Streaming Propagation Through Time: A New Computational Paradigm for Recurrent Neural Networks

A Additional Implementation Detail

All our experimental standards follow Tsinghua University's open-source Time Series Library. Our training equipment is based on RTX 4090. Unless otherwise specified, all our experiments use a default learning rate of $\eta = 0.0001$. The experimental data reported in our tables are average values obtained from at least 5 independent experiments, ensuring statistical significance. Our experimental setup includes training set, validation set, and test set, with early stopping strategy employed. For the loss curves of GRU networks, to provide more intuitive visualization, we only show the first 6000 iterations of training loss for JapaneseVowels, although in the actual training process, the SPTT-Endpoint training curve continues to decrease and reaches stability at approximately 25000 iterations. All experiments are implemented based on PyTorch.

B Dataset Description

For long-term forecasting tasks, to avoid randomness caused by the same sequence length and prediction length, we randomly set different sequence lengths and prediction lengths for each dataset. Tables 1 to 4 respectively show the detailed information of the datasets we used for the four tasks: long-term forecasting, classification, imputation, and anomaly detection.

Table 1: Dataset Overview for Long-Term Prediction. The values in the Sequence column represent (batch size, sequence length, prediction length), respectively.

Dataset	Channel	Information	Frequency.	Dataset Size	Sequence
Electricity	321	Consumption	1 Hour	(18317, 2633, 5261)	(128, 672, 384)
Weather	21	CO2-Concentration	10 Min	(36792, 5271, 10540)	(128, 672, 384)
ETTh	7	Oil Temperature	1 Hour	(8545, 2881, 2881)	(64, 384, 96)
ETTm	7	Oil Temperature	15 Min	(34465, 11521, 11521)	(256, 480, 192)
Traffic	862	Road Occupancy	1 Hour	(12185, 1757, 3509)	(64, 480, 192)
Illness	7	Influenza-like	1 Week	(966, 193, 193)	(32, 96, 96)
Exchange Rate	8	Economy	1 Day	(7588, 1518, 1518)	(64, 192, 96)

Table 2: Dataset descriptions for classification tasks. The values in the Series column represent (batch size, series length), respectively.

Dataset	Dim	Series	Dataset Size	Information (Frequency)	Classes
EthanolConcentration	3	(16, 1751)	(261, 0, 263)	Alcohol Industry	4
FaceDetection	144	(32, 62)	(5890, 0, 3524)	Face (250Hz)	2
Handwriting	3	(16, 152)	(150, 0, 850)	Handwriting	26
Heartbeat	61	(16, 405)	(204, 0, 205)	Heart Beat	2
JapaneseVowels	12	(16, 29)	(270, 0, 370)	Voice	9
PEMS-SF	963	(16, 144)	(267, 0, 173)	Transportation (Daily)	7
SelfRegulationSCP1	6	(16, 896)	(268, 0, 293)	Health (256Hz)	2
SelfRegulationSCP2	7	(16, 1152)	(200, 0, 180)	Health (256Hz)	2
SpokenArabicDigits	13	(64, 93)	(6599, 0, 2199)	Voice (11025Hz)	10

Table 3: Dataset Overview for Imputation tasks. The masking rate is set to 0.25.

Dataset	Channel	Information	Frequency.	Dataset Size	Sequence
ETTm1	7	Electricity	15 Min	(34465, 11521, 11521)	(128, 288, 288)
ETTm2	7	Electricity	15 Min	(34465, 11521, 11521)	(128, 384, 384)
ETTh1	7	Electricity	15 Min	(8545, 2881, 2881)	(64, 96, 96)
ETTh2	7	Electricity	15 Min	(8545, 2881, 2881)	(64, 192, 192)
Electricity	321	Electricity	15 Min	(18317, 2633, 5261)	(128, 480, 480)
Weather	21	CO2-Concentration	10 Min	(36792, 5271, 10540)	(128, 576, 576)

Table 4: Dataset Overview for Anomaly Detection tasks. The anomaly rate is set to 1%.

Dataset	Dim	Series Length	Dataset Size	Information (Frequency)	Batch Size
SMD	38	100	(566724, 141681, 708420)	Server Machine	128
MSL	55	100	(44653, 11664, 73729)	Spacecraft	128
SMAP	25	100	(108146, 27037, 427617)	Spacecraft	128
SWaT	51	100	(396000, 99000, 449919)	Infrastructure	256
PSM	25	100	(105984, 26497, 87841)	Server Machine	128

C Performance Analysis of SPTT: Long-term Forecasting Tasks

To establish the generalizability and robustness of SPTT beyond the LSTM results presented in the main manuscript, we conducted extensive performance evaluations across the same four core task categories: long-term forecasting, classification, imputation, and anomaly detection. The experimental design maintains identical configurations to the main manuscript to ensure fair comparison and meaningful cross-architecture analysis.

Table 5 presents the experimental results of SPTT on GRU networks for long-term forecasting tasks, which effectively validates the core findings based on LSTM in the main text and demonstrates the architecture-agnostic generality of SPTT's advantages. Highly consistent with the LSTM results reported in the main text, SPTT and its variants on GRU architectures similarly outperform traditional BPTT on the vast majority of datasets. This cross-architectural performance pattern consistency strongly supports the effectiveness of SPTT as a new computational paradigm.

Table 5: Long Term Forecast Results Comparison on GRU

Dataset →	ETTh1		ETTh2		ETTm1		ETTm2		Traffic		Weather		Electricity		Illness		Exchange	
Paradigm ↓	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
BPTT	1.077	0.829	1.534	1.029	0.779	0.668	1.145	0.887	**0.670**	**0.376**	**0.338**	**0.385**	**0.289**	**0.384**	4.623	1.488	0.675	0.687
SPTT-Hybrid	**0.712**	**0.604**	**0.858**	**0.711**	**0.642**	**0.552**	**0.456**	**0.497**	0.855	0.501	0.355	0.401	0.361	0.440	**3.506**	**1.218**	**0.174**	**0.297**
SPTT-Endpoint	0.669	0.596	0.941	0.763	0.820	0.645	0.513	0.538	0.732	0.428	0.374	0.414	0.322	0.411	3.556	1.231	0.203	0.351
TBPTT-4	0.932	0.756	3.460	1.624	0.990	0.758	1.857	1.168	**0.797**	**0.440**	0.611	0.529	**0.270**	**0.365**	**4.490**	1.481	0.930	0.771
SPTT-Window-4	**0.776**	**0.681**	**1.991**	**1.108**	**0.940**	**0.715**	**0.994**	**0.805**	0.816	0.455	**0.373**	**0.424**	0.308	0.401	4.663	**1.480**	**0.472**	**0.522**
TBPTT-8	0.973	0.775	3.427	1.605	1.049	0.793	2.712	1.289	0.901	0.493	0.587	0.528	**0.303**	**0.384**	**4.618**	**1.526**	0.852	0.732
SPTT-Window-8	**0.755**	**0.676**	**1.983**	**1.120**	**0.982**	**0.761**	**1.132**	**0.855**	**0.851**	**0.487**	**0.349**	**0.416**	0.312	0.397	5.016	1.542	**0.595**	**0.595**

Specifically, on the Exchange dataset, SPTT-Hybrid achieved a 74% MSE improvement relative to BPTT-GRU (from 0.675 to 0.174), which corresponds to the significant improvement magnitudes observed in the LSTM experiments in the main text, indicating that SPTT's decoupled optimization mechanism can consistently function across different gated structures.

The SPTT-Endpoint phenomenon emphasized in the main text has also been correspondingly validated in the GRU experiments. This method, using only the final time step information, surpassed BPTT-GRU utilizing complete time sequences on multiple datasets, as exemplified by the 0.203 vs 0.675 comparative results on the Exchange dataset. The consistent emergence of this counterintuitive superior performance across two different RNN architectures further confirms the core viewpoint proposed in the main text: SPTT's decoupled computational paradigm can extract more effective optimization signals from limited information.

The advantage of SPTT-Window over TBPTT in the GRU experiments similarly validates the important findings regarding gradient inheritance mechanisms in the main text. The significant difference between SPTT-Window-4 and TBPTT-4 on the ETTh2 dataset (1.991 vs 3.460) maintains consistency with similar performance comparison patterns in the main text, proving that SPTT's superiority in handling memory discontinuity problems in long-term forecasting tasks is architecture-independent. This cross-architectural consistency indicates that SPTT addresses fundamental computational problems in RNN training, rather than technical optimizations specific to certain gating mechanisms.

D Performance Analysis of SPTT: Classification Tasks

Table 6 presents the performance comparison of the SPTT computing paradigm applied to a GRU network for classification tasks against BPTT. The performance of SPTT on classification tasks is nearly identical to the LSTM results reported in the main text, further confirming that SPTT's advantage in temporal pattern recognition stems from the intrinsic characteristics of its computational paradigm rather than from incidental synergies with a specific architecture. Variants of SPTT also achieve higher classification accuracy on GRU networks across most datasets. The results on the JapVowels dataset are particularly illustrative: SPTT-Hybrid improves accuracy from 89.5% with BPTT-GRU to 94.4%, yielding a 4.9% gain. This improvement mirrors the magnitude and pattern observed in the LSTM experiments, demonstrating that SPTT's decoupling mechanism offers stable, cross-architecture effectiveness in capturing discriminative temporal features.

Table 6: Classification Results Comparison on GRU

Dataset →	EthanolCon	FaceDetec	Handwriting	Heartbeat	JapVowels	PEMS-SF	SCP1	SCP2	SAD
Paradigm ↓	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
BPTT	25.3%	61.9%	4.4%	**75.1%**	89.5%	**87.8%**	**78.8%**	50.0%	97.3%
SPTT-Hybrid	**30.1%**	**63.1%**	**5.3%**	71.2%	**94.4%**	82.8%	74.2%	52.1%	**98.6%**
SPTT-Endpoint	28.6%	55.0%	5.3%	72.8%	89.7%	78.3%	74.9%	**52.4%**	68.8%
TBPTT-4	26.2%	48.7%	4.0%	72.4%	89.7%	76.0%	**77.1%**	50.8%	98.4%
SPTT-Window-4	**27.6%**	**50.7%**	**4.1%**	**73.8%**	**92.4%**	**79.6%**	75.6%	**51.3%**	**98.5%**
TBPTT-8	22.2%	50.3%	**14.6%**	71.4%	89.9%	76.9%	74.7%	48.3%	96.8%
SPTT-Window-8	**27.9%**	**50.9%**	5.7%	**74.0%**	**90.8%**	**77.2%**	**75.3%**	**53.4%**	**98.3%**

The SPTT-Endpoint phenomenon highlighted in the main text is likewise validated in our GRU experiments. Despite relying solely on the final time step, this approach surpasses BPTT-GRU on several datasets. Yet, as with the LSTM case, we also observe that SPTT-Endpoint performs worse than BPTT in certain classification tasks. This divergence pattern aligns exactly with the LSTM results reported in the main text. Such cross-architecture consistency provides further evidence for our conclusion that, in some classification tasks, the decisive information is distributed across the entire sequence rather than concentrated at the final step—a property determined by the data itself rather than the specific RNN architecture.

The comparison of SPTT-Window and TBPTT in the GRU architecture reproduces another key finding of the main text, confirming the advantage of SPTT in truncated settings. In tasks requiring the capture of long-term temporal patterns, SPTT's dynamic adaptability and memory inheritance enable it to identify complex dependencies more effectively, an advantage equally evident in GRU experiments. This consistent improvement across architectures offers strong support for the central claim of the main text: that the decoupling mechanism of SPTT provides a fundamental advantage in addressing temporal pattern recognition.

E Performance Analysis of SPTT: Imputation Tasks

In the numerical imputation task, when both SPTT and BPTT utilize complete memory, their performances are comparable. However, under memory discontinuity challenges, SPTT exhibits a clear advantage, being far less affected than BPTT. For example, on the ETTm1 dataset, with full memory, SPTT underperforms BPTT (MSE of 0.347 vs. 0.260). Yet once memory gaps are introduced, the advantage of SPTT becomes evident: with a block size of 4, SPTT achieves an MSE of 0.456 compared to BPTT's 0.493, and with a block size of 8, SPTT maintains the lead (0.672 vs. 0.715).

Table 7: Imputation Results Comparison on GRU

Dataset → Paradigm ↓	ETTh1		ETTh2		ETTm1		ETTm2		Electricity		Weather	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
BPTT	0.313	0.400	0.958	0.732	**0.260**	**0.350**	0.498	0.524	**0.264**	**0.360**	**0.112**	**0.182**
SPTT-Hybrid	**0.302**	**0.390**	**0.539**	**0.529**	0.347	0.397	**0.273**	**0.370**	0.354	0.438	0.154	0.240
SPTT-Endpoint	0.341	0.398	1.091	0.791	0.433	0.438	0.640	0.596	0.317	0.402	0.200	0.291
TBPTT-4	0.445	0.485	1.500	0.936	0.493	0.491	0.784	0.665	**0.273**	**0.371**	0.214	0.298
SPTT-Window-4	**0.347**	**0.416**	**0.671**	**0.623**	**0.456**	**0.478**	**0.368**	**0.464**	0.333	0.419	**0.185**	**0.274**
TBPTT-8	0.709	0.602	1.268	0.885	0.715	0.594	0.661	0.605	**0.278**	**0.380**	**0.221**	**0.312**
SPTT-Window-8	**0.630**	**0.536**	**0.911**	**0.738**	**0.672**	**0.572**	**0.426**	**0.502**	0.346	0.432	0.267	0.358

On GRU networks, the stable advantage of SPTT across different truncation granularities mirrors the observations reported in the main text, further confirming the architecture-independent universality of SPTT's robustness and computational stability. SPTT thus represents a fundamentally superior computational paradigm for addressing sequence memory problems. Its unique decoupled inheritance mechanism effectively mitigates the adverse effects of memory discontinuities.

F Performance Analysis of SPTT: Anomaly Detection Tasks

The anomaly detection experiments conducted on GRU networks provide strong validation of the key finding reported in the main text: SPTT maintains stable performance under extreme class imbalance. Under the highly challenging setting of a 1% anomaly rate, SPTT variants outperform traditional BPTT-GRU on four out of five datasets. This success rate is consistent with the LSTM results presented in the main text, demonstrating that the computational advantage of SPTT in detecting rare anomaly patterns arises from its decoupling mechanism rather than any specific gating design. By leveraging limited temporal information, SPTT is able to extract more effective anomaly detection signals.

These experiments collectively demonstrate that SPTT's core mechanism—decoupling the optimization direction and the update magnitude—works effectively across different RNN implementations. In both the three-gate LSTM and the two-gate GRU, SPTT achieves systematic improvements over traditional methods through

Table 8: Anomaly Detection Results Comparison on GRU

Dataset → Paradigm ↓	MSL F1-Score	PSM F1-Score	SMAP F1-Score	SMD F1-Score	SWAT F1-Score
BPTT	81.7%	94.7%	66.6%	77.7%	83.2%
SPTT-Hybrid	**82.2%**	**94.8%**	**67.4%**	**78.0%**	82.9%
SPTT-Endpoint	**82.2%**	94.1%	67.2%	77.9%	**83.5%**
TBPTT-4	81.6%	91.6%	53.7%	85.4%	83.1%
SPTT-Window-4	**82.0%**	**93.1%**	**54.0%**	**85.8%**	**83.2%**
TBPTT-8	82.0%	**92.7%**	67.0%	78.4%	**85.0%**
SPTT-Window-8	**82.1%**	92.4%	**67.2%**	**78.5%**	83.6%

its distinctive gradient decomposition and inheritance strategy. This architecture-agnostic pattern of gains provides solid empirical support for the main text's claim that SPTT constitutes a foundational training paradigm for the next generation of recurrent neural networks.

G SPTT Achieves Fast and Efficient Computation

To verify the universality of SPTT's computational advantage, we repeated the runtime comparison experiments on GRU networks. Figure 1 presents the computation time of SPTT and its variants against traditional BPTT under the GRU architecture. The experiments were conducted on the Sogou News and AG News datasets, using the same sequence length settings as in the LSTM experiments reported in the main text.

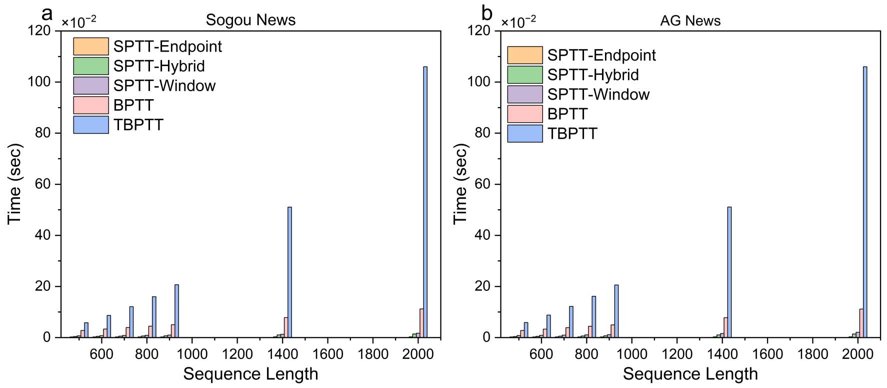

Fig. 1: Training time comparison across different paradigms with varying sequence lengths on Sogou News (a) and AG News (b).

In GRU networks, SPTT likewise demonstrates a significant speed advantage, which becomes increasingly pronounced as sequence length grows. Notably, the computational gains of SPTT under GRU even surpass those observed with LSTM. When the sequence length reaches 2000, SPTT-Hybrid runs approximately 8× faster than BPTT, while SPTT-Endpoint exhibits the most striking efficiency, with runtime nearly unaffected by sequence length and achieving about a 44× speedup over BPTT.

These findings indicate that SPTT's computational advantage is not tied to any particular RNN architecture but instead arises from its fundamentally innovative decoupled gradient computation mechanism. Whether in the complex gating structure of LSTM or the simplified design of GRU, SPTT leverages its streaming low-rank decomposition strategy to deliver substantial acceleration, thereby providing a unified and efficient solution for large-scale time series processing across diverse RNN architectures.

H SPTT Exhibits Robust Generalization Ability

To further examine the universality of SPTT's generalization advantage, we conducted corresponding experiments on GRU networks. Figure 2 compares the training and validation loss curves of SPTT and its variants against traditional BPTT under the GRU architecture. Two representative datasets were selected: Exchange Rate (prediction task) and Japanese Vowels (classification task).

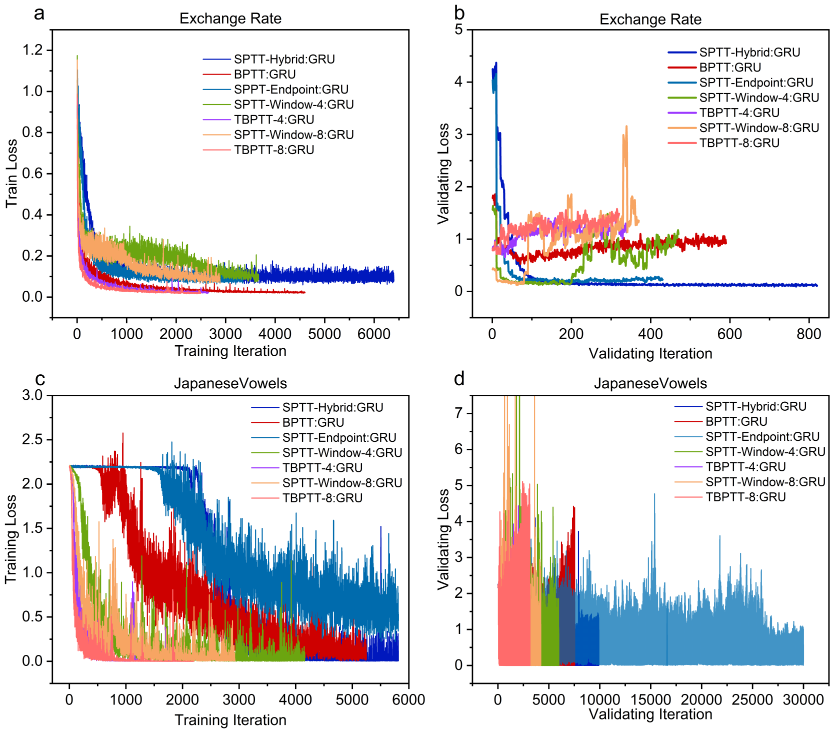

Fig. 2: Training and validation loss curves across paradigms. (a) Exchange Rate Training, (b) Exchange Rate Validation, (c) JapaneseVowels Training, (d) JapaneseVowels Validation.

The results remain highly consistent with those observed on LSTM networks, reaffirming the architecture-independent nature of SPTT's superior generalization performance. In the Exchange Rate prediction task (Figures 2a–b), SPTT variants achieve markedly lower validation losses than BPTT and its variants, while also exhibiting better convergence stability. In particular, SPTT-Hybrid and SPTT-Endpoint produce smoother validation curves with significantly reduced fluctuations, reflecting outstanding stability in generalization. Although the training loss of SPTT is slightly higher than that of BPTT, this outcome highlights the synergistic effect of SPTT's low-rank constraints, which in turn enhance the model's regularization capability.

In the Japanese Vowels classification task (Figures 2c–d), the generalization advantage of SPTT is even more pronounced. BPTT exhibits typical overfitting behavior, with training loss continuing to decrease while validation loss rises sharply, particularly in later stages where validation curves fluctuate heavily. By contrast, SPTT variants not only achieve lower validation losses but also maintain long-term training stability, with validation curves showing a relatively steady downward trajectory.

The decoupled computation mechanism of SPTT, together with the moderate randomness introduced by stochastic power iteration and Gaussian normalization, jointly contribute to stronger overfitting resistance in GRU networks.

Overall, these findings demonstrate that SPTT's generalization advantage is an inherent property of its algorithmic design rather than a byproduct of a specific network architecture. Whether applied to LSTM or GRU, SPTT leverages its distinctive low-rank decomposition and incremental learning mechanism to effectively mitigate overfitting, enabling the model to acquire more robust and transferable feature representations and thereby achieve superior generalization performance on unseen data.

I SPTT Exhibits Favorable Low-Rank Properties

To assess the effectiveness of SPTT's low-rank decomposition mechanism across different RNN architectures, we conducted a rank sensitivity analysis on GRU networks. Figure 3 illustrates the impact of varying rank settings on SPTT performance under the GRU architecture, covering both prediction tasks (Electricity, ETTh1, Traffic, Weather) and classification tasks (Face Detection, Heart Beat, Japanese Vowels, PEMS-SF).

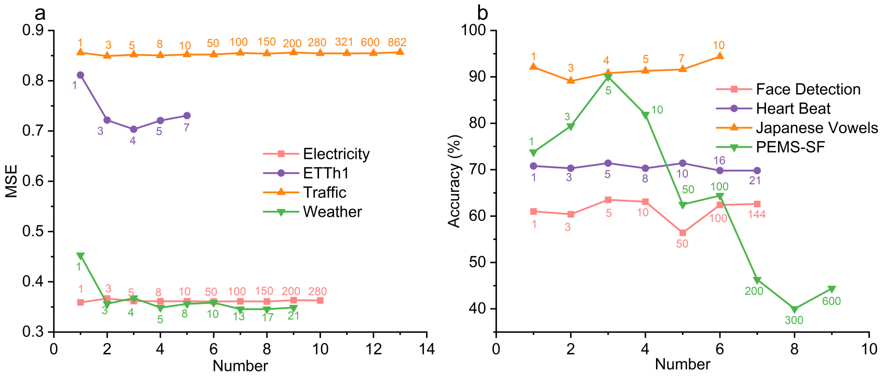

Fig. 3: The effect of different ranks on the performance of GRU. The numbers in the corresponding colors in the figure indicate the ranks under consideration for each dataset in the current experiment. The experimental setup follows the standard that the selected rank does not exceed the feature dimensionality of the corresponding dataset.

The results closely mirror the findings from LSTM experiments, further confirming the general prevalence of the low-rank property in neural network gradient matrices. Within the GRU architecture, SPTT also achieves optimal performance under relatively low-rank settings, with most datasets attaining their best results when the rank lies between 3 and 10. This observation once again validates our core theoretical assumption: regardless of whether the network is LSTM or GRU, the essential information of the gradient matrix is concentrated in only a few dominant directions.

For prediction tasks (Figure 3a), similar rank sensitivity patterns are observed across datasets. The ETTh1 and Weather datasets achieve optimal performance at ranks 3–5, after which the results stabilize; the Traffic dataset maintains consistently strong performance across the entire range of ranks, reflecting the highly decomposable low-rank structure of its data features; the Electricity dataset achieves stable optimal performance in the rank range of 3–8. Notably, when the rank exceeds a certain threshold, performance on some datasets shows a slight decline, which aligns with theoretical expectations—excessively high ranks tend to capture additional noise.

For classification tasks (Figure 3b), the influence of the rank parameter is more diverse, reflecting different requirements of feature complexity across tasks. The Japanese Vowels and Heart Beat datasets achieve strong performance under relatively low-rank settings, suggesting that their discriminative features are highly concentrated. The PEMS-SF dataset exhibits an inverted-U performance curve, peaking at rank 5 and then declining significantly, indicating the presence of a well-defined optimal rank range. The Face Detection dataset, by contrast, remains relatively stable across rank settings, suggesting that this task is less sensitive to the choice of rank.

These findings not only validate the theoretical foundation of SPTT but also provide practical guidance: when applying SPTT to GRU networks, one can adopt the same conservative rank-setting strategy as in LSTM, ensuring high performance while further reducing computational cost. This architecture-insensitive rank selection strategy highlights the flexibility and adaptability of the SPTT algorithm.

J Analysis of SPTT-Hybrid's Incremental Update Characteristics on GRU Networks

To verify the consistency of SPTT-Hybrid's incremental learning mechanism across different RNN architectures, we conducted time block partitioning sensitivity analysis on GRU networks. Figure 4 illustrates the impact of different time block settings on SPTT-Hybrid performance, with experiments covering both long-term prediction and classification tasks. The experimental results remain highly consistent with findings on LSTM networks, further confirming the architecture-agnostic nature of SPTT-Hybrid's incremental learning mechanism.

In long-term prediction tasks (Figure 4a), all datasets exhibit a pronounced performance degradation trend as the number of time blocks increases. The MSE of the ETTh1 dataset increases from 0.673 to 0.8081, representing approximately 20% performance deterioration; other datasets demonstrate similar degradation patterns. This reaffirms the strong dependence of prediction tasks on historical information completeness. Notably, the performance degradation magnitude under GRU architecture is slightly mitigated compared to LSTM, potentially attributable to the efficiency advantages of GRU's simplified gating mechanism.

In classification tasks (Figure 4b), GRU networks exhibit stability patterns similar to LSTM. The Japanese Vowels dataset maintains accuracy above 94% across the

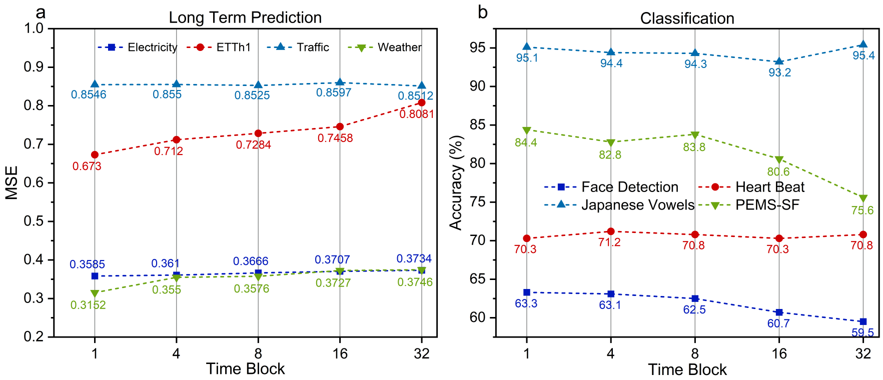

Fig. 4: Impact of Incremental Update Frequency on SPTT Performance in GRU Networks.

entire time block range, while the Heart Beat dataset also remains stable. However, the PEMS-SF dataset demonstrates exceptional sensitivity, with accuracy declining from 84.4% to 75.6%, which may reflect the special dependence of traffic flow classification tasks on temporal context completeness. These results indicate that SPTT-Hybrid's incremental update characteristics possess favorable architectural generalization capability.

Based on the aforementioned experimental observations, we recommend that in practical applications, when sufficient historical data is available, SPTT-Hybrid should adopt a one-shot computation strategy, i.e., performing complete learning based on all available historical data to compute optimal optimization directions and magnitudes in a single operation, thereby fully exploiting the performance potential of the SPTT algorithm.

K SPTT Mitigates the Impact of Memory Fragmentation

To further validate SPTT's robustness against memory fragmentation across different RNN architectures, we conducted comparative experiments between SPTT-Window and TBPTT on GRU networks under various time block partitioning scenarios. Figure 5 demonstrates the performance comparison when both methods face memory discontinuity challenges, encompassing long-term prediction and classification tasks.

The experimental results on GRU networks strongly corroborate our findings on LSTM architectures, reaffirming SPTT's superior resilience to memory fragmentation. In long-term prediction tasks (Figure 5a), traditional BPTT exhibits severe performance instability as time block partitioning becomes more granular. The ETTh2 dataset shows particularly dramatic performance collapse under BPTT, with MSE spiking from approximately 1.5 to over 4.2, demonstrating typical learning collapse

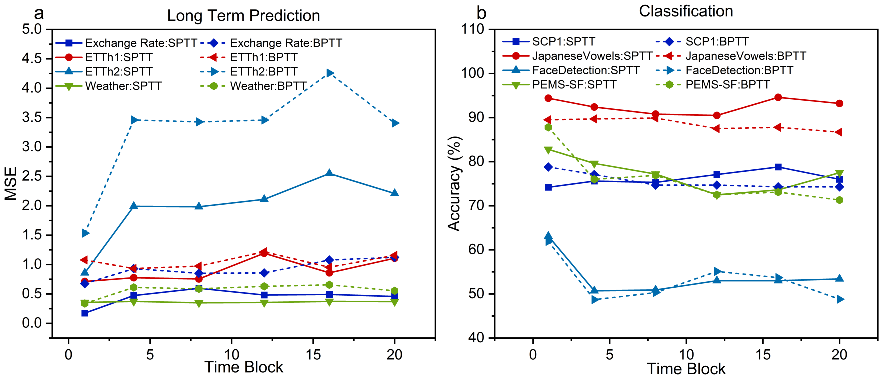

Fig. 5: Impact of Time Block Quantity on SPTT Performance

phenomenon caused by memory fragmentation. In contrast, SPTT maintains remarkably stable performance across different time block settings, with all datasets showing minimal performance fluctuation.

In classification tasks (Figure 5b), the advantage of SPTT's memory inheritance mechanism becomes even more pronounced. While BPTT suffers from significant accuracy degradation and high variance across different time block configurations, SPTT variants maintain consistent performance levels. Notably, the Japanese Vowels dataset shows that SPTT achieved stable accuracy above 90%, while BPTT's performance continuously deteriorated. These results demonstrate that SPTT's decoupled gradient computation and inheritance mechanism effectively mitigate the detrimental effects of memory fragmentation, providing a more robust solution for online learning scenarios where memory discontinuity is unavoidable.