CERTIFIED RATIONAL NEURAL LAYERS: STABILITY AND PRESCRIBED GAIN FOR SEQUENCE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-range sequence models require expressive memory together with efficient parallel computation. *Rational transfer-function layers* represent long-memory linear dynamics using finitely many coefficients and enable a state-free, full-sequence implementation through parallel FFT-based convolution. However, unconstrained rational filters may become unstable during training, while the Montel-certified RTF parameterization is conservative and neither formulation directly controls input–output amplification. We introduce *Certified Rational Neural Layers* (CERNEL), based on two complementary constructions. First, an unconstrained reflection-coefficient parameterization represents every real monic Schur-stable polynomial of a prescribed degree. It guarantees stability throughout training without computationally expensive and conservative projections. Second, a finite Schur-recursive construction enforces a user-specified induced ℓ_2-gain bound on the rational operator. Both constructions admit parallel implementation over the filter order and FFT-based full-sequence evaluation. We extensively evaluate CERNEL across several experiments and benchmark tasks spanning long-range sequence modeling, nonlinear system identification, and visuomotor control. Across these settings, CERNEL remains competitive with unconstrained rational models while providing stability or gain guarantees by construction, and improves over conservative certified parameterizations when their representational restrictions become limiting.

1 INTRODUCTION

Structured linear operators have become an important component of long-range sequence models because they combine dynamical memory with efficient sequence computation (Gu et al., 2021; Smith et al., 2022; Gu & Dao, 2023; Parnichkun et al., 2024; Zhu et al., 2024; Lahoti et al., 2026). A finite-dimensional linear time-invariant operator can be represented either through a state-space recurrence or, equivalently, by a rational transfer function (RTF). Stability and input–output amplification are properties of the underlying operator and must therefore be controlled regardless of which representation is used. The transfer-function representation makes these properties explicit through poles, zeros, and frequency response, while enabling direct convolutional evaluation through parallel FFT-based multiplication. The same operator admits an equivalent recurrent state-space realization, but the transfer-function implementation avoids explicit propagation of that latent state. Once the transfer kernel is constructed, its computational and memory costs during full-sequence inference are largely independent of the realization dimension (Forgione & Piga, 2021; Parnichkun et al., 2024).

In a state-space representation, Schur stability is determined by the eigenvalues of the state matrix; in the equivalent multi-input-multi-output (MIMO) transfer-function representation, it is determined by the roots of the denominator polynomial. Unconstrained parameterizations of either representation may therefore produce unstable operators during training, causing exploding activations and gradients, numerical overflow, and failed optimization (Bengio et al., 1994; Pascanu et al., 2013). Structured State-space Models (SSMs) typically address this issue by restricting the spectrum of the state matrix in a unit disk Gu et al. (2021; 2022); Orvieto et al. (2023), while rational layers can impose sufficient constraints directly on the denominator coefficients Parnichkun et al. (2024); Hayes et al. (2024); Subramani et al. (2022). For example, the Montel condition proposed in RTF

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

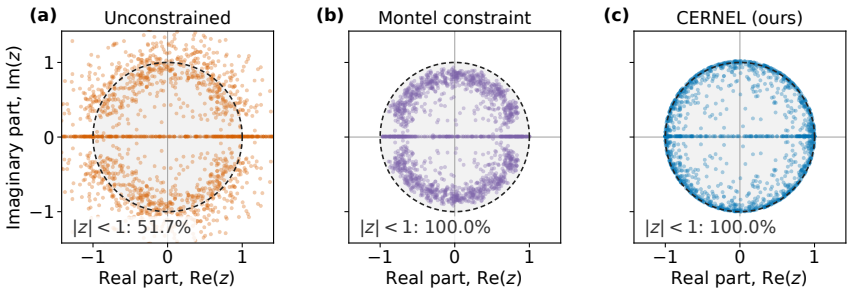

Figure 1: **Stable rational parameterizations.** Unconstrained denominator coefficients can yield unstable poles, while Montel's sufficient condition guarantees stability only over a conservative subset of the Schur-stable region. CERNEL uses reflection coefficients to parameterize the complete class of Schur-stable denominators of the chosen order: every parameter value is stable, and every such denominator is representable.

guarantees Schur stability (Parnichkun et al., 2024), but is conservative: it excludes valid stable denominators, including slowly decaying dynamics with poles close to the unit circle. As illustrated in Figure 1, it achieves certification at the cost of restricting the class of stable dynamics that can be learned.

Stability alone, however, does not control how strongly a dynamical operator amplifies its inputs. The induced ℓ_2 gain is the standard worst-case energy-amplification measure in input–output and $\mathcal{H}_\infty$ analysis. Bounding ℓ_2 gain limits disturbance-to-output amplification and supports compositional stability analysis through small-gain and dissipativity arguments (van Waarde & Sepulchre, 2022; Noori et al., 2024; Junnarkar et al., 2026; Massai et al., 2025; Zakwan & Ferrari-Trecate, 2026). Two Schur-stable systems can have identical poles yet markedly different induced ℓ_2 gains since it affects the response of the system to different frequencies in the input. Closely related, ℓ_2-Lipschitz bounds are used in learned operators to limit sensitivity to input perturbations and to certify interconnections of learned components (Wang & Manchester, 2023). Consequently, a small finite-energy input perturbation may produce a much larger output deviation even when the nominal responses of two stable systems are nearly indistinguishable. Figure 2 illustrates this distinction: two systems with the same pole radius, 0.985, have induced gains of approximately 2.0 and 21.1, and amplify the same 1% input perturbation very differently. This motivates rational sequence operators whose denominators are stable and whose induced ℓ_2 gain is bounded by a prescribed value.

To address these two limitations, we introduce Certified Rational Neural Layers (CERNEL). CERNEL uses classical reflection-coefficient coordinates to map unconstrained parameters bijectively onto the set of real monic Schur-stable denominator polynomials of a prescribed order. This removes the conservatism of sufficient coefficient constraints while guaranteeing stability at every optimization iterate. When explicit control of disturbance amplification is required, we additionally provide a Schur-recursive parameterization that guarantees $\|H\|_{\mathcal{H}_\infty} < \gamma$ for a user-specified bound γ. This gain-bounded construction is sufficient rather than complete, but its certificate also holds throughout training.

In practice, CERNEL serves as a certified linear temporal component within a broader nonlinear sequence model. For multivariate sequences, single-input-single-output (SISO) certified filters are assembled into a MIMO transfer matrix. Although the reflection-coefficient and Schur constructions are recursive in the filter order, we reformulate both as parallel scans, reducing their sequential depth from linear to logarithmic Blelloch (1990); Smith et al. (2022). Once constructed, the filters are applied over the sequence through parallel FFT-based convolution, without explicitly propagating a latent state. Figure 3 illustrates one possible instantiation, where the CERNEL operator is embedded in a gated residual block; other surrounding architectures can be used without changing the rational parameterization. The stability and gain guarantees apply directly to the rational operator. They can be extended to the complete nonlinear block by combining CERNEL with nonlinear components having prescribed incremental ℓ_2-gain, such as Lipschitz-bounded neural layers (Wang & Manchester, 2023).

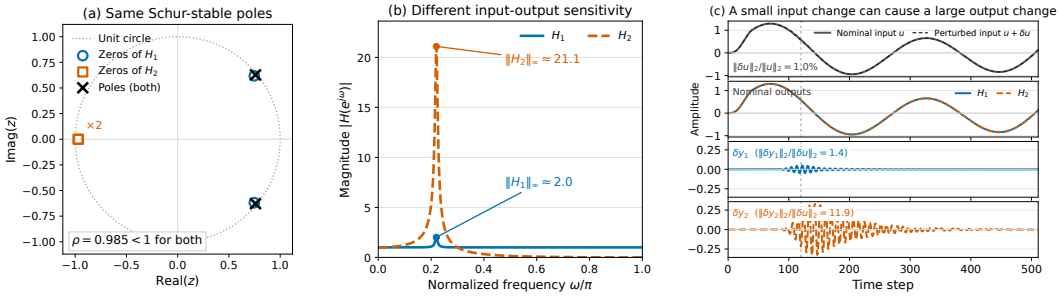

Figure 2: **Stability does not imply low input–output sensitivity.** **(a)** Two systems $H_1(z)$ and $H_2(z)$ share identical Schur-stable poles with maximum pole modulus 0.985. **(b)** Their induced ℓ_2 gains differ by more than an order of magnitude. **(c)** The same 1% input perturbation produces a much larger output deviation in the high-gain system.

Our main contributions are:

- We adapt classical reflection-coefficient coordinates to rational sequence layers named CERNEL, replacing the conservative Montel condition with a complete parameterization of real monic Schur-stable polynomials.
- We instantiate a complementary finite inverse-Schur construction with prescribed induced ℓ_2 gain and derive associative parallel-scan implementations of both recursions over the filter order.
- We evaluate the resulting layers on controlled long-memory tasks, Long Range Arena (Tay et al., 2020), nonlinear system identification, and visuomotor control (Chi et al., 2025).

A detailed discussion of related work and the positioning of CERNEL with respect to SSMs, rational neural architectures, and certified dynamical learning is provided in Appendix A.

2 Preliminaries

Rational sequence operators. Consider a discrete-time input sequence $u = (u_t)_{t\in\mathbb{N}}$, with $u_t \in \mathbb{R}^{d_{\text{in}}}$, and an output sequence $y = (y_t)_{t\in\mathbb{N}}$, with $y_t \in \mathbb{R}^{d_{\text{out}}}$. A causal linear time-invariant operator $\mathcal{G}$ acts as

$$y_t = (\mathcal{G}u)_t = \sum_{k=0}^{t} h_k u_{t-k}, \qquad h_k \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}},$$

where h_k is the MIMO *impulse response*. A finite-dimensional state-space realization is

$$x_{t+1} = Ax_t + Bu_t, \qquad y_t = Cx_t + Du_t,$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times d_{\text{in}}}$, $C \in \mathbb{R}^{d_{\text{out}}\times n}$, and $D \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$ are the state-space matrices and $x \in \mathbb{R}^n$ is called the hidden state. Unless stated otherwise, we use the standard zero-initial-state convention $x_0 = 0$ when defining the input–output operator and its induced gain. This is not a restriction on arbitrary initialized trajectories: a nonzero initial state contributes the separate homogeneous response $CA^t x_0$. Under $x_0 = 0$, the impulse response satisfies $h_0 = D$ and $h_k = CA^{k-1}B$ for $k \geq 1$, and the transfer function is

$$H(z) = D + C(zI - A)^{-1}B.$$

Each entry of $H(z)$ is a proper rational function, meaning that the degree of its numerator does not exceed the degree of its denominator. Such operators admit equivalent recurrent and convolutional implementations and can be used as differentiable sequence layers (Forgione & Piga, 2021; Parnichkun et al., 2024).

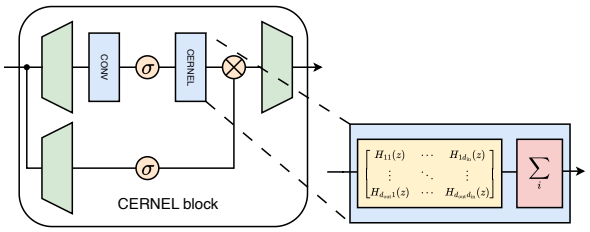

Figure 3: **One CERNEL architecture.** A MIMO bank of certified rational filters is embedded within a gated nonlinear sequence block. The certificates apply directly to the rational operator.

Stability and induced gain. Let $\ell_2^d := \left\{u = (u_t)_{t\in\mathbb{N}} : u_t \in \mathbb{R}^d,\ \sum_{t=0}^{\infty} \|u_t\|_2^2 < \infty\right\}$, $\|u\|_{\ell_2} := \left(\sum_{t=0}^{\infty} \|u_t\|_2^2\right)^{1/2}$, where $\|\cdot\|_2$ denotes the Euclidean vector norm. A rational operator is Schur stable when all poles of $H(z)$ lie strictly inside the unit disk and a state-space realization is Schur stable when $\rho(A) < 1$, where $\rho(\cdot)$ denotes the spectral radius. For a stable linear time-invariant (LTI) operator, its induced ℓ_2 gain is

$$\|G\|_{\ell_2\to\ell_2} = \|H\|_{\mathcal{H}_\infty}{}^{1} := \sup_{\omega\in[0,2\pi]} \bar{\sigma}\big(H(e^{\mathrm{i}\omega})\big)\,, \tag{1}$$

where $\bar{\sigma}(\cdot)$ denotes the largest singular value. Therefore, $\|H\|_{\mathcal{H}_\infty} \le \gamma$ guarantees

$$\|Gu\|_{\ell_2} \le \gamma\|u\|_{\ell_2}, \qquad u \in \ell_2^{d_{\mathrm{in}}}, \tag{2}$$

under zero initial conditions. Thus, γ directly bounds the worst-case energy amplification of finite-energy disturbances.

3 Certified Rational Neural Layers

We instantiate two complementary SISO constructions from unconstrained parameters. The first parameterizes all real monic Schur-stable polynomials of a prescribed degree, whereas the second constructs a sufficient family of stable rational operators satisfying $\|H\|_{\mathcal{H}_\infty} < \gamma$. Both certificates hold for every parameter value, and therefore throughout optimization, without projection or post-training verification. We then assemble the resulting SISO transfer functions into MIMO operators.

3.1 Complete stable-polynomial parameterization

Consider the SISO rational operator

$$H_\Theta(z) = h_0 + \frac{b_1 z^{-1} + \cdots + b_n z^{-n}}{1 + a_1 z^{-1} + \cdots + a_n z^{-n}}. \tag{3}$$

The feedthrough coefficient h_0 and numerator coefficients $b_1, \ldots, b_n$ are freely trainable. Since the poles of H_Θ depend only on its denominator, stability can be enforced by parameterizing the denominator independently of the numerator.

We use classical reflection coefficients (Nurges & Rustern, 2005; Subramani et al., 2022). For a fixed degree, these coefficients provide a one-to-one correspondence between real monic Schur-stable polynomials and vectors $(k_1, \ldots, k_n) \in (-1, 1)^n$. The following construction maps unconstrained trainable parameters to these bounded coordinates.

Proposition 1 (Complete stable-polynomial parameterization) *Let $\Theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ be the free parameters. Define the reflection coefficients $k_i = \tanh(\theta_i)$ for $i = 1, \ldots, n$. At recursion stage i, let $a_j^{(i)}$ denote the coefficient multiplying z^{i-j} in the intermediate polynomial*

$$p_i(z) = z^i + \sum_{j=1}^{i} a_j^{(i)} z^{i-j}.$$

[1] $\mathcal{H}_\infty$ is the Hardy space of bounded analytic transfer functions on the unit disk in $q = z^{-1}$; its norm equals the supremum of the boundary frequency response.

Initialize $a_1^{(1)} = -k_1$. *For* $i = 2, \dots, n$, *recursively define*

$$a_i^{(i)} = -k_i, \qquad a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \quad j = 1, \dots, i-1.$$

Then, $p_\Theta(z) = z^n + a_1^{(n)} z^{n-1} + \cdots + a_n^{(n)}$ *is Schur stable for every* $\Theta \in \mathbb{R}^n$. *Conversely, every real monic Schur-stable polynomial of degree* n *admits a unique representation of this form.*

Proposition 1 guarantees stability at every optimization iterate without projection. Because equation (3) is proper, stability also implies $\|H_\theta\|_{\mathcal{H}_\infty} < \infty$. Its value, however, remains uncontrolled because the numerator and feedthrough coefficients are freely trainable. The proof is provided in Appendix B.1.

3.2 Prescribed induced ℓ_2-gain

For applications requiring an explicit amplification bound, we use an alternative Schur-function parameterization. For a prescribed gain bound $\gamma > 0$, we parameterize the normalized transfer function using a finite inverse Schur recursion (Alpay & Vinnikov, 2010, The Schur Algorithm in Terms of System Realizations). The classical Schur algorithm associates contractive analytic functions on the unit disk with parameters lying in the open unit disk. We adapt this construction to obtain a differentiable rational parameterization with an explicit induced ℓ_2-gain bound.

Theorem 1 (Prescribed-gain Schur parameterization) *Given* $\gamma > 0$, *let* $\Theta = (\theta_0, \dots, \theta_n) \in \mathbb{R}^{n+1}$, $k_i = \tanh(\theta_i)$, *and* $q = z^{-1}$. *Define*

$$S_n(q) = k_n, \qquad S_i(q) = \frac{k_i + q S_{i+1}(q)}{1 + k_i q S_{i+1}(q)}, \quad i = n-1, \dots, 0. \tag{4}$$

Then,

$$H_\Theta(z) = \gamma S_0(z^{-1})$$

is a causal stable rational operator of degree at most n *satisfying*

$$\|H_\Theta\|_{\mathcal{H}_\infty} < \gamma, \qquad \|H_\Theta u\|_{\ell_2} < \gamma \|u\|_{\ell_2}, \quad 0 \neq u \in \ell_2.$$

Theorem 1 therefore defines an unconstrained, certified but non-complete family of gain-bounded rational operators. Remark 1 makes this representational restriction explicit. The proof is given in Appendix B.2.

Remark 1 (Expressivity of the gain-bounded family) *For fixed* γ, *the construction in Theorem 1 has* $n+1$ *scalar parameters, whereas a general real proper rational transfer function of order* n *has up to* $2n+1$ *independent coefficients. Hence, the Schur-recursive family is a lower-dimensional subset of the gain-bounded rational operators. For example, at order one,*

$$\frac{H(z)}{\gamma} = \frac{k_0 + k_1 q}{1 + k_0 k_1 q}, \qquad q = z^{-1},$$

so the numerator and denominator coefficients are algebraically coupled. The construction should therefore be viewed as a sufficient certified family rather than a complete parameterization of all rational operators satisfying $\|H\|_{\mathcal{H}_\infty} < \gamma$. *The restriction is particularly visible for strictly proper filters:* $h_0 = 0$ *forces* $k_0 = 0$, *so* $S_0(q) = q S_1(q)$ *and the denominator has degree at most* $n-1$. *Thus, each imposed leading delay consumes one stage of the finite Schur recursion.*

3.3 MIMO architecture and parallel implementation

For multivariate sequences, scalar CERNEL filters are assembled into the transfer matrix

$$\mathbf{H}(z) = \big[H_{ij}(z)\big]_{i=1,\dots,d_{\text{out}},\ j=1,\dots,d_{\text{in}}}, \qquad y^{(i)} = \sum_{j=1}^{d_{\text{in}}} H_{ij} u^{(j)}.$$

All $d_{\text{out}} \times d_{\text{in}}$ filters can be evaluated in parallel. If $\|H_{ij}\|_{\mathcal{H}_\infty} < \gamma_{ij}$ and $\Gamma = [\gamma_{ij}]$, then

$$\|\mathbf{H}\|_{\mathcal{H}_\infty} < \|\Gamma\|_2.$$

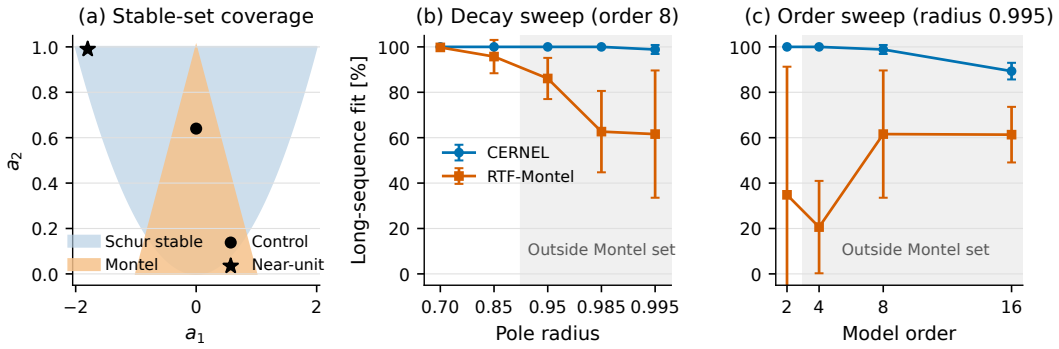

Figure 4: **Effect of the Montel constraint.** (a) The Montel set is a strict subset of the second-order Schur-stable region. (b) Long-sequence fit as the target poles approach the unit circle for order-eight systems. (c) Matched-order recovery for near-unit targets with $r_{\max} = 0.995$. Markers show means and sample standard deviations over three target systems.

A global gain bound γ can therefore be imposed by selecting the channelwise budgets such that $\|\Gamma\|_2 < \gamma$. See Appendix B.3 for more details.

Although the reflection-coefficient and Schur constructions are recursive in the filter order, their updates admit associative 2×2 polynomial-matrix representations. A parallel reduction therefore decreases their sequential depth from $\mathcal{O}(n)$ to $\mathcal{O}(\log n)$. Once the filters are constructed, full sequences can be evaluated through FFT-based convolution. An equivalent state-space-like realization also permits parallel affine scans with $\mathcal{O}(\log L)$ sequential depth over a sequence of length L. Both implementations are differentiable and algebraically equivalent to their sequential counterparts. Derivations and finite-horizon preservation of the gain bound are given in Appendix B.4.

The stability and gain guarantees above apply only to CERNEL. A certificate for the complete nonlinear block such as in Figure 3 requires the remaining components to be constrained accordingly, for example using nonlinear layers with prescribed incremental ℓ_2-gain (Wang & Manchester, 2023).

4 Experiments

We evaluate CERNEL along four dimensions: the expressivity of its stable polynomial parameterization, satisfaction of prescribed gain bounds, predictive performance across sequence modeling and nonlinear system identification, and robustness to observation perturbations. We also measure the computational benefit of the parallel implementation. Additional results and experimental protocols are provided in Appendix C and Appendix D, respectively.

4.1 Expressivity of the stable parameterization

We isolate the effect of the denominator parameterization by comparing CERNEL with the Montel-constrained RTF layer (RTF-Montel) of Parnichkun et al. (2024). Both models use matched orders, numerators, training data, optimizers, and training budgets; only their denominator coordinates differ. For a monic denominator $A(z) = 1 + \sum_{i=1}^{n} a_i z^{-i}$, RTF-Montel enforces $\sum_{i=1}^{n} |a_i| < 1$. This condition guarantees Schur stability but is not necessary and therefore excludes valid stable polynomials. CERNEL instead covers the complete class of real monic Schur-stable polynomials of the chosen degree.

Figure 4 summarizes the geometric and empirical consequences of this restriction. The Montel set covers only part of the second-order Schur-stable region, excluding systems with slowly decaying dynamics. Correspondingly, the performance gap between CERNEL and RTF-Montel widens as the target poles approach the unit circle and persists across the evaluated model orders.

We evaluate 60 stable target systems spanning multiple model orders and pole radii. Across this collection, CERNEL achieves an average long-sequence fit of $99.03\% \pm 2.89\%$, compared with $67.99\% \pm 33.02\%$ for RTF-Montel.

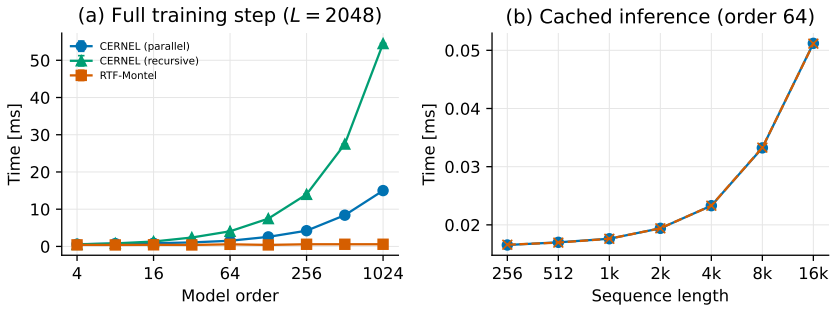

Figure 5: **Computational cost of CERNEL.** Left: forward-and-backward time for the explicit recursion, parallel CERNEL, and RTF-Montel at $L = 2048$. Right: cached FFT inference versus sequence length for order $n = 64$. Markers show medians and error bars show interquartile ranges over synchronized CUDA measurements.

When the target lies inside the Montel set, both parameterizations recover it to numerical precision; see Appendix C.1. Control experiments inside the Montel set and against complete stable SOS coordinates are provided in Appendices C.1 and C.3, respectively.

4.2 Computational efficiency

The explicit reflection-coefficient recursion requires $\mathcal{O}(n^2)$ work and $\mathcal{O}(n)$ sequential depth. By expressing its updates as an ordered product of 2×2 polynomial matrices, our parallel implementation reduces the depth to $\mathcal{O}(\log n)$ with $\mathcal{O}(n)$ work per frequency; see Appendix B.4.

As shown in Figure 5, the parallel formulation substantially reduces training time at large filter orders, achieving a $3.48\times$ speedup at $n = 512$, although parameter updates remain more expensive than for RTF-Montel. After training, the transfer kernel is cached and both models use the same FFT convolution, resulting in comparable inference cost and $\mathcal{O}(BL \log L)$ full-sequence complexity.

4.3 Validation and conservatism of the prescribed gain

We examine the representational cost of the prescribed-gain construction using the stable second-order teacher

$$H^\star(z) = \frac{0.0075 z^{-1}}{(1 - 0.95 z^{-1})(1 - 0.90 z^{-1})}, \qquad \|H^\star\|_{\mathcal{H}_\infty} = 1.5.$$

Although the teacher satisfies the prescribed budget $\gamma = 2$, it does not belong to the matched-order finite-Schur family; see Appendix D.3.

Figure 6 shows that increasing the model order progressively reduces this representational gap. The test MSE decreases from 0.995 at matched order to 3.1×10^{-4} at order 128, while every learned filter satisfies the prescribed gain bound. Thus, for this example, overparameterization recovers much of the lost expressivity, but at the cost of additional parameters, which exposes the conservatism.

4.4 Long-range sequence modeling

We first evaluate CERNEL on the copying and continuous-delay tasks of Parnichkun et al. (2024). CERNEL and RTF-Montel use matched architectures, parameter counts, initialization, data, and training budgets, differing only in their denominator parameterizations.

The advantage is largest when the required memory approaches the available filter order. On copying, CERNEL achieves 85.12% and 99.61% accuracy at orders 64 and 128, compared with 1.64% and 54.49% for RTF-Montel; at higher orders, both parameterizations solve the task. CERNEL also improves upon RTF-Montel at every evaluated order on continuous delay, although by smaller margins. These results indicate that the Montel restriction is most consequential in the memory-limited regime.

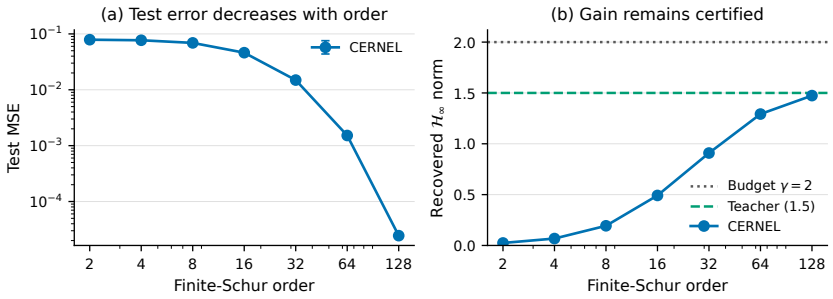

Figure 6: **Certification versus expressivity.** An order sweep for a second-order teacher satisfying $\|H^\star\|_{\mathcal{H}_\infty} = 1.5$ under the prescribed budget $\gamma = 2$. (a) Test MSE decreases with model order. (b) The numerically evaluated gain remains below the prescribed bound.

Table 1: **Synthetic long-memory tasks.** Published S4 and RTF results are from Parnichkun et al. (2024). Best results are **bold**, and second-best results are underlined.

	Copying accuracy (%) ↑				Continuous-delay RMSE ↓			
n	S4	RTF	Montel	CERNEL	S4	RTF	Montel	CERNEL
64	29.3	22.1	1.64	**85.12**	**0.410**	0.450	0.4477	0.4474
128	34.2	93.3	54.49	**99.61**	**0.390**	0.450	0.4467	0.4393
256	35.0	**100.0**	**100.0**	**100.0**	**0.330**	0.440	0.4470	0.4389
512	33.1	**100.0**	**100.0**	**100.0**	**0.220**	0.380	0.3864	0.3788
1024	33.2	**100.0**	**100.0**	**100.0**	0.029	**0.006**	0.00838	0.00815

We compare CERNEL with a matched RTF rerun on four LRA tasks (Tay et al., 2020; Parnichkun et al., 2024).[2] The models differ only in their denominator parameterization. CERNEL achieves comparable accuracy while guaranteeing denominator stability; comparisons with published sequence-model baselines are provided in Appendix C.2. Results are single-seed, so the small differences are not statistically meaningful.

4.5 Nonlinear system identification

We next evaluate whether the stable polynomial parameterization remains effective when rational operators are embedded within nonlinear dynamical models. Starting from the released dynoNet architectures (Forgione & Piga, 2021), we replace only the unconstrained denominator coefficients of each linear block with the reflection-coefficient parameterization of Proposition 1. Numerators, nonlinearities, model sizes, data splits, and training budgets are otherwise matched. Performance is measured through free-run simulation over three paired random seeds.

CERNEL achieves lower mean RMSE on four of the five benchmarks, with the largest reductions on EMPS and Parallel Wiener–Hammerstein. On Silverbox, the two parameterizations are effectively indistinguishable, indicating that the stability constraint does not impose a systematic accuracy penalty. The unconstrained dynoNet models happened to remain stable in these runs; the result is therefore not that unconstrained training must fail, but that CERNEL guarantees stability while retaining, and in several cases improving the predictive accuracy.

4.6 Robustness to observation perturbations

We finally examine whether prescribing the induced ℓ_2-gain of the temporal operator reduces the amplification of observation errors in a closed-loop policy. We augment the official low-dimensional Diffusion Policy with a residual temporal module at the bottleneck of its conditional U-Net (Chi et al., 2023; 2024; 2025). The module is implemented either as an unconstrained RTF or as gain-constrained CERNEL, with matched parameter counts. All other architectural and training choices

[2]We did not obtain converged matched results for Pathfinder or Path-X and therefore do not report a full-suite average.

Table 2: **Long Range Arena test accuracy (%).** Both models use matched architectures and training settings.

Method	ListOps	Text	Retrieval	Image	Certified
RTF	60.03	89.38	**92.02**	90.48	No
CERNEL	**60.33**	**89.40**	91.80	**90.64**	Yes

Table 3: **Nonlinear system identification.** Free-run RMSE, reported as mean ± standard deviation over three seeds. Lower is better.

Benchmark	dynoNet	CERNEL	Change
Wiener–Hammerstein	1.253 ± 0.166	$\mathbf{1.226 \pm 0.254}$	-2.1%
Bouc–Wen	$4.99\times10^{-5} \pm 3.93\times10^{-6}$	$\mathbf{4.70\times10^{-5} \pm 4.17\times10^{-6}}$	-5.8%
EMPS	18.02 ± 4.89	$\mathbf{11.43 \pm 5.91}$	-36.6%
Silverbox	$\mathbf{0.009549 \pm 0.000033}$	0.009560 ± 0.000012	$+0.1\%$
Parallel Wiener–Hammerstein	10.23 ± 1.85	$\mathbf{6.03 \pm 0.80}$	-41.1%

remain unchanged, and the residual branch is initialized at zero to preserve the original policy at initialization.

Perturbations are applied only during inference after observation normalization. Clean and corrupted evaluations use identical environment and diffusion-sampling seeds. We consider independent Gaussian noise. Exact corruption definitions and training details are provided in Appendix D.

Table 4: **PushT robustness.** Mean score over 50 evaluation environments. Corrupted entries report *score (relative degradation from clean, %)*, where $\mathcal{L}_c = 100(1 - S_c/S_{\text{clean}})$. **Bold** denotes the highest absolute score within each evaluation condition; relative degradation from the corresponding clean score is shown in parentheses.

Method	Clean	G. 0.05	G. 0.10	G. 0.15
Diffusion Policy	**0.9022**	0.8151 (9.7)	0.6284 (30.4)	0.4777 (47.1)
Diffusion Policy + RTF	0.8266	**0.8215** (0.6)	0.6518 (21.1)	0.4299 (48.0)
Diffusion Policy + CERNEL	0.8084	0.8080 (0.05)	**0.7236** (10.5)	**0.5314** (34.3)

Despite lower clean performance, CERNEL performs best under stronger Gaussian corruption, providing preliminary component-level evidence of reduced noise sensitivity.

5 Conclusion, Limitations and Future work

We presented Certified Rational Neural Layers (CERNEL), adapting classical reflection coefficients to parameterize all real monic Schur-stable denominators and a finite-Schur recursion to prescribe induced ℓ_2 gain. Both support parallel construction and FFT-based evaluation. Experiments show improved recovery of slowly decaying dynamics and competitive predictive performance. PushT provides preliminary evidence under Gaussian noise, not a system-level robustness guarantee.

The prescribed-gain construction is sufficient rather than complete, and its certificate applies directly to the rational operator rather than automatically to the complete nonlinear architecture. Future work will consider less conservative gain-bounded parameterizations, improved numerical conditioning at high filter orders, broader empirical evaluation, and extensions beyond fixed linear time-invariant operators.

AI USE STATEMENT

We used generative AI tools to provide feedback on experimental design, assist with code development and debugging, check and refine mathematical derivations, support the interpretation and presentation of experimental results, and improve the clarity, structure, and concision of the manuscript. These tools were also used for literature search. All AI-generated outputs were treated as preliminary suggestions and subsequently reviewed and refined by the authors. The authors verified all mathematical claims, checked citations against the original sources, reviewed and tested AI-assisted code, validated the reported results, and made all final scientific and editorial decisions. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

All authors have read and agree to follow the ICLR Code of Ethics. The experiments use synthetic systems and established benchmark datasets; we did not conduct studies involving human participants or collect personal or sensitive data. CERNEL is intended to improve the reliability of learned sequence models through stability and input–output gain certificates. These certificates apply to the rational layer itself and should not be interpreted as guarantees of stability, safety, or adversarial robustness for an entire nonlinear architecture or deployed control system. Similarly, the visuomotor experiments demonstrate reduced sensitivity to the tested observation noise, not general robustness in real-world deployment. Any safety-critical application would therefore require additional system-level analysis, validation, and appropriate domain-specific safeguards.

REPRODUCIBILITY STATEMENT

The main text provides the definitions, algorithms, assumptions, and guarantees underlying CERNEL. The appendix contains the corresponding proofs and derivations, details of the parallel implementations, complexity analysis, controlled validation experiments, and additional results. It also documents the datasets and synthetic-system generation procedures, model architectures, preprocessing, initialization, hyperparameters, training protocols, evaluation metrics, random seeds, and computational resources used in the experiments. The supplementary material includes an anonymous implementation of both CERNEL constructions together with the training configurations and scripts required to reproduce the reported tables and figures.

REFERENCES

Daniel Alpay and Victor Vinnikov. *Characteristic functions, scattering functions and transfer functions: the Moshe Livsic memorial volume*. Springer Science & Business Media, 2010.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Guy E. Blelloch. Prefix sums and their applications. 1990. URL `https://api.semanticscholar.org/CorpusID:60459178`.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Loris Di Natale, Muhammad Zakwan, Philipp Heer, Giancarlo Ferrari-Trecate, and Colin Neil Jones. Simba: System identification methods leveraging backpropagation. *IEEE Transactions on Control Systems Technology*, 33(2):418–433, 2024.

N Benjamin Erichson, Omri Azencot, Alejandro Queiruga, Liam Hodgkinson, and Michael W Mahoney. Lipschitz recurrent neural networks. *arXiv preprint arXiv:2006.12070*, 2020.

Marco Forgione and Dario Piga. dynonet: A neural network architecture for learning dynamical systems. *International Journal of Adaptive Control and Signal Processing*, 35(4):612–626, 2021.

Keith Glover and John C Doyle. A state space approach to H_∞ optimal control. In *Three decades of mathematical system theory: a collection of surveys at the occasion of the 50th birthday of Jan C. Willems*, pp. 179–218. Springer, 2005.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35971–35983. Curran Associates, Inc., 2022. doi: 10.52202/068431-2607. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e9a32fade47b906de908431991440f7c-Paper-Conference.pdf.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22982–22994. Curran Associates, Inc., 2022. doi: 10.52202/068431-1670. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9156b0f6dfa9bbd18c79cc459ef5d61c-Paper-Conference.pdf.

Vaibhav Gupta, Muhammad Zakwan, Giancarlo Ferrari Trecate, and Alireza Karimi. Ensuring stability in bilinear structured state-space models via iqcs: A free parameterisation approach. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 7481–7486. IEEE, 2025.

Ben Hayes, Jordie Shier, György Fazekas, Andrew McPherson, and Charalampos Saitis. A review of differentiable digital signal processing for music and speech synthesis. *Frontiers in Signal Processing*, 3:1284100, 2024.

Neelay Junnarkar, Murat Arcak, and Peter Seiler. Synthesizing neural network controllers with closed-loop dissipativity guarantees. *Automatica*, 193:113185, 2026.

Boris Kuznetsov, Julian D Parker, and Fabián Esqueda. Differentiable iir filters for machine learning applications. In *Proc. Int. Conf. Digital Audio Effects (eDAFx-20)*, pp. 297–303, 2020.

Aakash Lahoti, Kevin Y Li, Berlin Chen, Caitlin Wang, Aviv Bick, J Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. *arXiv preprint arXiv:2603.15569*, 2026.

Eric Martin and Chris Cundy. Parallelizing linear recurrent neural nets over sequence length. *arXiv preprint arXiv:1709.04057*, 2017.

Leonardo Massai, Muhammad Zakwan, and Giancarlo Ferrari-Trecate. L2ru: a structured state space model with prescribed l2-bound. *arXiv preprint arXiv:2503.23818*, 2025.

Shahan Nercessian. Neural parametric equalizer matching using differentiable biquads. In *Proc. Int. Conf. Digital Audio Effects (eDAFx-20)*, pp. 265–272, 2020.

Shahan Nercessian, Andy Sarroff, and Kurt James Werner. Lightweight and interpretable neural modeling of an audio distortion effect using hyperconditioned differentiable biquads. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 890–894. IEEE, 2021.

Sahel Vahedi Noori, Bin Hu, Geir Dullerud, and Peter Seiler. Stability and performance analysis of discrete-time relu recurrent neural networks. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 8626–8632. IEEE, 2024.

U Nurges and E Rustern. New stability conditions for discrete polynomials. In *Proceedings of the 2005, American Control Conference, 2005.*, pp. 4149–4153. IEEE, 2005.

Yuji Okamoto and Ryosuke Kojima. Learning deep dissipative dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19749–19757, 2025.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International conference on machine learning*, pp. 26670–26698. PMLR, 2023.

Rom N Parnichkun, Stefano Massaroli, Alessandro Moro, Jimmy TH Smith, Ramin Hasani, Mathias Lechner, Qi An, Christopher Ré, Hajime Asama, Stefano Ermon, et al. State-free inference of state-space models: The transfer function approach. *arXiv preprint arXiv:2405.06147*, 2024.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. Pmlr, 2013.

Patricia Pauli, Dennis Gramlich, and Frank Allgöwer. Lipschitz constant estimation for 1d convolutional neural networks. In *Learning for dynamics and control conference*, pp. 1321–1332. PMLR, 2023.

Dario Piga, Marco Forgione, and Manas Mejari. Deep learning with transfer functions: new applications in system identification. *IFAC-PapersOnLine*, 54(7):415–420, 2021.

Achuth Rao and Prasanta Kumar Ghosh. Sfnet: A computationally efficient source filter model based neural speech synthesis. *IEEE Signal Processing Letters*, 27:1170–1174, 2020.

Max Revay, Ruigang Wang, and Ian R Manchester. Recurrent equilibrium networks: Flexible dynamic models with guaranteed stability and robustness. *IEEE Transactions on Automatic Control*, 69(5):2855–2870, 2023.

Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

Krishna Subramani, Jean-Marc Valin, Umut Isik, Paris Smaragdis, and Arvindh Krishnaswamy. End-to-end lpcnet: A neural vocoder with fully-differentiable lpc estimation. *arXiv preprint arXiv:2202.11301*, 2022.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

Henk van Waarde and Rodolphe Sepulchre. Training lipschitz continuous operators using reproducing kernels. In *Learning for Dynamics and Control Conference*, pp. 221–233. PMLR, 2022.

Ruigang Wang and Ian Manchester. Direct parameterization of lipschitz-bounded deep networks. In *International Conference on Machine Learning*, pp. 36093–36110. PMLR, 2023.

Jan C Willems. Dissipative dynamical systems part i: General theory. *Archive for rational mechanics and analysis*, 45(5):321–351, 1972.

Muhammad Zakwan and Giancarlo Ferrari-Trecate. Neural distributed controllers with port-hamiltonian structures. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 8633–8638. IEEE, 2024.

Muhammad Zakwan and Giancarlo Ferrari-Trecate. Neural port-hamiltonian models for nonlinear distributed control: An unconstrained parametrization approach. *IEEE Transactions on Automatic Control*, 2026.

Muhammad Zakwan, Liang Xu, and Giancarlo Ferrari-Trecate. Robust classification using contractive hamiltonian neural odes. *IEEE Control Systems Letters*, 7:145–150, 2022.

Muhammad Zakwan, Vaibhav Gupta, Alireza Karimi, Efe C Balta, and Giancarlo Ferrari-Trecate. Controller design for structured state-space models via contraction theory. *arXiv preprint arXiv:2604.07069*, 2026a.

Muhammad Zakwan, Leonardo Massai, Efe C Balta, and Giancarlo Ferrari-Trecate. Free parametrization of l_2-bounded structured state-space controllers for nonlinear control with stability guarantees. *arXiv preprint arXiv:2606.11049*, 2026b.

George Zames. On the input-output stability of time-varying nonlinear feedback systems part one: Conditions derived using concepts of loop gain, conicity, and positivity. *IEEE transactions on automatic control*, 11(2):228–238, 1966.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

A Related Work

A.1 Rational neural sequence models

Rational transfer functions (RTFs) provide finite-dimensional linear operators with infinite impulse responses and admit both recurrent and convolutional implementations. In system identification, dynoNet introduced differentiable rational transfer-function layers as building blocks of nonlinear dynamical networks, combining linear dynamical operators with static neural nonlinearities (Forgione & Piga, 2021). Related work incorporated such layers into prediction-error and likelihood-based identification methods (Piga et al., 2021). Differentiable infinite-impulse-response (IIR) filters have also appeared in audio and differentiable signal-processing architectures, typically using low-order filter structures or application-specific stability constraints (Kuznetsov et al., 2020; Hayes et al., 2024).

Differentiable IIR filters have also been studied extensively in differentiable digital signal processing, particularly for audio and speech synthesis (Hayes et al., 2024). Existing approaches include direct recursive optimization (Kuznetsov et al., 2020), frequency-sampled IIR filters (Nercessian, 2020), and cascaded biquad parameterizations with stability-preserving activations (Nercessian et al., 2021). We provide a controlled comparison in Appendix C.3. In neural linear-prediction models, stable all-pole filters have additionally been parameterized through reflection coefficients followed by a differentiable Levinson recursion (Rao & Ghosh, 2020; Subramani et al., 2022).

CERNEL builds on these classical ideas but targets a different problem: general-purpose rational sequence modeling. Rather than introducing stability-preserving reflection coefficients themselves, CERNEL makes their role as a complete coordinate system for real monic Schur-stable polynomial explicit, combines them with freely trainable numerators, and develops a complementary finite-Schur construction for prescribed induced ℓ_2 gain. The resulting filters further admit MIMO gain bounds and an associative parallel construction over the filter order.

More recently, the Rational Transfer Function (RTF) architecture brought the transfer-function representation to long-range sequence modeling (Parnichkun et al., 2024). RTF evaluates the frequency response directly and applies the resulting kernel through FFT convolution, avoiding explicit propagation of the realization state. It also considers the Montel coefficient condition as a sufficient mechanism for enforcing stability. This condition is computationally simple but represents only a subset of the Schur-stable polynomials. CERNEL retains the state-free transfer-function implementation while replacing this sufficient constraint with reflection-coefficient coordinates that parameterize every real monic Schur-stable polynomial of the selected degree. CERNEL additionally introduces a separate Schur-recursive construction for prescribing the induced ℓ_2 gain. Unlike unconstrained dynoNet and the Montel-constrained RTF formulation, CERNEL guarantees the corresponding property throughout optimization. The complete stable construction is, however, more expensive during parameter updates than the Montel normalization used by RTF.

A.2 Stable state-space and recurrent models

Modern structured state-space models (SSMs) obtain long-range memory through carefully designed state matrices, spectral parameterizations, and initializations (Gu et al., 2021; Gupta et al., 2022; Gu et al., 2022; Smith et al., 2022; Orvieto et al., 2023). Stable parameterizations have also been developed specifically for learning dynamical systems. SIMBa uses LMI-based free parameterizations of Schur state matrices to guarantee stability throughout gradient-based system identification (Di Natale et al., 2024). Related work extends stability-by-design to bilinear structured state-space models using integral quadratic constraints (Gupta et al., 2025), and studies controllability, observability, and contraction-based output-feedback synthesis for SSM surrogate models (Zakwan et al., 2026a).

Most closely related to CERNEL, L_2-recurrent unit (L2RU) constructs structured state-space layers with a prescribed induced ℓ_2-gain and uses the resulting certificate to establish closed-loop stability through small-gain arguments (Zakwan et al., 2026b; Massai et al., 2025). Its parameterization is expressed through a state-space realization and is designed primarily for certified nonlinear control. CERNEL instead operates directly in transfer-function coordinates. It provides a bijective parameterization of all real monic Schur-stable polynomials of a prescribed degree and a complementary finite-Schur construction for gain-bounded rational filters. This transfer-function representation enables direct frequency-domain construction and FFT-based application in general sequence-modeling architectures. L2RU provides a complete prescribed-gain parameterization for square state-space cores, but its complete construction requires dense system matrices and cubic-cost spectral computations; its computationally efficient structured construction is instead sufficient rather than complete. CERNEL targets scalar rational filters evaluated directly in the frequency domain, providing efficient FFT-based application but sacrificing fixed-order expressivity in its gain-constrained form. These methods therefore make different trade-offs between completeness, structure, and computational cost.

A.3 Gain-bounded and Lipschitz neural networks

Input–output gain bounds are classical tools for quantifying worst-case disturbance amplification and analyzing interconnected systems. The small-gain framework relates finite input–output gains to feedback stability (Zames, 1966), while dissipativity provides a general energy-based framework for stability and performance analysis (Willems, 1972). For linear systems, bounded-real and $\mathcal{H}_\infty$ methods characterize dynamical operators and controllers satisfying prescribed disturbance-attenuation bounds (Glover & Doyle, 2005).

These ideas have increasingly been incorporated into learned models, although the resulting approaches differ in the object being certified and in how the certificate is imposed. Lipschitz-continuous kernel operators provide tunable input–output robustness and support incremental small-gain analysis (van Waarde & Sepulchre, 2022). Lipschitz RNNs derive sufficient conditions for stable nonlinear recurrent dynamics (Erichson et al., 2020), whereas Recurrent Equilibrium Networks (RENs) directly parameterize contracting recurrent models satisfying incremental quadratic constraints, including gain and passivity conditions (Revay et al., 2023). For static feedforward networks, sandwich layers provide a direct parameterization satisfying an SDP-based ℓ_2-Lipschitz certificate (Wang & Manchester, 2023). Complementary approaches estimate Lipschitz constants through dissipativity and semidefinite optimization (Pauli et al., 2023), or enforce dissipativity by projecting learned dynamics onto a certified set (Okamoto & Kojima, 2025).

Related work has developed stability- and robustness-certified neural architectures for continuous-time dynamics and control. Contractive Hamiltonian neural ODEs guarantee contraction and non-exploding gradients by design, thereby limiting sensitivity to perturbations of the initial condition or input (Zakwan et al., 2022). Neural port-Hamiltonian controllers provide closed-loop stability together with finite or incremental ℓ_2-gain guarantees independently of the trainable controller parameters (Zakwan & Ferrari-Trecate, 2024; 2026). Most closely related to CERNEL, L2RU gives an unconstrained parameterization of structured state-space layers with a prescribed induced ℓ_2-gain bound (Massai et al., 2025).

CERNEL differs from these approaches in both representation and scope. Unlike static Lipschitz networks, it certifies a dynamical temporal operator. Unlike Lipschitz RNNs and RENs, it works directly with an explicit finite-dimensional rational transfer function rather than a nonlinear or implicit

state recurrence. Unlike contractive neural ODEs and port-Hamiltonian systems, it targets discrete-time sequence-modeling layers rather than continuous-time dynamics that are not easy to implement for a faster inference. Relative to L2RU, which constructs gain-bounded operators through a state-space parameterization, CERNEL uses a finite inverse-Schur recursion in transfer-function coordinates.

The resulting rational operator satisfies $\|H\|_{\mathcal{H}_\infty} < \gamma$ for every value of its unconstrained trainable parameters. The certificate therefore holds throughout optimization without solving an SDP, applying a projection, or estimating the gain after training. The construction is differentiable and compatible with standard gradient-based learning and FFT-based sequence evaluation. It is, however, a sufficient rather than complete parameterization of gain-bounded rational operators. Moreover, the certificate applies directly to the rational temporal component; certifying the complete nonlinear sequence block additionally requires compatible incremental-gain bounds on the surrounding nonlinear layers.

A possible alternative is to normalize a stable rational filter using the maximum response observed on an FFT grid. This retains greater fixed-order expressivity, but without an additional inter-grid or Lipschitz bound it provides only a grid-dependent approximate certificate. We therefore focus on an exact finite-Schur certificate and expose its resulting conservatism.

A.4 Parallel scans for sequence models

Parallel scans evaluate ordered compositions under an associative binary operation with logarithmic sequential depth (Blelloch, 1990). Linear affine recurrences can be represented by such associative compositions, enabling recurrent neural networks with linear sequential dependencies to be parallelized over the sequence dimension (Martin & Cundy, 2017). This principle is used in modern state-space architectures. S5 applies associative scans to a MIMO state-space recurrence (Smith et al., 2022), and LRU uses parallel linear recurrence evaluation to match the computational properties of structured SSMs (Orvieto et al., 2023). Input-dependent models such as Mamba develop hardware-aware scan implementations for selective state-space recurrences that can no longer be evaluated through a fixed convolutional kernel (Gu & Dao, 2023).

The scan used in CERNEL addresses a different recursive dimension. The reflection-coefficient and finite Schur constructions are recursive in the filter order n, even before the resulting operator is applied to a sequence. CERNEL expresses each recursion as an ordered product of constant-size polynomial matrices and uses associativity to reduce the construction depth from $\mathcal{O}(n)$ to $\mathcal{O}(\log n)$. The operation is vectorized independently over channels and frequency points. Once the transfer function has been constructed, a length-L sequence can be processed through FFT convolution, as in RTF, or through an equivalent affine state-space scan. Thus, CERNEL does not introduce a new parallel-scan algorithm; its contribution is the associative reformulation of the reflection-coefficient and Schur recursions needed to construct certified rational filters efficiently.

B Mathematical Proofs and Derivations

B.1 Proof of the complete stable-polynomial parameterization

The result follows from the classical reflection-coefficient characterization of real Schur polynomials (Nurges & Rustern, 2005). Unlike a direct root test or a sufficient coefficient inequality, this characterization recursively removes one degree of the polynomial and is necessary and sufficient for strict Schur stability. We recall the construction to make both the parameterization and its completeness explicit.

For each $i = 1, \dots, n$, define the intermediate monic polynomial

$$p_i(z) = z^i + a_1^{(i)} z^{i-1} + \cdots + a_i^{(i)}.$$

Its ith reflection coefficient is

$$k_i = -a_i^{(i)},$$

and the classical degree-reduction recursion is

$$a_j^{(i-1)} = \frac{a_j^{(i)} + k_i a_{i-j}^{(i)}}{1 - k_i^2}, \qquad j = 1, \dots, i-1.$$

Thus, stage i reduces a degree-i polynomial to degree $i-1$. A real monic polynomial is strictly Schur stable if and only if every reflection coefficient produced by this reduction satisfies $|k_i| < 1$.

The recursion in Proposition 1 is the inverse of this degree-reduction map. Indeed,

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}$$

and

$$a_{i-j}^{(i)} = a_{i-j}^{(i-1)} - k_i a_j^{(i-1)}.$$

Consequently,

$$\begin{aligned} a_j^{(i)} + k_i a_{i-j}^{(i)} &= a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} + k_i a_{i-j}^{(i-1)} - k_i^2 a_j^{(i-1)} \\ &= (1-k_i^2) a_j^{(i-1)}, \end{aligned}$$

which recovers the degree-reduction recursion.

Since $k_i = \tanh(\theta_i)$, every reflection coefficient satisfies $|k_i| < 1$. The reflection-coefficient stability criterion therefore implies that p_θ is strictly Schur stable for every $\Theta \in \mathbb{R}^n$.

Conversely, let p be an arbitrary real monic strictly Schur-stable polynomial of degree n. The degree-reduction recursion uniquely determines reflection coefficients $k_1, \dots, k_n \in (-1, 1)$. Since $\tanh$ is a bijection from $\mathbb{R}$ onto $(-1, 1)$, there exists a unique

$$\theta_i = \operatorname{artanh}(k_i)$$

for each i. Applying the forward recursion reconstructs p. Thus, the parameterization is bijective onto the class of real monic strictly Schur-stable polynomials of degree n.

B.2 Proof of the prescribed-gain parameterization

The proof adapts the classical disk-preserving argument underlying the Schur recursion (Alpay & Vinnikov, 2010). We additionally establish causality, rational degree, Schur stability under the change of variables $q = z^{-1}$, and the induced ℓ_2-gain certificate required for the sequence operator.

Let

$$\mathbb{D} := \{q \in \mathbb{C} : |q| < 1\}, \qquad \overline{\mathbb{D}} := \{q \in \mathbb{C} : |q| \le 1\}.$$

Since $k_i = \tanh(\theta_i)$, we have

$$|k_i| < 1, \qquad i = 0, \dots, n.$$

For $k \in (-1, 1)$, define the Möbius map

$$\phi_k(w) := \frac{k + w}{1 + kw}.$$

For every $w \in \mathbb{C}$, direct calculation gives

$$|1 + kw|^2 - |k + w|^2 = (1 - k^2)(1 - |w|^2), \tag{5}$$

and hence

$$1 - |\phi_k(w)|^2 = \frac{(1 - k^2)(1 - |w|^2)}{|1 + kw|^2}. \tag{6}$$

Therefore, whenever $|k| < 1$ and $|w| < 1$, the denominator $1 + kw$ is nonzero and

$$|\phi_k(w)| < 1.$$

Thus, ϕ_k maps the open unit disk into itself.

We now prove by backward induction that, for every $i = 0, \dots, n$,

1. S_i is rational and has no poles in $\overline{\mathbb{D}}$;
2. $$|S_i(q)| < 1, \qquad q \in \overline{\mathbb{D}}.$$

For $i = n$, the definition $S_n(q) = k_n$ gives

$$|S_n(q)| = |k_n| < 1,$$

so S_n is constant and has no poles in $\overline{\mathbb{D}}$.

Suppose that the claims hold for S_{i+1} and define

$$w_i(q) := qS_{i+1}(q).$$

For every $q \in \mathbb{D}$,

$$|w_i(q)| = |q|\,|S_{i+1}(q)| < 1.$$

Repeating the recursion from Theorem 1,

$$S_i(q) = \frac{k_i + qS_{i+1}(q)}{1 + k_i qS_{i+1}(q)} = \phi_{k_i}(w_i(q))\,.$$

The recursion can be written as

$$S_i(q) = \phi_{k_i}(w_i(q))\,.$$

Since $|k_i| < 1$ and $|w_i(q)| < 1$, the denominator satisfies

$$|k_i w_i(q)| < 1,$$

and therefore

$$1 + k_i w_i(q) \neq 0$$

on $\overline{\mathbb{D}}$. Hence S_i has no poles in $\overline{\mathbb{D}}$. Moreover, applying equation (6) yields

$$1 - |S_i(q)|^2 = \frac{(1 - k_i^2)\big(1 - |qS_{i+1}(q)|^2\big)}{|1 + k_i qS_{i+1}(q)|^2} > 0.$$

Consequently,

$$|S_i(q)| < 1, \qquad q \in \overline{\mathbb{D}}.$$

The induction is complete.

We next verify rationality and degree. Write

$$S_i(q) = \frac{N_i(q)}{D_i(q)}.$$

Starting from

$$N_n(q) = k_n, \qquad D_n(q) = 1,$$

the recursion gives

$$N_i(q) = k_i D_{i+1}(q) + qN_{i+1}(q), \tag{7}$$
$$D_i(q) = D_{i+1}(q) + k_i qN_{i+1}(q). \tag{8}$$

If N_{i+1} and D_{i+1} have degree at most $n - i - 1$, then N_i and D_i have degree at most $n - i$. It follows inductively that

$$\deg N_0 \leq n, \qquad \deg D_0 \leq n.$$

Any pole–zero cancellation can only reduce this degree. In addition,

$$D_i(0) = D_{i+1}(0) = \cdots = D_n(0) = 1,$$

so $S_0(q)$ defines a causal rational transfer function in the delay variable $q = z^{-1}$.

Because $D_0(q) \neq 0$ for every $|q| \leq 1$, all finite poles of S_0 in the q-plane satisfy

$$|q_p| > 1.$$

Under the change of variables $q = z^{-1}$, the corresponding poles of $H_\Theta(z) = \gamma S_0(z^{-1})$ satisfy

$$|z_p| = \frac{1}{|q_p|} < 1.$$

Thus H_Θ is a causal Schur-stable rational operator of degree at most n.

Finally, S_0 is continuous on the compact unit circle and satisfies

$$|S_0(q)| < 1, \qquad |q| = 1.$$

Therefore,

$$\max_{|q|=1} |S_0(q)| < 1.$$

For $q = e^{-\mathrm{i}\omega}$,

$$\|H_\Theta\|_{\mathcal{H}_\infty} = \sup_{\omega\in[0,2\pi]} \left|\gamma S_0(e^{-\mathrm{i}\omega})\right| \tag{9}$$

$$= \gamma \max_{|q|=1} |S_0(q)| \tag{10}$$

$$< \gamma. \tag{11}$$

Using the equality between the induced ℓ_2-gain and the $\mathcal{H}_\infty$-norm of a stable discrete-time LTI operator, we obtain, for every nonzero $u \in \ell_2$ under zero initial conditions,

$$\|H_\Theta u\|_{\ell_2} \le \|H_\Theta\|_{\mathcal{H}_\infty} \|u\|_{\ell_2} < \gamma \|u\|_{\ell_2}.$$

This proves the result.

B.3 Proof of the MIMO gain bound

Proposition 2 (MIMO gain from channelwise bounds) *Let*

$$\mathbf{H}(z) = \big[H_{ij}(z)\big] \in \mathbb{C}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$$

be a stable transfer matrix. Suppose that, for nonnegative constants γ_{ij},

$$\|H_{ij}\|_{\mathcal{H}_\infty} \le \gamma_{ij}, \qquad i = 1, \dots, d_{\mathrm{out}}, \quad j = 1, \dots, d_{\mathrm{in}}.$$

Define

$$\Gamma = [\gamma_{ij}] \in \mathbb{R}_{\ge 0}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}.$$

Then

$$\|\mathbf{H}\|_{\mathcal{H}_\infty} \le \|\Gamma\|_2.$$

Consequently, if $\|\Gamma\|_2 \le \gamma$, then

$$\|\mathbf{H}u\|_{\ell_2} \le \gamma \|u\|_{\ell_2}$$

for every $u \in \ell_2^{d_{\mathrm{in}}}$ under zero initial conditions.

Fix $\omega \in [0, 2\pi]$ and write $H_\omega := H(e^{\mathrm{i}\omega})$. For every $x \in \mathbb{C}^{d_{\mathrm{in}}}$, the channelwise bounds imply, componentwise,

$$|H_\omega x|_i \le \sum_{j=1}^{d_{\mathrm{in}}} |H_{ij}(e^{\mathrm{i}\omega})|\, |x_j| \le \sum_{j=1}^{d_{\mathrm{in}}} \gamma_{ij} |x_j| = (\Gamma |x|)_i.$$

Because the Euclidean norm is monotone under componentwise inequalities between nonnegative vectors, and because $\||x|\|_2 = \|x\|_2$, we obtain

$$\|H_\omega x\|_2 \le \|\Gamma |x|\|_2 \le \|\Gamma\|_2 \, \|x\|_2.$$

This is precisely the defining induced-norm inequality; equivalently, it uses submultiplicativity of the spectral norm. Taking the supremum over $x \neq 0$ gives

$$\bar\sigma\big(H(e^{\mathrm{i}\omega})\big) \le \|\Gamma\|_2.$$

Since the inequality holds for every frequency,

$$\|H\|_{\mathcal{H}_\infty} = \sup_{\omega\in[0,2\pi]} \bar\sigma\big(H(e^{\mathrm{i}\omega})\big) \le \|\Gamma\|_2.$$

Finally, stability of all scalar entries implies stability of the finite transfer matrix. Hence its induced ℓ_2 gain equals its $\mathcal{H}_\infty$ norm, and if $\|\Gamma\|_2 \le \gamma$, then

$$\|Hu\|_{\ell_2} \le \|H\|_{\mathcal{H}_\infty} \|u\|_{\ell_2} \le \gamma \|u\|_{\ell_2}.$$

The entrywise MIMO construction is a sufficient bound and may become conservative as the number of channels increases. The PushT implementation uses diagonal filters, for which the MIMO gain is simply the maximum channelwise gain.

B.4 PARALLEL CERNEL CONSTRUCTION

The reflection-coefficient and prescribed-gain constructions are recursive in the filter order. We show that both recursions can be expressed as ordered products of 2×2 matrices. Matrix multiplication is associative, allowing the products to be evaluated through a parallel scan rather than sequentially.

Proposition 3 (Parallelization of the CERNEL recursions) *At any fixed evaluation point $q \in \mathbb{C}$, both the stable-denominator recursion of Proposition 1 and the Schur recursion of Theorem 1 can be evaluated using an associative ordered matrix product. For a filter of order n, the resulting construction requires $\mathcal{O}(n)$ work and $\mathcal{O}(\log n)$ parallel depth per frequency.*

We treat the two constructions separately.

Stable-denominator recursion. Define the intermediate denominator polynomial

$$A_i(q) := 1 + \sum_{j=1}^{i} a_j^{(i)} q^j, \qquad A_0(q) := 1,$$

and its coefficient-reversed polynomial

$$A_i^{\mathrm{rev}}(q) := q^i A_i(q^{-1}).$$

The reflection-coefficient update

$$a_i^{(i)} = -k_i, \qquad a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}$$

is equivalent to

$$A_i(q) = A_{i-1}(q) - k_i q A_{i-1}^{\mathrm{rev}}(q).$$

Reversing this identity gives

$$A_i^{\mathrm{rev}}(q) = -k_i A_{i-1}(q) + q A_{i-1}^{\mathrm{rev}}(q).$$

Therefore,

$$\begin{bmatrix} A_i(q) \\ A_i^{\mathrm{rev}}(q) \end{bmatrix} = T_i(q) \begin{bmatrix} A_{i-1}(q) \\ A_{i-1}^{\mathrm{rev}}(q) \end{bmatrix}, \qquad T_i(q) := \begin{bmatrix} 1 & -k_i q \\ -k_i & q \end{bmatrix}.$$

Since

$$A_0(q) = A_0^{\mathrm{rev}}(q) = 1,$$

the final denominator satisfies

$$\begin{bmatrix} A_n(q) \\ A_n^{\mathrm{rev}}(q) \end{bmatrix} = T_n(q) T_{n-1}(q) \cdots T_1(q) \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Thus, the sequential reflection recursion is exactly an ordered matrix product.

Prescribed-gain Schur recursion. Write each intermediate Schur function as

$$S_i(q) = \frac{N_i(q)}{D_i(q)}.$$

Starting from

$$N_n(q) = k_n, \qquad D_n(q) = 1,$$

the recursion

$$S_i(q) = \frac{k_i + q S_{i+1}(q)}{1 + k_i q S_{i+1}(q)}$$

gives

$$N_i(q) = q N_{i+1}(q) + k_i D_{i+1}(q),$$

and

$$D_i(q) = k_i q N_{i+1}(q) + D_{i+1}(q).$$

Hence,

$$\begin{bmatrix} N_i(q) \\ D_i(q) \end{bmatrix} = G_i(q) \begin{bmatrix} N_{i+1}(q) \\ D_{i+1}(q) \end{bmatrix}, \qquad G_i(q) := \begin{bmatrix} q & k_i \\ k_i q & 1 \end{bmatrix}.$$

It follows that

$$\begin{bmatrix} N_0(q) \\ D_0(q) \end{bmatrix} = G_0(q) G_1(q) \cdots G_{n-1}(q) \begin{bmatrix} k_n \\ 1 \end{bmatrix},$$

and therefore

$$H_\Theta(z) = \gamma \frac{N_0(z^{-1})}{D_0(z^{-1})}.$$

Parallel evaluation. Both constructions reduce to an ordered product of n constant-size matrices at every frequency. Because matrix multiplication is associative,

$$(M_3M_2)M_1 = M_3(M_2M_1),$$

the product can be evaluated by a balanced binary-tree reduction, or equivalently by an associative parallel scan. Such a tree contains $\mathcal{O}(n)$ matrix multiplications but has only $\mathcal{O}(\log n)$ sequential depth.

When the transfer function is evaluated at F frequency points, the scans are performed independently and simultaneously across frequencies and channels. The resulting total work is $\mathcal{O}(nF)$, while the depth associated with the filter order remains $\mathcal{O}(\log n)$. Constructing the time-domain kernel through an inverse FFT adds $\mathcal{O}(F \log F)$ work.

The matrix products reproduce exactly the corresponding sequential recursions. Consequently, in exact arithmetic, parallelization changes neither the resulting transfer function nor its stability and induced ℓ_2-gain certificates.

Finite-horizon implementation. Let $\mathcal{T}_H$ denote the causal convolution operator induced by the gain-constrained transfer function H, and let E_L and P_L denote zero extension and truncation to the first L samples. For a length-L input, the implemented operator is

$$T_L = P_L \mathcal{T}_H E_L.$$

Because $\|P_L\|_2 = \|E_L\|_2 = 1$,

$$\|T_L\|_2 \leq \|\mathcal{T}_H\|_{\ell_2 \to \ell_2} = \|H\|_{\mathcal{H}_\infty} < \gamma.$$

In the prescribed-gain experiments, the first L coefficients of the causal impulse response are applied using $2L$-point zero-padded FFT convolution, which evaluates linear rather than circular convolution. In PushT, the same zero-state IIR operator is evaluated through a parallel affine scan. Thus, neither gain-certified experiment relies on an IFFT approximation obtained by sampling H on a frequency grid.

C Additional Experimental Results

C.1 Control target inside the Montel set

To separate representational limitations from possible optimization differences, we consider an order-eight target whose denominator lies close to, but inside, the Montel boundary:

$$\sum_{i=1}^{n} |a_i| = 0.986 < 1, \qquad r_{\max} = 0.85.$$

As shown in Figure 7, both models recover the unseen trajectory and pole locations to numerical precision. CERNEL and RTF-Montel obtain test fits of 99.999995% and 99.999997%, respectively, with pole RMSE below 10^{-7} for both models. Therefore, the degradation observed for slowly decaying systems arises when the Montel constraint excludes the target denominator, rather than from an intrinsic inability to optimize the RTF-Montel model.

C.2 Additional Long Range Arena results

For context, Table 5 compares CERNEL with the four-task results compiled by Parnichkun et al. (2024). These published values were not reproduced in our environment; the controlled comparison in our study is between RTF (rerun) and CERNEL. CERNEL provides competitive performance in LRA tasks. Therefore, stability by design does not compromise the generalization capabilities in long-range sequence modelling.

C.3 Comparison with a Complete Stable SOS Parameterization

We compare CERNEL with a complete cascade of stable second-order sections (SOS). Each quadratic section is parameterized as

$$A_s(q) = 1 + a_{1,s}q + a_{2,s}q^2, \qquad a_{2,s} = \tanh(u_s), \qquad a_{1,s} = (1 + a_{2,s})\tanh(v_s),$$

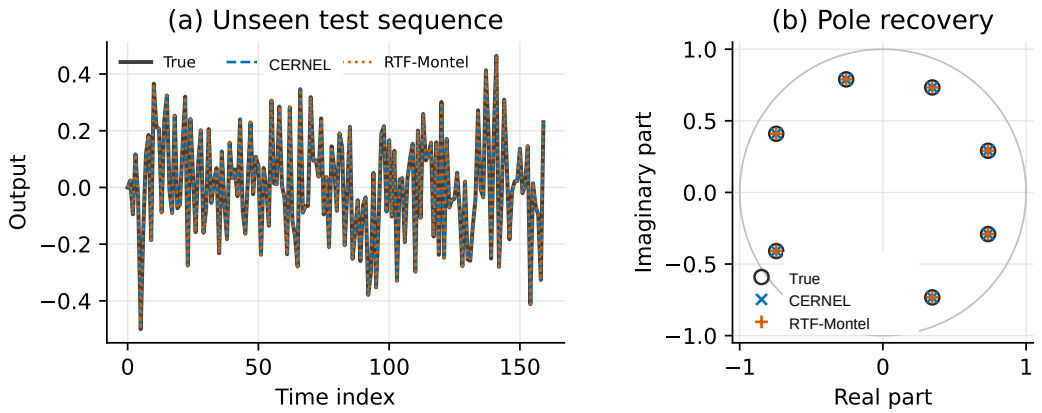

Figure 7: **Control experiment for a target inside the Montel set.** Left: predictions on an unseen input sequence. Right: true and recovered poles. Both parameterizations recover the matched-order system to numerical precision when the target denominator satisfies the Montel constraint.

Table 5: **Accuracy (%) on four LRA tasks.** Published values are reproduced from Parnichkun et al. (2024). RTF (rerun) and CERNEL are our matched single-seed runs. $\times$ denotes a reported failure.

Model	ListOps	Text	Retrieval	Image
Transformer	36.37	64.27	57.46	42.44
Luna-256	37.25	64.57	79.29	47.38
Transformer + Causal SPT	59.15	88.81	90.38	76.00
Mega $\mathcal{O}(\ell^2)$	**63.14**	**90.43**	91.25	90.44
H3	57.50	88.20	91.00	87.30
CCNN	43.60	84.08	$\times$	88.90
Liquid-S4	62.75	89.02	91.20	89.50
S5	62.15	89.31	91.40	88.00
S4	61.29	88.25	90.90	89.20
S4D	60.74	87.03	90.68	89.18
SpaceTime	56.40	87.80	91.45	86.27
RTF (rerun)	60.03	89.38	**92.02**	90.48
CERNEL	60.33	89.40	91.80	**90.64**

which spans the strictly Schur-stable second-order region. Thus, CERNEL and SOS both parameterize every real monic strictly Schur-stable denominator and use the same number of denominator parameters.

We evaluate 60 stable target systems spanning orders $n \in \{2, 4, 8, 16\}$, pole radii $r_{\max} \in \{0.70, 0.85, 0.95, 0.985, 0.995\}$, and real, complex, and mixed pole configurations. To avoid permutation-symmetric SOS initialization, the SOS sections are initialized at distinct points and the resulting denominator is converted to CERNEL coordinates; both models therefore start from the same transfer function. All other parameters, training data, optimization budgets, and three paired restarts are matched.

The median paired difference in test fit is 1.83×10^{-6} percentage points, showing essentially identical recovery on a typical target. CERNEL converges more frequently on the difficult systems, with restart- and target-level convergence rates of 83.3% and 86.7%, compared with 67.8% and 73.3% for SOS. These results establish no expressivity advantage over SOS, but provide limited evidence that the reflection-coefficient coordinates yield more reliable optimization under the tested protocol.

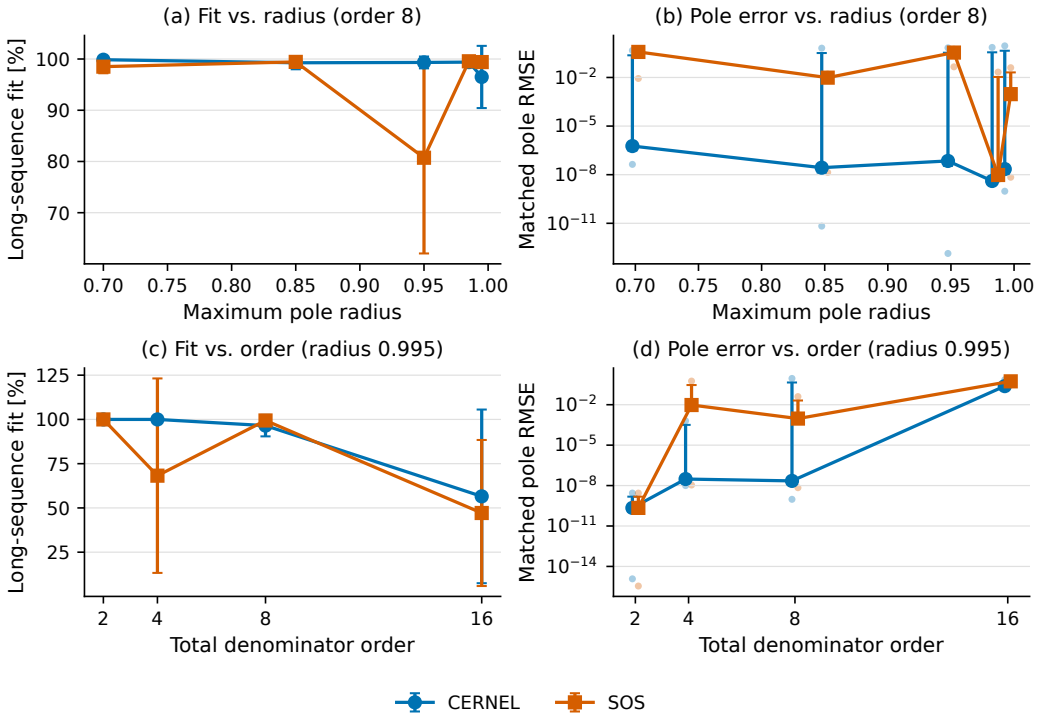

Figure 8: **CERNEL versus a complete stable SOS parameterization.** Long-sequence fit and matched-pole error as the target pole radius increases at order 8 (top) and as the model order increases at $r_{\max} = 0.995$ (bottom). Fit shows the mean and sample standard deviation; pole errors show individual targets, medians, and interquartile ranges.

D Experimental Details

Implementation. Experiments were run on an NVIDIA GeForce RTX 4090 using Python 3.10.12, PyTorch 2.7.1, and CUDA 11.8. Stable-system recovery experiments used double precision on a single CPU thread; all neural-network experiments used single precision on the GPU. The supplementary code provides the complete software environment and configuration files.

D.1 Stable-system recovery

We compare SISO CERNEL filters with the ℓ_1-montel parameterization from the official RTF implementation. Unless stated otherwise, trajectory fit is

$$100\left(1-\frac{\|\hat{\mathbf{y}}-\mathbf{y}\|_2}{\|\mathbf{y}\|_2}\right),$$

and pole error is computed after optimal one-to-one matching of the learned and target poles.

Second-order illustration. The Montel-feasible control has pole radius 0.8 and angle 0.5π. The near-unit target has radius 0.995 and angle 0.14π, and lies outside the Montel set. Both models are fitted to the first 1024 impulse-response samples using four Adam restarts, each run for 2500 iterations with learning rate 10^{-2}. Testing uses 32 independently generated Gaussian sequences of length 8192.

Order–decay sweep. We generate three systems for every combination of order $n \in \{2, 4, 8, 16\}$ and maximum pole radius $r_{\max} \in \{0.70, 0.85, 0.95, 0.985, 0.995\}$, for 60 systems in total. Complex poles occur in conjugate pairs, with radii in $[r_{\max} - 0.06, r_{\max}]$; independently sampled numerators are rescaled to give comparable output variance. Each model is fitted to 1024 impulse samples using

three restarts of 1000 Adam iterations at learning rate 10^{-2} followed by at most 150 strong-Wolfe L-BFGS iterations. Restarts are selected by training loss. Testing uses 16 Gaussian sequences of length 4096 (seed 37). Error bars show the mean and sample standard deviation over the three systems.

The Montel-feasible control uses an order-eight system with four conjugate pole pairs having radii between 0.79 and 0.85. It follows the same training and evaluation protocol as the sweep.

D.2 Computational benchmarks

Training time is measured with batch size $B = 8$, sequence length $L = 2048$, and orders $n \in \{4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. Cached inference is measured at order 64 for sequence lengths from 256 to 16384; kernel construction is excluded because the parameters are fixed at deployment. After warm-up, we use PyTorch's blocked-autorange timer with a minimum measurement duration of 0.2 seconds. We report the median and interquartile range.

D.3 Gain-constrained order sweep

The teacher used in Section 4.3 is

$$H^\star(z) = \frac{0.0075 z^{-1}}{(1 - 0.95 z^{-1})(1 - 0.90 z^{-1})}, \qquad \|H^\star\|_{\mathcal{H}_\infty} = 1.5.$$

All CERNEL models use the prescribed bound $\gamma = 2$ and the model orders reported in Figure 6. Training data, optimization, and evaluation are held fixed across orders; the corresponding configuration is included with the supplementary code.

To verify that the teacher is outside the matched-order family, let $q = z^{-1}$. For $n = 2$, the finite-Schur recursion gives

$$\frac{H_\Theta(z)}{\gamma} = \frac{k_0 + k_1(1 + k_0 k_2)q + k_2 q^2}{1 + k_1(k_0 + k_2)q + k_0 k_2 q^2}.$$

Because the teacher numerator is proportional to q, exact representation requires $k_0 = k_2 = 0$. The resulting denominator is one and therefore cannot represent poles at 0.95 and 0.90. Thus the teacher satisfies the gain budget but does not belong to the order-two finite-Schur family.

D.4 Synthetic memory tasks

We follow the copying and continuous-delay protocols of Parnichkun et al. (2024). CERNEL and RTF-Montel use identical data, numerators, surrounding architectures, optimization, and training budgets. For RTF-Montel, the denominator is

$$\mathbf{a} = \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_1 + |s| + 10^{-6}},$$

where s is learned. CERNEL uses the reflection-coefficient recursion from Section 3; no gain constraint or gain regularizer is applied. Matched models are initialized to the same effective transfer function by converting the initial Montel denominator to reflection coordinates.

Continuous delay. Each input is a band-limited white-noise sequence of length 4000, sampled at $\Delta t = 2.5 \times 10^{-4}$ with cutoff frequency 1000. The target is a 1000-sample delay. Training data are generated online, with $16{,}384$ sequences per epoch; evaluation uses 1024 fixed sequences. The linear model consists of a 1-to-4 input projection, one causal rational layer with a shared denominator and four independent numerators, and a 4-to-1 output projection. We evaluate orders $n \in \{64, 128, 256, 512, 1024\}$.

Models are trained for 20 epochs with AdamW, learning rate 10^{-3}, batch size 64, and zero weight decay. The objective is sample-wise MSE, and the checkpoint with the lowest evaluation MSE is reported. All runs use seed 1112. Published S4 and RTF results are taken from Parnichkun et al. (2024); the CERNEL and RTF-Montel results are controlled single-seed runs in our environment.

Table 6: Nonlinear system-identification protocols.

Data	Architecture	Split and transient	Optimization
WH	$G(1,1;8,8,1) \to F(1,1;10) \to G(1,1;8,8,0)$	Train: first 20,000 samples, skip 5000. Test: samples 100,000:188,000, skip 1000.	Adam, 4×10^4 full-batch steps, lr 2×10^{-4}; loss scaled by 10^3.
Bouc–Wen	$G(1,8;3,3,1) \to F(8,4;10) \to G(4,4;3,3,0) \to F(4,1;10)$, summed with $G(1,1;2,2,1)$	Released train/test multisines; input divided by 50, output by 6×10^{-4}; skip 300.	Adam, 10^4 full-batch steps, lr 2×10^{-3}.
EMPS	$G(1,10;2,2,1) \to F(10,5;20) \to G(5,1;2,2,0)$, followed by the fixed cumulative sum	`DATA_EMPS.mat` for training and `DATA_EMPS_PULSES.mat` for testing; no transient.	Adam, 5×10^4 full-batch steps, lr 10^{-4}; reflection coordinates use lr 10^{-3}.
Silverbox	$G(1,1;3,3,0) \to F_{\mathrm{res}}(1,1;20) \to G(1,1;3,3,0)$	Train: first 40,000 samples. Holdout: unused continuation, skip 50; input mean-centered.	Adam, 10^4 full-batch steps, lr 10^{-3}.
Parallel WH	$G(1,2;6,6,0) \to F(2,2;20) \to G(2,1;6,6,0)$	95/5 train/validation realizations, batch 16, skip 100. Test: levels 1–5 and Arrow, concatenated after the transient.	Adam for 1000 epochs at lr 10^{-3}.

D.5 Long Range Arena

We evaluate ListOps, Text, Image, and Retrieval using the official RTF codebase. CERNEL replaces only the denominator coordinates of each rational layer; the data pipeline, task-specific architecture, numerator and direct-feedthrough parameters, optimizer, schedule, pooling, and evaluation code remain unchanged. CERNEL uses $k_i = \tanh(\theta_i)$ without additional coefficient decay. Exact task-specific configurations are included in the supplementary code.

Both methods use seed 3333, and the checkpoint with the highest validation accuracy is retained. ListOps, Image, and Retrieval are evaluated on their official test splits. For Text, the official IMDb pipeline exposes the held-out split as `val/accuracy`, which we report. Each result is from one completed seed; the comparison therefore supports no claim of a statistically significant improvement over RTF.

D.6 Nonlinear system identification

We use paired seeds $\{0, 1, 2\}$, resetting all random-number generators before constructing each model. Paired models therefore share their initial transfer functions and static-network weights and have identical parameter counts. Test records are not used for model or hyperparameter selection. We report RMSE and FIT from free-run simulations after the transients listed below. Models are evaluated at the final repository-prescribed iterate.

Let $G(c_{\mathrm{in}}, c_{\mathrm{out}}; n_b, n_a, n_k)$ denote a linear dynamical block and $F(c_{\mathrm{in}}, c_{\mathrm{out}}; h)$ a one-hidden-layer tanh network.

The EMPS and Parallel-WH coordinate learning-rate multipliers were selected in preliminary screens and then fixed for all three seeds. All other parameters retain the learning rates of the released implementations.

D.7 PushT observation-noise experiment

Dataset and policy. We use the official low-dimensional PushT benchmark dataset from Diffusion Policy. We retain the demonstrations, dynamics, reward, initialization, termination, normalization, and conditional one-dimensional U-Net. Observations and actions have dimensions 20 and 2, respectively; the prediction horizon is 16, the observation horizon is 2, and eight actions are executed per policy query. We use at most 90 training episodes, validation ratio 0.02, and seed 42.

RTF and CERNEL add one residual temporal module after the two middle U-Net blocks,

$$\widetilde{h} = h + \alpha \mathcal{T}(h), \qquad h \in \mathbb{R}^{B \times 1024 \times 4},$$

where α is learned and initialized to zero. Both modules are diagonal order-16 rational filters with identical parameter counts (18,433 additional parameters, including α). The rational filters operate along the four-step temporal axis at the U-Net bottleneck. RTF uses unconstrained coefficients; CERNEL uses the finite-Schur construction with prescribed gain $\gamma = 1$. This certificate applies only to $\mathcal{T}$, not to the residual connection or the complete diffusion policy.

The gain certificate applies only to the CERNEL branch T. The complete residual insertion satisfies the generic bound

$$\|I + \alpha T\|_2 \leq 1 + |\alpha|\gamma,$$

and is not claimed to have gain γ. For the reported EMA checkpoint, $\alpha = 0.0599$. The surrounding U-Net and complete diffusion policy are also unconstrained.

Training and checkpointing. All models use the official diffusion loss and DDPM configuration. They are trained with AdamW (learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.95, 0.999)$, weight decay 10^{-6}), batch size 256, a cosine schedule with 500 warm-up steps, and exponential moving averages. Results use one training seed, so the reported episode-level variation does not measure variation across trained policies.

Paired evaluation. We evaluate 50 environments with seeds $e_i = 100000 + i$. Diffusion and corruption seeds are $200000 + e_i$ and $300000 + e_i$, respectively, and are shared across models and paired clean/corrupted rollouts. Gaussian noise is added in normalized observation coordinates,

$$\widetilde{x}_k = x_k + \sigma\epsilon_k, \qquad \epsilon_k \sim \mathcal{N}(0, I), \qquad \sigma \in \{0.05, 0.10, 0.15\}.$$

The validity mask is not corrupted.

For episode i, the PushT score is $S_i = \max_t r_{i,t}$, and success means $S_i \geq 0.95$. For corruption condition c, we report the mean score, success rate, and relative score loss

$$\mathcal{L}_c = 100 \left(1 - \frac{\overline{S}_c}{\overline{S}_{\text{clean}}}\right).$$

Smaller relative loss indicates less degradation under the tested noise.