A APPENDIX

A.1 PROOFS

Proof of Proposition 1 Denote the sampled $u_t' = u_t + \varepsilon_t$, where ε_t is the sampling error caused by variation in the sampling points. Consider the propagation of the error in the output values $\{y_k\}_{k=1}^L$:

$$\begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_L' \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{CB}} & \mathbf{0} & \cdots & \mathbf{0} \\ \overline{\boldsymbol{CAB}} & \overline{\boldsymbol{CB}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\boldsymbol{CA}}^{T-1}\overline{\boldsymbol{B}} & \overline{\boldsymbol{CA}}^{T-2}\overline{\boldsymbol{B}} & \cdots & \overline{\boldsymbol{CB}} \end{bmatrix} \begin{bmatrix} u_1 + \varepsilon_1 \\ u_2 + \varepsilon_2 \\ \vdots \\ u_T + \varepsilon_t \end{bmatrix} \tag{29}$$

then

$$\begin{aligned} \|y_t' - y_t\| &= \left\| \overline{\boldsymbol{CA}}^{t-1}\overline{\boldsymbol{B}}\varepsilon_1 + \overline{\boldsymbol{CA}}^{t-2}\overline{\boldsymbol{B}}\varepsilon_2 + \cdots + \overline{\boldsymbol{CB}}\varepsilon_t \right\| \\ &\le \left\|\overline{\boldsymbol{A}}^{t-1}\right\| \left\|\overline{\boldsymbol{B}}\right\| |\varepsilon_1| + \left\|\overline{\boldsymbol{A}}^{t-2}\right\| \left\|\overline{\boldsymbol{B}}\right\| |\varepsilon_2| + \cdots + \left\|\overline{\boldsymbol{B}}\right\| |\varepsilon_t| \\ &\le |\lambda_{\max}|^{t-1} b\varepsilon_1 + |\lambda_{\max}|^{t-2} b\varepsilon_2 + \cdots + b\varepsilon_t \end{aligned} \tag{30}$$

Note that if $\lambda_{\max} \ge 1$, $\lim_{t\to\infty} \|y_t' - y_t\|$ becomes unbounded. If $|\lambda_{\max}| < 1$, then we have

$$\begin{aligned} \|\boldsymbol{x}_t\| &= \left\| \overline{\boldsymbol{A}}^{L-1}\overline{\boldsymbol{B}}u_1 + \overline{\boldsymbol{A}}^{L-2}\overline{\boldsymbol{B}}u_2 + \cdots + \overline{\boldsymbol{B}}u_t \right\| \\ &\le \left\|\overline{\boldsymbol{A}}^{L-1}\right\| \left\|\overline{\boldsymbol{B}}\right\| |u_1| + \left\|\overline{\boldsymbol{A}}^{L-2}\right\| \left\|\overline{\boldsymbol{B}}\right\| |u_2| + \cdots + \left\|\overline{\boldsymbol{B}}\right\| |u_t| \\ &\le |\lambda_{\max}|^{L-1} b\zeta + |\lambda_{\max}|^{L-2} b\zeta + \cdots + b\zeta, \end{aligned} \tag{31}$$

thus

$$\lim_{t\to\infty} \|\boldsymbol{x}_t\| \le \lim_{t\to\infty} \left(|\lambda_{\max}|^{L-1} b\zeta + |\lambda_{\max}|^{L-2} b\zeta + \cdots + b\zeta \right) = \frac{b\zeta}{1 - |\lambda_{\max}|} < \lim_{t\to\infty} \|\boldsymbol{x}_t\| \tag{32}$$

contradicts the assumption, therefore there must be $|\lambda_{\max}| >= 1$, which also implies that $\lim_{t\to\infty} \|y_t' - y_t\|$ is unbounded.

Remark Note that imposing the constraint $|\lambda_{\max}| < 1$ on the state space model will cause the initial input u_{t_0} to tend to zero as it propagates ($\overline{\boldsymbol{A}}^{t-t_0}\overline{\boldsymbol{B}}u_{t_0} \underset{t-t_0\to\infty}{\longrightarrow} 0$). This causes all previous states to rapidly decay to 0 during the propagation, thus severely limits the long-term memory capacity of the model.

Proof of Theorem1 Taking into account the error propagation in latent states of the S4 model, the grid deviation error emerges from signal misalignment and can be considered as an additional disturbance term. Assuming that the actual sampled value, denoted as u', satisfies the relationship $u_t' = u_t + \varepsilon_t$, where ε_t represents the error term, we can have

$$\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_T \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{B}} & \mathbf{0} & \cdots & \mathbf{0} \\ \overline{\boldsymbol{AB}} & \overline{\boldsymbol{B}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\boldsymbol{A}}^{T-1}\overline{\boldsymbol{B}} & \overline{\boldsymbol{A}}^{T-2}\overline{\boldsymbol{B}} & \cdots & \overline{\boldsymbol{B}} \end{bmatrix} \begin{bmatrix} u_1 + \varepsilon_1 \\ u_2 + \varepsilon_2 \\ \vdots \\ u_T + \varepsilon_t \end{bmatrix} \tag{33}$$

observe that

$$\begin{aligned} \boldsymbol{x}_t &= \overline{\boldsymbol{A}}^{t-1}\overline{\boldsymbol{B}}(u_1 + \varepsilon_1) + \overline{\boldsymbol{A}}^{t-2}\overline{\boldsymbol{B}}(u_2 + \varepsilon_2) + \cdots + \overline{\boldsymbol{B}}(u_t + \varepsilon_t) \\ &= \overline{\boldsymbol{A}}^{t-1}\overline{\boldsymbol{B}}u_1 + \overline{\boldsymbol{A}}^{t-2}\overline{\boldsymbol{B}}u_2 + \cdots + \overline{\boldsymbol{B}}u_t + L(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_t), \end{aligned} \tag{34}$$

where $L(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_t) = \overline{\boldsymbol{A}}^{t-1}\overline{\boldsymbol{B}}\varepsilon_1 + \overline{\boldsymbol{A}}^{t-2}\overline{\boldsymbol{B}}\varepsilon_2 + \cdots + \overline{\boldsymbol{B}}\varepsilon_t$. Consider its continuous form and drawing upon the controller concept in EMC theory, we consider the following state propagation:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\left(\boldsymbol{x}(t) + \int_0^t \boldsymbol{k}(t-l)\varepsilon(l)dl\right) + \boldsymbol{B}u(t), \tag{35}$$

where $\boldsymbol{k}$ is a coefficient matrix that varies over time, and has the same shape as $\overline{\boldsymbol{B}}$. Owing to the accumulation of errors in the time domain, we introduce a modifiable factor denoted as $h([t-\tau,t])$ with backtracking capability to regulate the input. Specifically, the controlled input is defined as $u_{adj}(t) = h([l-\tau,l])u(t)$. then we have

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\left(x(t) + \int_0^t \boldsymbol{k}(t-l)h([l-\tau,l])\varepsilon(l)dl\right) + \boldsymbol{B}h([t-\tau,t])u(t), \tag{36}$$

then $h_\tau(t)$ has the ability to adjust the errors with coefficients carrying temporal phases. Taking into account the following observer used for sampling:

$$\dot{\boldsymbol{z}}(t) = \boldsymbol{A}\boldsymbol{z}(t) + \boldsymbol{B}h([t-\tau,t])(u(t)+\varepsilon(t)), \tag{37}$$

denote $\boldsymbol{e}(t) = \boldsymbol{x}(t) - \boldsymbol{z}(t)$, we have

$$\dot{\boldsymbol{e}}(t) = \boldsymbol{A}\boldsymbol{e}(t) + \boldsymbol{A}\int_0^t \boldsymbol{k}(t-l)h([l-\tau,l])\varepsilon(l)dl - \boldsymbol{B}h([t-\tau,t])\varepsilon(t) \tag{38}$$

Consider the Lyapunov function $\mathcal{L}_{\boldsymbol{e}}(t) = \boldsymbol{e}^\top(t)\boldsymbol{P}\boldsymbol{e}(t)$, where $\boldsymbol{P}$ is a positive definite symmetric matrix, we can obtain

$$\begin{aligned}\frac{d\mathcal{L}_{\boldsymbol{e}}(t)}{dt} &= 2e^\top(t)\boldsymbol{P}\dot{\boldsymbol{e}}(t)\\ &= 2e^\top(t)\boldsymbol{P}\left(\boldsymbol{A}\boldsymbol{e}(t) + \boldsymbol{A}\int_0^t \boldsymbol{k}(t-l)h([l-\tau,l])\varepsilon(l)dl - \boldsymbol{B}h([t-\tau,t])\varepsilon(t)\right)\\ &= \boldsymbol{e}^\top(t)\left(\boldsymbol{P}\boldsymbol{A} + \boldsymbol{e}^\top(t)\boldsymbol{A}^\top\boldsymbol{P}\right)\boldsymbol{e}(t) + \Lambda(t)\\ &= \boldsymbol{e}^\top(t)\left(\boldsymbol{P}\boldsymbol{A} + \boldsymbol{A}^\top\boldsymbol{P}\right)\boldsymbol{e}(t) + \Lambda(t),\end{aligned} \tag{39}$$

where

$$\begin{aligned}\Lambda(t) &= \boldsymbol{e}^\top(t)\boldsymbol{A}\int_0^t \boldsymbol{k}(t-l)h([l-\tau,l])\varepsilon(l)dl - \boldsymbol{e}^\top(t)\boldsymbol{B}h([t-\tau,t])\varepsilon(t)\\ &= \|\boldsymbol{e}(t)\|\,\|\boldsymbol{A}\|\int_0^t \|\boldsymbol{k}(t-l)\|\,|h([l-\tau,l])|\,|\varepsilon(l)|\,dl + \|\boldsymbol{e}(t)\|\,\|\boldsymbol{B}\|\,|h([t-\tau,t])|\,|\varepsilon(t)|\\ &\le \|h_\tau\|\,\|\boldsymbol{e}(t)\|\left(\int_0^t \|\boldsymbol{k}(t-l)\|\,|\varepsilon(l)|\,dl + \|\boldsymbol{B}\|\,|\varepsilon(t)|\right)\end{aligned} \tag{40}$$

Hence, selecting a value of $|h([t-\tau,t])| < 1$ strengthens the stability of the system, while $h([t-\tau,t]) \equiv 1$ corresponds to the case without a controller. Additionally, choosing a larger τ value can further enhance the control performance.

A.2 NSS IN 5-LAYERS S4

Due to space constraints, we present the analysis of the deep S4 model here. Specifically, we conducted an experiment on a 5-layer S4 model, extending from the experiment described in Section 2.3. We plotted the results of the hidden states in the first layer and observed the presence of the NSS issue in the 5-layer S4 model, as depicted in Figure 8.2. Notably, the S4 model without Memory Replay exhibited a significant NSS phenomenon. In contrast, the S4+ model with Memory Replay demonstrated highly stable hidden states, as illustrated in Figure 8.4. The sum of absolute values of the states at each time step decreased from 10^2 to 10^1, and the output error under perturbation was also reduced (Figure 8.2).

A.3 EXPERIMENT DETAILS

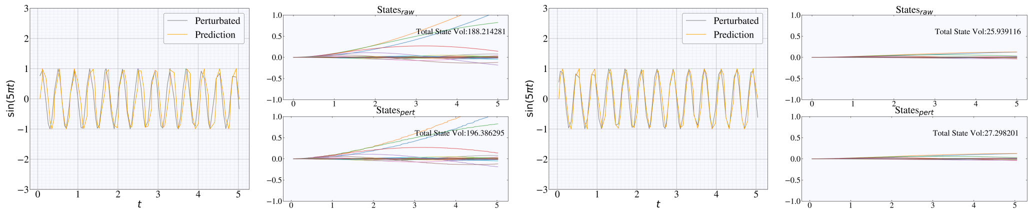

Figure 6: Comparative results for Memory Reply.

Table 5: Detailed training settings used in our experiments.

	Autoregressive language modelling	**Bidirectional language modelling**
Data used	Wikitext-103	Wikitext-103
Tokenizer method	BPE	BPE
Vocab size	50265	50265
Sequence length	512	512
Batch size	64	64
Total updates	50,000	50,000
Warmup steps	3,000	3,000
Peak learning rate	5e-4	5e-4
Lr scheduler	Inverse sqrt	Polynomial decay
Optimizer	Adam	Adam
Adam ϵ	1e-8	1e-6
Adam (β_1, β_2)	(0.9, 0.98)	(0.9, 0.98)
Weight decay	0.2 for TNN, 0.1 for others	0.2 for TNN, 0.1 for others
Gradient clip norm	1.0	1.0
Dropout	0.1	0.1

Table 6: Detailed training settings used in LRA tasks.

	Retrieval	ListOps	Text	Image	Pathfinder
Num blocks	6	6	4	8	4
Embedding dimension	128	80	128	128	128
Max length	4000	2048	4096	1024	1024
Batch size	20	50	16	64	64
Total epochs	20	40	32	200	200
Learning rate	1e-3	1e-4	1e-3	4e-3	2e-4
Weight decay	0.1	0.0	5e-2	5e-2	0.0
Dropout	0.1	0.0	0.1	0.1	0.1