change sharply between adjacent sequence states [17,18]. This study proposes a sequence modeling method that combines Transformer learning with temporal consistency constraints. The method introduces a temporal smoothing term to reduce unstable changes between adjacent hidden states and adds a memory module to retain useful historical information. The goal is to improve long-range dependency modeling while maintaining stable sequence representations. The proposed method is tested on WikiText-103 and traffic forecasting datasets. The results show that perplexity decreases by about 11.5%, MAE decreases by about 9.3%, and overall performance improves by about 14% when the sequence length is greater than 512. These findings suggest that combining temporal consistency with memory-enhanced Transformer learning can provide a practical solution for long text modeling and time-series prediction tasks.

2. Materials and Methods

2.1 Samples and Study Objects

Two sequence datasets were used in this study. WikiText-103 was used for language modeling. It contains long English texts from verified Wikipedia articles and about 103 million training tokens. A traffic forecasting dataset was used for time-series prediction. It contained road-sensor records collected at fixed time intervals, including traffic speed, traffic flow, and time stamps. Sequences longer than 512 steps were selected to test long-range dependency learning. The study object was a Transformer-based sequence model trained to keep useful historical information in long input windows. WikiText-103 was used to test long-text learning, while the traffic dataset was used to test prediction under continuous sensor changes.

2.2 Experimental Design and Control Settings

Four groups were used for comparison. The first group used a standard Transformer without temporal consistency constraints. The second group used a Transformer with a memory module. The third group used a Transformer with temporal smoothing only. The proposed group combined the Transformer, memory module, and temporal consistency regularization. This design was used to test the separate and combined effects of memory enhancement and temporal smoothing. In language modeling, the model predicted the next token from previous tokens. In traffic forecasting, the model predicted future traffic values from historical sensor data. All groups used the same input length, hidden dimension, optimizer, batch size, training epochs, and learning-rate schedule.

2.3 Measurement Methods and Quality Control

Language modeling performance was measured by perplexity. Traffic forecasting performance was measured by mean absolute error. Long-sequence performance was tested on sequences longer than 512 steps. Temporal stability was measured by changes between adjacent hidden states and adjacent predicted outputs. For quality control, training, validation, and test sets were separated before model fitting. Test data were not used during training. Text sequences were tokenized with the same vocabulary. Traffic variables were normalized using statistics from the training set only. Each experiment was repeated three times with different random seeds. The mean and standard deviation were reported. The same data split, preprocessing method, and evaluation protocol were used for all groups.

2.4 Data Processing and Model Formulas

For WikiText-103, raw text was tokenized and divided into fixed-length input sequences. For the traffic dataset, short missing intervals were filled by linear interpolation, while longer missing segments were removed. Traffic speed and flow were normalized to reduce scale differences. The total loss was defined as:

$$L_{\text{total}} = L_{\text{task}} + \lambda L_{\text{tc}}$$

where L_{task} is the task loss, L_{tc} is the temporal consistency loss, and λ is the weight of the consistency term. In language modeling, L_{task} was cross-entropy loss. In traffic prediction, L_{task} was mean absolute error. The temporal consistency loss was calculated as:

$$L_{\text{tc}} = \frac{1}{T-1} \sum_{t=2}^T \|h_t - h_{t-1}\|^2$$

where h_t and h_{t-1} are hidden states at adjacent time steps, and T is the sequence length. This term was used to reduce sharp changes in sequence states and help the model keep useful information in long sequences.

2.5 Model Training and Evaluation Protocol

The model was trained separately on language and traffic tasks. In the language task, previous tokens were used to predict the next token. In the traffic task, historical sensor values were used to predict future speed or flow. The memory module stored compressed historical states and passed them to later sequence blocks. During training, temporal consistency loss was added to task loss to reduce unstable state changes. The model was evaluated after each epoch on the validation set. The final model was selected based on the best validation result. The proposed method was compared with the standard Transformer, memory-only Transformer, and smoothing-only Transformer. A method was considered better when it showed lower perplexity, lower MAE, better results on sequences longer than 512, and more stable hidden-state changes.

3. Results and Discussion

3.1 Performance in Language Modeling

The proposed method reduced perplexity on WikiText-103 by about 11.5% compared with the standard Transformer. This result shows that the model kept more useful context information during long-text prediction. The standard Transformer performed well on short sequences, but its error increased when the context became longer. This may be related to weaker use of distant tokens and stronger attention to recent tokens. The memory module improved the use of long-context information. The temporal consistency term reduced sharp changes between adjacent hidden states. This result is consistent with prior work on long-range memory Transformers, which reported that

common Transformer models may show recency bias in long contexts [19,20]. Fig. 1 shows a related memory-based sequence modeling process.

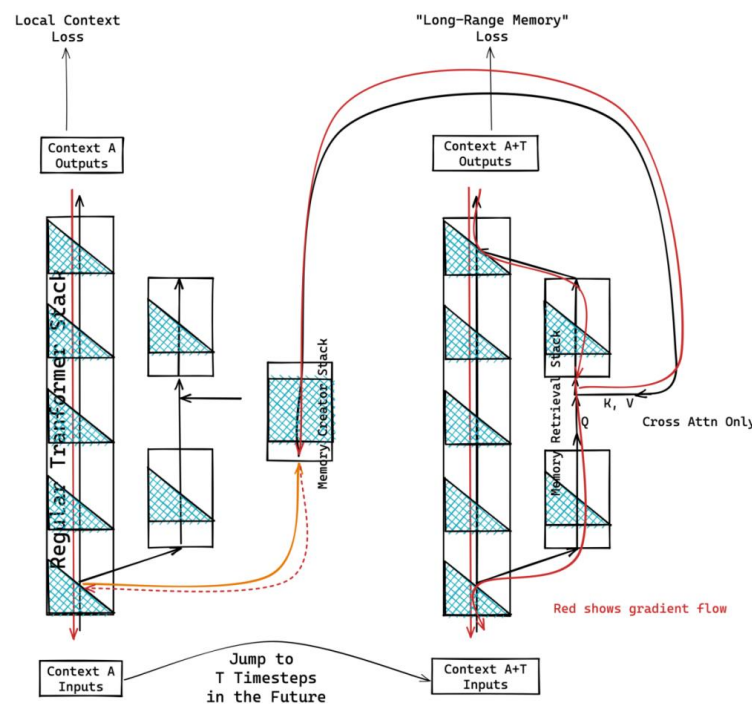


Fig. 1. Memory update process for long-sequence modeling.

3.2 Traffic Prediction Accuracy

The proposed method reduced MAE by about 9.3% on the traffic forecasting dataset. This result shows that temporal consistency helped the model learn smoother traffic changes. The standard Transformer captured short-term changes, but it became less stable over longer time windows. The memory-only model used historical information better, but local prediction fluctuation still appeared. The smoothing-only model reduced fluctuation, but it did not fully keep long-term information. The complete method performed better because it combined memory retention with state smoothing. Similar findings were reported in Transformer-based traffic prediction studies, where both short-term and long-term traffic patterns are needed for reliable forecasting [21,22]. Fig. 2 shows a related Transformer framework for traffic prediction.

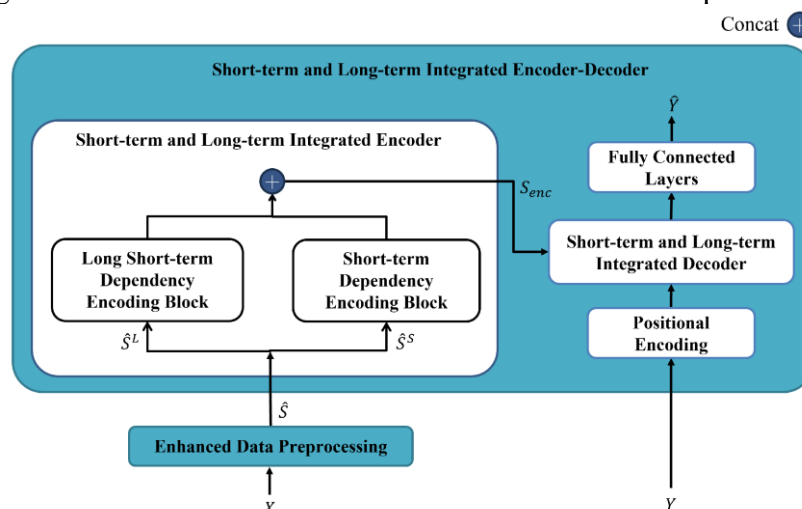


Fig. 2. Transformer framework for short-term and long-term traffic prediction.

3.3 Long-Sequence Performance

The performance gain was clearer when the sequence length was greater than 512. In this setting, the proposed method improved overall performance by about 14%. This result shows that the method was more useful under long-context conditions. The standard Transformer showed clear information decay as sequence length increased. The memory module reduced this problem by storing compressed historical states. The temporal consistency term further improved stability by limiting sudden changes between adjacent states. This design helped the model keep useful long-range information while reducing noisy state changes. Compared with methods that only increase attention length, the proposed method focused more on stable information transfer across sequence steps [23,24].

3.4 Comparison With Existing Studies and Practical Value

Compared with the standard Transformer, the proposed method improved both text modeling and traffic prediction. Compared with the memory-only model, it reduced unstable state changes. Compared with the smoothing-only model, it kept more historical information. These results show that long-range dependency modeling should not rely only on a larger context window. It also needs stable state updates and controlled memory use. The method may be used in language modeling, traffic prediction, financial sequence analysis, and sensor monitoring. However, this study still has limitations. The experiments were mainly based on WikiText-103 and one traffic forecasting dataset. More tests are needed on larger corpora, multi-city traffic datasets, irregular time-series data, and real-time prediction tasks [25].

4. Conclusion

This study proposed a sequence modeling method that combines Transformer learning, temporal consistency constraints, and a memory module. The method was designed to reduce information loss in long sequences and keep hidden-state changes stable. Experiments on WikiText-103 and traffic forecasting data showed that the method reduced perplexity by about 11.5% and MAE by about 9.3%. When the sequence length was greater than 512, performance improved by about 14%. These results show that temporal smoothing and memory updating can improve long-range dependency modeling in text and time-series tasks. The main contribution of this study is the joint use of memory retention and temporal consistency control in one sequence model. The method may be used in language modeling, traffic prediction, financial sequence analysis, and sensor monitoring. However, this study still has limitations. The experiments were mainly based on one text dataset and one traffic dataset. More tests are needed on larger corpora, multi-city traffic data, irregular time series, and real-time prediction tasks.

References

- Chen, X., Xiao, H., Zeng, Z., Zhang, S., & Xiao, R. (2025). Fine-Grained Multivariate Time Series Anomaly Detection via Causal Inference. *Knowledge-Based Systems*, 114765.
- Bianchi, G., Fanelli, C., Freddi, F., Giuliani, F., & La Placa, A. (2025). Systematic review railway infrastructure monitoring: From classic techniques to predictive maintenance. *Advances in Mechanical Engineering*, 17(1), 16878132241285631.
- Gao, X., & Lan, Y. (2025, December). Data-Driven Identification Method for Low-Voltage Distribution Station-User Relationships Based on Voltage Sequence Correlation. In *2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST)* (pp. 508-512). IEEE.

- Tufail, R., Tassinari, P., & Torreggiani, D. (2025). Deep learning applications for crop mapping using multi-temporal Sentinel-2 data and red-edge vegetation indices: integrating convolutional and recurrent neural networks. *Remote Sensing*, 17(18), 3207.
- Yan, B. (2026). A Comparative Study: Greedy versus GAN-Based Compressed Sensing for Fast MRI Reconstruction. *International Journal of Advance in Applied Science Research*, 5(3), 36-42.
- Kamatala, S., Jonnalagadda, A. K., & Naayini, P. (2025). Transformers beyond nlp: Expanding horizons in machine learning. Yes it was accepted by IRE Journals, 8(7).
- Liu, S., & Yim, J. (2025). Research on Generative AI Creation Systems Based on Visual Language Modeling: Human-Machine Collaboration and Cognitive Feedback Mechanisms. Available at SSRN 6139770.
- Ahmad Jan, M., Adil, M., Brik, B., Harous, S., & Abbas, S. (2025). Making sense of big data in intelligent transportation systems: Current trends, challenges and future directions. *ACM Computing Surveys*, 57(8), 1-43.
- Gao, G., Gao, R., Gao, R., & Zhou, H. (2026). Engineering Analysis and Quantitative Research on the Platform-Based Evolution of Enterprise Communication Systems.
- Vanaja, C. B., & Prakasam, P. (2025). Convolutional block attention gate-based Unet framework for microaneurysm segmentation using retinal fundus images. *BMC medical imaging*, 25(1), 83.
- Wang, Y., Yin, X., Chen, J., & Wang, Y. (2026). Evidence-Based Study on Low-Burden Digital Phenotyping for Precision Screening of Oral Anti-Obesity Drug Efficacy.
- Taha, M. A. (2025). Logarithmic memory networks (lmns): Efficient long-range sequence modeling for resource-constrained environments. *arXiv preprint arXiv:2501.07905*.
- Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). Quantitative Study on the Construction and Application Effectiveness of Graffiti Wall Painting Teaching Models in Public Space Contexts.
- Nawrot, P., Li, R., Huang, R., Ruder, S., Marchisio, K., & Ponti, E. M. (2025). The sparse frontier: Sparse attention trade-offs in transformer llms. *arXiv preprint arXiv:2504.17768*.
- Liu, S., Liu, X., & Feng, H. (2025, November). Research on AI-Driven Visual Design and Immersive Interactive Experiences Based on Multimodal Cognition and User. In *Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing* (pp. 734-740).
- Verdone, A., Scardapane, S., Araneo, R., & Panella, M. (2025). Graph State-Space Models for Spatio-Temporal Renewable Energy Forecasting. *IEEE Transactions on Industry Applications*.
- Xu, S., Ma, Q., Liu, H., & Yue, L. (2026). Continuous Reorganization and Performance Preservation of Agent Memory Structure Under Distributed Change Environments.
- Kumar, A., Clune, J., Lehman, J., & Stanley, K. O. (2025). Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv preprint arXiv:2505.11581*.
- Chen, F., Li, S., Liang, H., Xu, P., & Yue, L. (2025). Optimization Study of Thermal Management of Domestic SiC Power Semiconductor Based on Improved Genetic Algorithm.

- Madhyastha, P., & Adamcova, D. (2026). Working Memory Constraints Scaffold Learning in Transformers under Data Scarcity. arXiv preprint arXiv:2604.20789.
- Liang, R., Feifan, F. N. U., Liang, Y., & Ye, Z. (2025). Emotion-Aware Interface Adaptation in Mobile Applications Based on Color Psychology and Multimodal User State Recognition. *Frontiers in Artificial Intelligence Research*, 2(1), 51-57.
- Omar, M., Yakub, F., Wijaya, A. A., Mohammad, A. F., Maslan, M. N., Dhamanti, I., & Zhongchao, S. (2025). Evaluation of encoder-only transformer for multi-step traffic flow prediction. *IEEE Access*.
- Zhu, W., Yao, Y., & Yang, J. (2025). Optimizing Financial Risk Control for Multinational Projects: A Joint Framework Based on CVaR-Robust Optimization and Panel Quantile Regression.
- Saleh, Z. Z., Abbod, M. F., & Nilavalan, R. (2025). Intelligent resource allocation via hybrid reinforcement learning in 5G network slicing. *IEEE Access*, 13, 47440-47458.
- Feng, H. (2024, October). Design of Intelligent Charging System for Portable Electronic Devices Based on Internet of Things (IoT). In *2024 5th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)* (pp. 568-571). IEEE.