Spectral Identifiability for World Models: Polynomial Projectors, Resolvent Stability, and a Krylov Bottleneck

Hung Mai[1,2], **Hai Nguyen**[2,3], **Luong Doan**[2], **Ngoc Vu**[2], **Khanh Nguyen**[2,3,5], **Nhung Duong**[2], **Naeem Ul Islam**[4*], **Tuan Do**[5*]

[1]National Economics University, Vietnam
[2]N2TP Technology, Vietnam
[3]Nanyang Technological University, Singapore
[4]Yuan Ze University, Taiwan
[5]Phenikaa University, Vietnam

Abstract

World models are central to model-based reinforcement learning and planning, but their latent dimensions, architectural spectra, and downstream behavior remain poorly understood. In this work, we develop a spectral theory of linear world models in three settings. First, for shared multi-horizon prediction in linear dynamical systems, we identify the finite-horizon block-Krylov operator $\mathcal{K}_H$ as the exact latent bottleneck and give a perturbation condition for recovering its empirical rank elbow. Second, for stable diagonal state-space models with fixed spectra, we prove a Rademacher complexity bound governed by the spectral memory energy $\sum_j L_j(T)^2$ under a product-norm budget, rather than the parameter count alone. Third, for jointly trained diagonal latent dynamics with square encoder/decoder maps, we show that disjoint factor spectra imply factor-aligned latent coordinates at zero multi-lag loss, with a Riesz-projector stability extension for approximate recovery and a local gradient-flow companion. To relate these structural results to downstream evaluation, separate supporting bounds specify additional conditions for compositional OOD prediction, rollout envelopes, and fixed-policy evaluation. Finally, generated-system experiments illustrate the predicted spectral behavior under the stated assumptions, while motivating nonlinear observations show that pixel prediction and control returns can rank world models differently without establishing the mechanism behind those rankings.

1 Introduction

World models are learned representations that predict future states given actions, central to model-based reinforcement learning, sequence prediction, and planning [Ha and Schmidhuber, 2018, Hafner et al., 2025b]. Despite empirical progress, formal understanding of why certain architectures, latent dimensions, or training regimes produce good world models remains limited. The standard approach scores world models largely by reconstruction loss, leaving open how latent dimension relates to intrinsic dynamics, how architecture biases shape learned transitions, and when compositional generalization succeeds.

In this paper, we study three orthogonal spectral questions in linear settings where each can be answered exactly. The first concerns representation: what information a shared latent state must preserve for multi-horizon prediction. The second concerns statistical complexity: how stable diagonal

*Correspondence: Tuan Do (tuan.dongoc@phenikaa-uni.edu.vn); Naeem Ul Islam (naeem@saturn.yzu.edu.tw).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

state-space model (SSM) spectra govern generalization. The third concerns identifiability: when jointly trained diagonal latent dynamics recover compositional factors. A downstream motivation, why pixel prediction, rollout quality, and downstream control rank world models differently, ties these questions to model-based reinforcement learning practice.

Throughout the paper we work with a linear dynamical system (LDS) of the form $s_{t+1} = As_t + Ba_t + \xi_t$, where $s_t \in \mathbb{R}^D$ is the state with covariance Σ_s, a_t is the observed action, $A \in \mathbb{R}^{D\times D}$ is the state transition matrix, and ξ_t is process noise. For the first question we form the finite-horizon block-Krylov operator $\mathcal{K}_H$ whose blocks are the powers $A^h\Sigma_s^{1/2}$ for $h = 1, \ldots, H$. Theorem 1 identifies $\mathcal{K}_H$ as the shared-latent predictive object for multi-horizon LDS prediction and gives a perturbation condition under which the empirical elbow in its singular spectrum recovers the true bottleneck rank. For the second question, with diagonal SSM spectrum $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ and stability $|\lambda_j| < 1$, Theorem 2 gives a Rademacher complexity bound governed by the memory energy $\sum_j L_j(T)^2$ where $L_j(T)^2 = (1 - |\lambda_j|^{2T})/(1 - |\lambda_j|^2)$. For the third question, with square encoder E, decoder C, and learned diagonal dynamics Λ, Theorem 3 establishes that joint encoder-decoder factor alignment follows from disjoint factor spectra at zero multi-lag loss, while Theorem 4 extends this to approximate factor-aligned coordinate recovery under approximate one-step similarity, near-identity encoder-decoder composition, and pole localization. Theorem 5 provides a local projected-gradient-flow companion near the spectral solution. To relate these structural results to downstream evaluation, we also give three standalone propositions that clarify the additional assumptions needed for compositional OOD prediction, observation-space rollouts, and fixed-policy evaluation.

The contribution lies in identifying which spectral object answers each of these world-model questions. The objectives considered here lead to specific statements about predictive representations. In particular, Theorem 1 characterizes shared-code regression, including feedback residualization and empirical spectral-threshold recovery, whereas classical subspace identification and balanced truncation address realization estimation and dynamical input-output reduction. Theorem 2 then shows how individual pole locations enter a capacity bound at a fixed product-norm budget. Finally, Theorems 3–4 describe which factor-coordinate blocks can be recovered, what ambiguity remains, and how approximation errors affect the guarantee. These results provide a common perspective on three distinct linear problems, while their interpretation for nonlinear architectures remains conditional, as discussed in Section 5.2.

2 Related Work

World models, architectures, and latent dynamics. World models learn predictive latent dynamics for planning and control. Ha and Schmidhuber [2018] introduced the VAE+RNN world model, and Hafner et al. [2025a] scaled this idea to diverse domains. Recent architectures increasingly use Transformer and SSM backbones, motivating questions about latent memory, transition spectra, and generalization. Dao and Gu [2024] unified SSMs and linear attention through structured state-space duality, while 8B-scale empirical studies [Shoeybi et al., 2019] showed that SSMs can match Transformers on many tasks while exhibiting distinctive long-context tradeoffs. Our work asks which spectral quantities these architectural choices expose in a controlled linear world-model setting.

Statistical and system-theoretic foundations. Our analysis connects statistical learning theory with linear-system theory. Suh and Cheng [2025] surveyed statistical learning theory for deep networks, including approximation, NTK and mean-field regimes, and training dynamics; Nakada and Imaizumi [2020] studied adaptive approximation with intrinsic dimensionality; and Kuznetsov and Mohri [2017] gave guarantees for time-series forecasting under β-mixing. Existing RNN and dynamical-system generalization bounds often scale through spectral norms, parameter counts, unrolling length, stability constants, or PAC-Bayes terms [Chen et al., 2020, Eringis et al., 2024]. In contrast, Theorem 2 isolates the fixed diagonal-SSM case and gives exact pole sensitivity through the memory energy $\sum_j L_j(T)^2$.

Representation bottlenecks and subspace identification. Learning low-dimensional latent dynamics from high-dimensional observations has been studied for LTI systems with provable sample complexity bounds, and Vafa et al. [2024] applied Myhill–Nerode theory to implicit world models in generative sequence models. Classical subspace identification estimates realizations or

Markov/Hankel parameters [Van Overschee and De Moor, 1996], while balanced truncation studies reduced-order input-output approximation through controllability and observability Gramians [Antoulas, 2005]. Recent finite-sample LDS analyses give guarantees for spectral filtering, gradient methods, Ho–Kalman recovery, and least-squares prediction [Hazan et al., 2017, Hardt et al., 2018, Simchowitz et al., 2018, Oymak and Ozay, 2019, Sarkar and Rakhlin, 2019]. Theorem 1 targets a complementary object: the best rank-r approximation of a single present-state encoding $z_t = Es_t$ that must serve all horizons after observed actions are separated. For non-normal systems this can differ from one-step truncation; for example, with

$$A = \begin{bmatrix} 0.95 & 1 \\ 0 & 0.95 \end{bmatrix}, \qquad \Sigma_s = I, \qquad H = 8,$$

the one-step singular values are approximately $(1.57, 0.57)$, while the shared-horizon operator has singular values approximately $(11.34, 1.18)$.

Compositional generalization and identifiability. Compositional generalization and disentanglement provide another route to understanding world-model representations. Wiedemer et al. [2023] proved sufficient conditions for compositional generalization under factored data-generating processes, while Montero et al. [2022] showed that disentangled representations often fail at combinatorial generalization. More broadly, Locatello et al. [2019] showed that unsupervised disentanglement is impossible without inductive bias or supervision, and nonlinear ICA restores identifiability using auxiliary variables or temporal structure [Hyvarinen and Morioka, 2016, Khemakhem et al., 2020]. We build on this lesson by making decoder-side factor alignment an explicit spectral condition. Theorem 3 studies jointly trained linear encoder/decoder maps with diagonal latent dynamics, zero multi-lag population loss, and disjoint factor spectra, while Theorem 4 extends the mechanism to approximate recovery through contour-resolvent stability.

3 Theoretical Framework

We first ask what information a shared latent code must preserve, then examine how fixed diagonal filters affect statistical complexity, and finally study when learned coordinates recover compositional factors. These questions lead to Theorems 1–4, with Theorem 5 providing a local optimization companion. We then turn to downstream evaluation through three separate consequences of Assumption 4. The assumptions are collected below, and each result is accompanied by a proof sketch, with full proofs in Appendix A.

3.1 Definitions

Definition 1 (World model)**.** *A world model is a tuple $(\mathcal{E}, \mathcal{D}, f_\theta)$, where $\mathcal{E} : \mathcal{S} \to \mathbb{R}^{d_z}$ is an encoder, $\mathcal{D} : \mathbb{R}^{d_z} \to \mathcal{S}$ is a decoder, and $f_\theta : \mathbb{R}^{d_z} \times \mathcal{A} \to \mathbb{R}^{d_z}$ is a latent transition. It predicts $\hat{s}_{t+1} = \mathcal{D}(f_\theta(\mathcal{E}(s_t), a_t))$.*

Definition 2 (Shared multi-horizon latent prediction)**.** *For horizon H, let $Y_{t,H} := (s_{t+1}^\top, \ldots, s_{t+H}^\top)^\top \in \mathbb{R}^{HD}$. A rank-$r$ shared-latent linear predictor encodes the present state once, $z_t = Es_t \in \mathbb{R}^r$, and predicts all horizons by $\hat{Y}_{t,H} = Uz_t + Va_{t:t+H-1}$, with $U \in \mathbb{R}^{HD \times r}$. This is stricter than fitting a separate rank-r predictor per horizon, because the same latent coordinates must serve all horizons.*

Definition 3 (Compositional dynamics)**.** *A system is K-compositional if $s = (s^{(1)}, \ldots, s^{(K)})$ and $T(s, a) = (T_1(s^{(1)}, a), \ldots, T_K(s^{(K)}, a))$, with orthogonal factor subspaces in the squared prediction loss.*

Definition 4 (Contractive world model)**.** *A world model is ρ-contractive if $\|f_\theta(z, a) - f_\theta(z', a)\|_2 \leq \rho\|z - z'\|_2$ for all z, z', a, with $\rho < 1$. Its one-step latent error is $\epsilon := \sup_{(s,a)} \|f_\theta(\mathcal{E}(s), a) - \mathcal{E}(T(s, a))\|_2$.*

Standing assumptions. Each result invokes the relevant subset of the following assumptions.

Assumption 1 ((S1) Shared-latent LDS setting)**.** *The state follows $s_{t+1} = As_t + Ba_t + \xi_t$, where $a_t \in \mathbb{R}^{d_a}$, s_t is zero-mean with covariance $\Sigma_s \succ 0$, and $W_t := a_{t:t+H-1}$ is exogenous and mean-independent of s_t, with $\mathbb{E}[s_t W_t^\top] = 0$. Actions are observed without latent compression. The stacked*

horizon noise $\eta_{t,H}$ is independent, zero-mean, finite-covariance, and orthogonal to both present state and future actions: $\mathbb{E}[\eta_{t,H} s_t^\top] = 0$ and $\mathbb{E}[\eta_{t,H} W_t^\top] = 0$. For empirical elbow recovery, $\widehat{\mathcal{K}}_H$ is estimated from independent sub-Gaussian trajectory windows.

For non-exogenous or feedback-generated actions, $[A\tilde{\Sigma}_s^{1/2}; \dots; A^H \tilde{\Sigma}_s^{1/2}]$ is generally not the shared-latent bottleneck, because future actions may be coupled with future process noise and $A, \dots, A^H$ need not describe the regression of future states on the present after partialling actions out. Let $\tilde{s}_t := s_t - \Pi_W s_t$ and $\tilde{Y}_{t,H} := Y_{t,H} - \Pi_W Y_{t,H}$, where Π_W is the L_2 projection onto W_t. With $\tilde{\Sigma}_s := \mathrm{Cov}(\tilde{s}_t)$, define $M_H := \mathrm{Cov}(\tilde{Y}_{t,H}, \tilde{s}_t)\tilde{\Sigma}_s^\dagger$ and $\mathcal{K}_{H,\mathrm{FWL}} := M_H \tilde{\Sigma}_s^{1/2}$. This is the correct bottleneck for general action sequences. Under exogenous actions and process-noise orthogonality, $M_H = [A; A^2; \dots; A^H]$ and $\mathcal{K}_{H,\mathrm{FWL}} = \mathcal{K}_H$.

Assumption 2 ((S2) Stable diagonal-SSM class)**.** *The SSM spectrum $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_m)$ is fixed with $|\lambda_j| < 1$. Input sequences satisfy $\sum_t \|x_t\|_2^2 \le R_X^2$, and the product norm satisfies $\sum_j c_j^2 \|b_j\|_2^2 \le A^2$. For risk bounds, the loss is bounded and Lipschitz and the trajectory is stationary β-mixing.*

Assumption 3 ((S3) Spectral factor setting)**.** *The autonomous state decomposes orthogonally as $\mathbb{C}^D = \bigoplus_{k=1}^K U_k$, the dynamics are block diagonal $A = \bigoplus_k A_k$, and A is diagonalizable over $\mathbb{C}$ with pairwise disjoint factor spectra. If q is the degree of the minimal polynomial of A, set $q_+ := \max\{q, 2\}$ and include lags $h = 0, \dots, q_+ - 1$, so that $h = 0, 1$ are always present. The learned transition is diagonal and $E, C \in \mathbb{C}^{D \times D}$ are square. In real-valued implementations, nonreal conjugate eigenpairs are represented by real invariant two-dimensional blocks. Controlled variants require additional input-separation or controllability assumptions.*

Assumption 4 ((S4) Derived-risk regularity)**.** *For compositional OOD, the predictor is factored with orthogonal squared loss and support-uniform factor errors. For rollouts, the world model is ρ-contractive with finite one-step latent error, Lipschitz decoder, and finite reconstruction error. For control, the policy is fixed, rewards have bounded model error, and the true value function is Lipschitz in a task metric.*

3.2 Shared-Latent Multi-Horizon Bottleneck

Theorem 1 (Shared-latent finite-horizon Krylov bottleneck and elbow recovery)**.** *Under Assumption 1, define*

$$\mathcal{K}_H := \begin{bmatrix} A\Sigma_s^{1/2} \\ A^2\Sigma_s^{1/2} \\ \vdots \\ A^H\Sigma_s^{1/2} \end{bmatrix} \in \mathbb{R}^{HD \times D}.$$

Let Γ_H be the covariance of the stacked process noise in $(s_{t+1}, \dots, s_{t+H})$. Then $\inf_{E,U,V} \mathbb{E}\|Y_{t,H} - UEs_t - Va_{t:t+H-1}\|_2^2 = \sum_{i>r} \sigma_i(\mathcal{K}_H)^2 + \mathrm{tr}(\Gamma_H)$. Thus the systematic multi-horizon error is zero iff $r \ge r_\star(H) := \mathrm{rank}(\mathcal{K}_H)$. If $\|\widehat{\mathcal{K}}_H - \mathcal{K}_H\|_{\mathrm{op}} \le \varepsilon$, then every singular value is recovered to accuracy ε. Hence, if $\sigma_r(\mathcal{K}_H) > \tau + \varepsilon$ and $\sigma_{r+1}(\mathcal{K}_H) < \tau - \varepsilon$, then $\#\{i : \sigma_i(\widehat{\mathcal{K}}_H) > \tau\} = r$. Under independent sub-Gaussian trajectory windows, $\varepsilon = O\Big(L_H \sqrt{(HD + D + Hd_a + \log(1/\delta))/n}\Big)$ with probability at least $1 - \delta$, where L_H absorbs sub-Gaussian scale, covariance conditioning, and finite-horizon action/noise energy.

Concretely, from windows $(s_i, a_{i:i+H-1}, Y_{i,H})_{i=1}^n$, estimate $(\widehat{\mathcal{A}}_H, \widehat{\mathcal{B}}_H) \in \arg\min_{M,N} \sum_i \|Y_{i,H} - Ms_i - Na_{i:i+H-1}\|_2^2$, set $\widehat{\Sigma}_s = n^{-1} \sum_i s_i s_i^\top$, and define $\widehat{\mathcal{K}}_H = \widehat{\mathcal{A}}_H \widehat{\Sigma}_s^{1/2}$. For feedback data, residualize s_i and $Y_{i,H}$ against $W_i := a_{i:i+H-1}$, estimate $M_H = \mathrm{Cov}(\tilde{Y}_{t,H}, \tilde{s}_t)\tilde{\Sigma}_s^\dagger$, and set $\widehat{\mathcal{K}}_{H,\mathrm{FWL}} = \widehat{M}_H \widehat{\tilde{\Sigma}}_s^{1/2}$. Then

$$\inf_{E,U,V} \mathbb{E}\|Y_{t,H} - UEs_t - VW_t\|_2^2 = \sum_{i>r} \sigma_i(\mathcal{K}_{H,\mathrm{FWL}})^2 + \mathrm{tr}(\Omega_H),$$

where $\Omega_H := \mathrm{Cov}(\tilde{Y}_{t,H}) - M_H \tilde{\Sigma}_s M_H^*$. Under exogeneity, $M_H = [A; A^2; \dots; A^H]$ and $\Omega_H = \Gamma_H$.

Proof sketch. Stacking futures gives $Y_{t,H} = \mathcal{A}_H s_t + \mathcal{B}_H a_{t:t+H-1} + \eta_{t,H}$. Since actions are observed without compression and mean-independent of s_t, optimizing over V removes the action term and leaves a reduced-rank regression in s_t. The shared-latent predictor UEs_t has systematic error equal to the best rank-r approximation error of $\mathcal{A}_H \Sigma_s^{1/2} = \mathcal{K}_H$, so Eckart–Young–Mirsky gives the tail singular-value formula. Weyl's inequality gives singular-value stability and hence elbow recovery. For feedback data, Frisch–Waugh–Lovell reduces the joint regression to the partialled regression of $\tilde{Y}_{t,H}$ on $\tilde{s}_t$, giving the same argument with $\mathcal{K}_{H,\mathrm{FWL}}$. The finite-sample statement follows from standard sub-Gaussian matrix-regression concentration; see [Vershynin, 2018, Ch. 4]. □

Interpreting the horizon dependence. Theorem 1 distinguishes the dimension needed for exact prediction from the quality of a compressed representation. Under Assumption 1, the first-horizon block and invertibility of $\Sigma_s^{1/2}$ give

$$\ker(\mathcal{K}_H) = \ker(A\Sigma_s^{1/2}), \qquad \operatorname{rank}(\mathcal{K}_H) = \operatorname{rank}(A), \qquad H \geq 1.$$

Thus increasing the horizon does not change the rank needed for zero systematic error in this setting. It does, however, change the singular spectrum, the optimal rank-r error, and the subspace that should be retained. This also clarifies the relationship to other prediction objectives. The case $H = 1$ recovers one-step prediction, while separate horizon-specific codes remove the requirement that the same coordinates serve every horizon. Conversely, an open-loop recursive linear model initialized once from an r-dimensional code is more constrained than our shared-code predictor. Its optimal risk is therefore at least $\sum_{i>r} \sigma_i(\mathcal{K}_H)^2 + \operatorname{tr}(\Gamma_H)$, but equality need not hold because its horizon decoders and action maps must arise from a common recurrence.

3.3 Architecture-Specific Generalization for Stable Diagonal SSMs

Theorem 2 (Stable diagonal-SSM memory-norm generalization)**.** *Under Assumption 2, let* $L_j(T)^2 := \sum_{t=0}^{T-1} |\lambda_j|^{2t} = (1 - |\lambda_j|^{2T})/(1 - |\lambda_j|^2)$. *For scalar outputs, consider* $f_{B,c}(X) = \sum_{j=1}^{m} c_j \sum_{t=0}^{T-1} \lambda_j^t b_j^\top x_{T-1-t}$. *For any sample* $X_1, \ldots, X_n$ *satisfying Assumption 2,*

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_{\Lambda,A}) \leq \frac{AR_X}{\sqrt{n}} \left(\sum_{j=1}^{m} L_j(T)^2 \right)^{1/2}.$$

For dependent data, partition a stationary β*-mixing trajectory into* 2μ *alternating blocks of length* a*, retain one parity, and let* $\widehat{\mathfrak{R}}_\mu^{\mathrm{block}}(\ell \circ \mathcal{F}_{\Lambda,A})$ *be the empirical Rademacher complexity over the* μ *retained block summaries. If* $4(\mu - 1)\beta(a) \leq \delta/2$*, then every* L_ℓ*-Lipschitz loss bounded by* M *satisfies, with probability at least* $1 - \delta$*,* $R(f) \leq \widehat{R}_{\mathrm{ret}}(f) + 2\widehat{\mathfrak{R}}_\mu^{\mathrm{block}}(\ell \circ \mathcal{F}_{\Lambda,A}) + M\sqrt{\log(2/\delta)/\mu}$*, and* $\widehat{\mathfrak{R}}_\mu^{\mathrm{block}}(\ell \circ \mathcal{F}_{\Lambda,A}) \lesssim L_\ell A R_X (\sum_j L_j(T)^2/\mu)^{1/2}$*. For vector outputs, if coordinate* k *has budget* A_k *and* $\ell(y, \hat{y}) = d_y^{-1} \sum_{k=1}^{d_y} \phi(y_k, \hat{y}_k)$ *with* ϕ *bounded and* 1*-Lipschitz in its second argument, then*

$$\widehat{\mathfrak{R}}_n(\ell \circ \mathcal{F}^{\mathrm{vec}}_{\Lambda,A_{1:d_y}}) \leq \frac{R_X}{\sqrt{n}} \left(\sum_{j=1}^{m} L_j(T)^2 \right)^{1/2} \frac{1}{d_y} \sum_{k=1}^{d_y} A_k.$$

For a general L_ℓ*-Lipschitz vector loss in* ℓ_2*, Maurer's vector contraction gives* $\widehat{\mathfrak{R}}_n(\ell \circ \mathcal{F}^{\mathrm{vec}}_{\Lambda,A}) \leq CL_\ell R_X n^{-1/2} \left(\sum_{j=1}^{m} L_j(T)^2 \right)^{1/2} B_{\mathrm{out}}$*, where* C *is universal and* B_{out} *is the output-operator budget.*

Proof sketch. Unroll the diagonal SSM into exponential filters $\psi_j(X) = \sum_t \lambda_j^t x_{T-1-t}$, for which $\|\psi_j(X)\|_2 \leq L_j(T) R_X$ by Cauchy–Schwarz. The product-norm constraint turns the Rademacher supremum into $A(\sum_j \|g_j\|_2^2)^{1/2}$ with $g_j = \sum_i \sigma_i \psi_j(X_i)$. Taking expectation cancels cross terms and gives the memory-energy bound. Vector outputs follow coordinatewise or by Maurer's contraction. Dependent data follow by coupling alternating β-mixing blocks to independent blocks, with retained sample size μ and coupling error $4(\mu - 1)\beta(a)$. □

Memory energy and familiar capacity measures. Memory energy complements the usual width and norm controls. Indeed, for $T \geq 1$ and $|\lambda_j| \leq \rho < 1$,

$$m \leq \sum_j L_j(T)^2 \leq m \min\{T, (1-\rho^2)^{-1}\}.$$

When the poles remain uniformly away from the unit circle, the theorem therefore recovers the familiar $O(\sqrt{m/n})$ dependence, up to a stability factor. Its additional value appears when models have the same width and product-norm budget but different pole locations. In that case, the channelwise energies distinguish their memory profiles without assigning every channel the worst-case cost $\sqrt{m}\, L_{\max}(T)$. Poles closer to the unit circle increase this upper bound, although this alone does not imply a larger observed generalization error.

Selecting the spectrum from data. The fixed-spectrum result also helps distinguish the cost of using a spectrum from the cost of choosing one. If the spectrum is selected using independent data, the theorem applies conditionally. When selection uses the same sample, however, the candidate spectra must also be taken into account. For a nonempty, pre-specified finite family $\mathcal{L}$, write $L_j^{\Lambda}(T)^2 = \sum_{t=0}^{T-1} |\lambda_j|^{2t}$ and define the scalar output envelope $B := AR_X \max_{\Lambda \in \mathcal{L}} (\sum_j L_j^{\Lambda}(T)^2)^{1/2}$. Then

$$\widehat{\mathfrak{R}}_n \left(\bigcup_{\Lambda \in \mathcal{L}} \mathcal{F}_{\Lambda, A} \right) \leq \max_{\Lambda \in \mathcal{L}} \frac{AR_X}{\sqrt{n}} \left(\sum_j L_j^{\Lambda}(T)^2 \right)^{1/2} + B \sqrt{\frac{2 \log |\mathcal{L}|}{n}}.$$

The first term captures memory within a fixed spectrum, while the second accounts for selection among candidates; Appendix A.2 gives the full proof. Extending this argument to a continuous family requires additional uniform stability and covering or Lipschitz control. Consequently, a bound based only on the learned spectrum's realized memory energy does not follow. Input-dependent or selective poles require a different analysis and remain outside the present result.

3.4 Joint Encoder/Decoder Factorization by Spectral Projectors

Theorem 3 (Spectral-projector identifiability for jointly trained diagonal SSMs)**.** *Under Assumption 3, let q be the minimal-polynomial degree of A and $q_+ := \max\{q, 2\}$. If $\mathcal{L}(C, \Lambda, E) := \sum_{h=0}^{q_+-1} \|C\Lambda^h E - A^h\|_F^2 = 0$, then the $h = 0, 1$ terms imply $CE = I$ and $C\Lambda E = A$. The inclusion of $h = 1$ is necessary when $q = 1$, since otherwise only $CE = I$ would be enforced. The learned latent coordinates split into disjoint index sets $I_1, \ldots, I_K$ such that $CP_{I_k}E = \Pi_k$ for every k, where P_{I_k} is the latent coordinate projector. Thus every zero-loss global optimum is factor-aligned at the level of coordinate blocks.*

The remaining ambiguity is the invertible commutant of Λ: after fixing the diagonal ordering, a reparametrization $C \to C_0 R$, $E \to R^{-1}E_0$ preserves zero loss only if $R\Lambda = \Lambda R$. Thus repeated-pole eigenspaces may mix arbitrarily, while simple spectra allow only nonzero diagonal rescaling; simultaneous relabeling of coordinates and diagonal entries recovers the usual permutation ambiguity. More generally, if $p_k(A) = \Pi_k$ with $p_k(x) = \sum_{h=0}^{q-1} a_{k,h} x^h$ and $\max_{0 \leq h \leq q_+-1} \|C\Lambda^h E - A^h\|_F \leq \varepsilon$, then $\|Cp_k(\Lambda)E - \Pi_k\|_F \leq \varepsilon \sum_{h=0}^{q-1} |a_{k,h}|$. This gives approximate spectral-filter recovery; identifying $p_k(\Lambda)$ with a coordinate selector additionally requires the pole-localization and near-inversion assumptions of Theorem 4.

Proof sketch. Zero loss gives $C\Lambda^h E = A^h$ for all $h \leq q_+ - 1$, hence $CE = I$ and $C\Lambda E = A$. Since the factor spectra are disjoint, polynomial functional calculus gives degree-$(q-1)$ projectors p_k with $p_k(A) = \Pi_k$. Applying $Cr(\Lambda)E = r(A)$ to $r = p_k$ yields $Cp_k(\Lambda)E = \Pi_k$. As Λ is diagonal, $p_k(\Lambda)$ selects exactly the latent coordinates whose poles belong to factor k. Since $CE = I$, $E = C^{-1}$, so $CP_{I_k}E = \Pi_k$ implies $P_{I_k}E = E\Pi_k$, meaning those coordinates encode only factor k. The approximate statement follows by expanding p_k and applying the triangle inequality. □

Contribution. Theorem 3 shows that factor alignment follows from diagonal latent dynamics and multi-lag spectral separation. The exact result partitions learned coordinates by factor spectra; the approximate result recovers factor projectors through learned spectral filters.

Extension to rectangular maps. The square assumption allows the theorem to identify factor-coordinate blocks throughout the latent space. A related result remains available for rectangular maps if the active subspace is invariant. To see this, retain the assumptions on A and its factor spectra, and let $E : \mathbb{C}^D \to \mathbb{C}^m$ and $C : \mathbb{C}^m \to \mathbb{C}^D$ satisfy $CE = I_D$ and $C\Lambda E = A$, with diagonal $\Lambda \in \mathbb{C}^{m\times m}$. If $\mathcal{V} := \mathrm{im}(E)$ is Λ-invariant, then

$$\Lambda E = EA, \qquad p_k(\Lambda)E = Ep_k(A) = E\Pi_k.$$

Indeed, invariance gives $\Lambda E = EB$ for some B, and multiplying by C yields $B = A$. Applying the polynomial p_k then gives the second identity, so $E(U_k)$ are factor-aligned invariant subspaces within $\mathcal{V}$. This extension has two limits. First, $CE = I_D$ forces $m \geq D$, so it cannot cover compressive exact reconstruction. Second, when $m > D$, these identities constrain the active subspace without identifying all ambient latent coordinates: decoder-null directions may remain outside $\mathcal{V}$, where $p_k(\Lambda)$ need not be a coordinate projector.

3.5 Robust Approximate Recovery by Contour Resolvents

Theorem 4 (Riesz-projector stability under approximate similarity and near-inversion)**.** *Assume Assumption 3. Let $C, E \in \mathbb{C}^{D\times D}$ be square, define $\Delta := A - C\Lambda E$, $\eta_{CE} := \|CE - I\|_{\mathrm{op}}$, and $\eta_{EC} := \|EC - I\|_{\mathrm{op}}$. For factor k, let Γ_k be a positively oriented simple closed contour with $\Gamma_k \subset \rho(A) \cap \rho(\Lambda)$, enclosing $\mathrm{spec}(A_k)$ and no other factor spectrum, and enclosing exactly the learned poles indexed by $I_k := \{i : \lambda_i \text{ lies inside } \Gamma_k\}$. Define $d_{\Lambda,k} := \min_{z\in\Gamma_k} \mathrm{dist}(z, \mathrm{spec}(\Lambda))$ and $\mathfrak{R}_{A,k} := \sup_{z\in\Gamma_k} \|(zI - A)^{-1}\|_{\mathrm{op}}$. Then*

$$\|CP_{I_k}E - \Pi_k\|_{\mathrm{op}} \leq \frac{\mathrm{len}(\Gamma_k)}{2\pi}\mathfrak{R}_{A,k}\left[\eta_{CE} + \frac{\|C\|_{\mathrm{op}}\|E\|_{\mathrm{op}}}{d_{\Lambda,k}}\left(\|\Delta\|_{\mathrm{op}} + \|\Lambda\|_{\mathrm{op}}\eta_{EC}\right)\right].$$

If $A = S\Lambda_\star S^{-1}$, then $\mathfrak{R}_{A,k} \leq \kappa(S)/d_{A,k}$ with $d_{A,k} := \min_{z\in\Gamma_k} \mathrm{dist}(z, \mathrm{spec}(A))$. When $E = C^{-1}$, both near-inverse defects vanish and the bound becomes

$$\|CP_{I_k}E - \Pi_k\|_{\mathrm{op}} \leq \frac{\mathrm{len}(\Gamma_k)}{2\pi}\mathfrak{R}_{A,k}\frac{\|C\|_{\mathrm{op}}\|E\|_{\mathrm{op}}}{d_{\Lambda,k}}\|\Delta\|_{\mathrm{op}}.$$

Proof sketch. Use the contour representations $\Pi_k = (2\pi i)^{-1}\oint_{\Gamma_k}(zI - A)^{-1}dz$ and $P_{I_k} = (2\pi i)^{-1}\oint_{\Gamma_k}(zI - \Lambda)^{-1}dz$. With $R_\Lambda(z) := (zI - \Lambda)^{-1}$ and $M(z) := CR_\Lambda(z)E$, an approximate-intertwining calculation gives

$$(zI - A)M(z) = CE - C\Lambda(EC - I)R_\Lambda(z)E - \Delta CR_\Lambda(z)E.$$

Therefore $(zI - A)^{-1} - M(z) = (zI - A)^{-1}[I - CE + C\Lambda(EC - I)R_\Lambda(z)E + \Delta CR_\Lambda(z)E]$. Integrating over Γ_k and using $\|R_\Lambda(z)\|_{\mathrm{op}} \leq d_{\Lambda,k}^{-1}$ gives the bound. Diagonalizable A gives $\|(zI - A)^{-1}\|_{\mathrm{op}} \leq \kappa(S)/d_{A,k}$. □

The bound explains why a small prediction residual alone is not enough to interpret learned coordinates as factors. The inverse defects, spectral margins, and conditioning of the dynamics also enter the guarantee. In particular, $\mathfrak{R}_{A,k} \leq \kappa(S)/d_{A,k}$ shows how small margins or poorly conditioned eigenvectors can weaken the bound even when the similarity defect is small. As long as these quantities remain controlled and a separating contour stays valid, the theorem bounds the effect of perturbations on recovery. If a learned pole crosses the contour, that contour choice must be reconsidered; if no contour separates the true factor spectra, the theorem gives no factor-coordinate guarantee. These conditions must therefore be checked in a learned model, since the theorem does not establish that training will achieve them.

3.6 Local Gradient-Flow Companion

Theorem 5 (Local projected gradient flow near the spectral factorization)**.** *Assume Assumption 3 and that $A = S\Lambda_\star S^{-1}$ has simple spectrum. In spectral coordinates, write $C = S(I + X)$, $E = (I + Y)S^{-1}$, and $\Lambda = \Lambda_\star$. Consider $\Phi(X, Y) := \frac{1}{2}\sum_{h=0}^{q_+-1}\|(I + X)\Lambda_\star^h(I + Y) - \Lambda_\star^h\|_F^2$, where $q_+ := \max\{q, 2\}$. Equip complex matrix space with $\langle U, V\rangle_{\mathbb{R}} := \mathrm{Re}\,\mathrm{tr}(U^*V)$, and let $\mathcal{G} := \{(X, Y) : \mathrm{diag}(X) = \mathrm{diag}(Y)\}$ be the gauge slice removing diagonal rescalings. Projected gradient*

flow is $\dot{Z}(t) = -\Pi_{\mathcal{G}} \nabla \Phi(Z(t))$*, where* $Z(t) = (X(t), Y(t))$*. There exist* $r_0, \mu > 0$*, depending on spectral gaps, spectral radius,* q_+*, and* $c_{\Lambda,q_+} := \sqrt{2 \sum_{h=0}^{q_+-1} \|\Lambda_\star^h\|_{\mathrm{op}}^2}$*, such that if* $Z(0) \in \mathcal{G}$ *and* $\|Z(0)\|_F \le r_0$*, then* $\Phi(X(t), Y(t)) \le e^{-\mu t} \Phi(X(0), Y(0))$*. Moreover,* $(X(t), Y(t)) \to (0, 0)$ *in the gauge slice, so* $(C, E) \to (S, S^{-1})$*, which is factor-aligned by Theorem 3.*

Proof sketch. We first linearize the residual as $\mathcal{J}(X, Y)_h = X\Lambda_\star^h + \Lambda_\star^h Y$. The first two lags, together with simple spectrum and the gauge constraint, make this map injective on $\mathcal{G}$. The remaining term, $X\Lambda_\star^h Y$, is quadratic and can therefore be controlled in a sufficiently small neighborhood. This yields a local PL inequality, from which exponential decay under projected gradient flow follows. The full argument is given in Appendix A.5. □

Contribution. This result complements algebraic identifiability with a local convergence guarantee when the poles are fixed and the gauge is controlled. It does not, however, establish convergence from arbitrary initialization or under unrestricted joint training.

3.7 Supporting Consequences for OOD Prediction, Rollout, and Control

The spectral results describe what a representation preserves and when its coordinates align with factors. Connecting these properties to downstream performance requires further assumptions. The next three propositions make those assumptions explicit and follow independently from Assumption 4. For example, factor alignment from Theorem 3 can support Proposition 1, but only when the required support-uniform factor errors are also controlled. The rollout and control bounds similarly require information about the decoder, prediction errors, and task metric.

Proposition 1 (Support-uniform compositional OOD bound)**.** *Under Assumption 4, if a factored predictor satisfies* $\sup_{(s^{(k)},a) \in \operatorname{supp} Q_k} \|h_k(s^{(k)}, a) - T_k(s^{(k)}, a)\|_2 \le \epsilon_k$ *for every factor-marginal test support* Q_k*, then every joint OOD distribution* Q *over those factor supports satisfies* $\mathbb{E}_Q \|h(s, a) - T(s, a)\|_2^2 \le \sum_{k=1}^K \epsilon_k^2$.

Proof sketch. Orthogonality gives $\|h(s, a) - T(s, a)\|_2^2 = \sum_k \|h_k(s^{(k)}, a) - T_k(s^{(k)}, a)\|_2^2$. Each summand is pointwise at most ϵ_k^2 on the test support, so expectation under any joint Q gives the result. □

Proposition 2 (Observation-space rollout envelope)**.** *Under Assumption 4, the true observation-space rollout error satisfies* $\|\mathcal{D}(\hat{z}_H) - s_{t+H}\|_2 \le \alpha + L_D \epsilon (1 - \rho^H)/(1 - \rho)$*. Thus lower contraction alone does not rank pixel rollouts; the full envelope* $\alpha + L_D \epsilon (1 - \rho^H)/(1 - \rho)$ *must be compared.*

Proof sketch. Let $e_h = \hat{z}_h - z_h^\star$ with $z_h^\star = \mathcal{E}(s_{t+h})$. Contractivity and one-step latent error imply $\|e_{h+1}\|_2 \le \rho \|e_h\|_2 + \epsilon$, hence $\|e_H\|_2 \le \epsilon (1 - \rho^H)/(1 - \rho)$. Decoder Lipschitzness converts latent error to observation error, and reconstruction contributes α. □

Proposition 3 (Prediction-control mismatch)**.** *Under Assumption 4, for a fixed policy in a discounted control problem,* $|J(\pi) - \hat{J}(\pi)| \le (\eta_r + \gamma L_V \eta_T)/(1 - \gamma)$*. Pixel MSE or SSIM implies control quality only when it controls reward error and task-relevant transition error.*

Proof sketch. Evaluate both Bellman operators on V^π. The one-step discrepancy is at most $\eta_r + \gamma L_V \eta_T$, where the first term is reward error and the second follows from Lipschitzness of V^π in the task metric. Since $\|V^\pi - \hat{V}^\pi\|_\infty \le \|T^\pi V^\pi - \hat{T}^\pi V^\pi\|_\infty + \gamma \|V^\pi - \hat{V}^\pi\|_\infty$, rearranging gives the factor $(1 - \gamma)^{-1}$. □

4 Numerical Illustrations and Results

The generated-system suite constructs small systems satisfying the assumptions of Section 3, with each main theorem paired with a numerical illustration whose data-generating process matches the corresponding assumptions. Appendix-level support is provided for Theorems 4–5 and Propositions 1–3, and saved DMControl [Tassa et al., 2018] pixel tables are recorded in Appendix C as motivating observations for Propositions 2 and 3.

For Theorem 1, we generate an LDS with state dimension $D = 6$, rank-3 transition, horizon $H = 4$, independent Gaussian actions, and $n = 8000$ trajectory windows. The stacked Krylov operator is estimated by least squares and its empirical singular values are compared with the population operator. For Theorem 2, we compare two fixed-spectrum diagonal SSM filter classes on bounded-energy sequences, with fast poles in $[0.15, 0.55]$ and slow poles in $[0.72, 0.96]$, estimating empirical Rademacher complexity using 400 random sign draws. For Theorem 3, we build a block-diagonal linear system with three disjoint factor spectra and compute the spectral-projector polynomials recovered through diagonal latent dynamics. Appendix B records a local projected-gradient instance for Theorem 5, together with illustrations for Theorem 4 and Propositions 1–3.

The headline numerical outcomes are summarized in Table 1.

Table 1: Generated-system illustrations under the theorem assumptions.

Result	Construction	Observed quantity
Theorem 1	Rank-3 LDS, $H = 4$, $n = 8000$	Population rank 3, empirical rank 3, max singular-value error 5.53×10^{-3}
Theorem 2	Fast vs. slow fixed SSM poles	Slow/fast memory-energy ratio $4.04\times$, empirical complexity ratio $2.02\times$, both estimates below the bound
Theorem 3	Block system with disjoint spectra	Max projector error 1.64×10^{-13}, max multi-lag residual 3.39×10^{-15}

The Theorem 1 illustration recovers the shared-latent elbow from data: thresholding the empirical singular values at the population mid-gap gives rank 3, matching $\mathrm{rank}(\mathcal{K}_4)$.

The Theorem 2 illustration isolates the spectral quantity in the bound. Slow poles increase memory energy from 9.414 to 38.020 and increase the empirical Rademacher estimate from 0.0201 to 0.0405, consistent with the increase in the stable-filter memory term $\sum_j L_j(T)^2$ while parameter count, width, and product-norm budget remain unchanged. On the same construction, the coarser worst-channel summary $\sqrt{m} L_{\max}(T)$ equals 3.39 for fast poles and 9.91 for slow poles, while the exact memory-energy term gives 3.07 and 6.17, showing that the theorem avoids worst-channel overestimation inside heterogeneous diagonal classes.

The Theorem 3 illustration confirms the algebraic mechanism behind factor recovery. With disjoint factor spectra, polynomial projectors recover the three factor subspaces up to numerical precision. Appendix B records the contour-separation illustration for Theorem 4, the local projected-gradient instance for Theorem 5, and the supporting illustrations for Propositions 1–3.

5 Discussion & Conclusion

5.1 Discussion

What the Spectral View Explains The generated-system illustrations reproduce the numerical behavior predicted by the proofs under their stated assumptions, and the measured quantities are the formal objects appearing in the results: Krylov singular values, stable-filter memory energy, spectral projectors, rollout envelopes, and discounted control residuals. Across the paper, the main results reduce world-model questions to spectral objects on known operators: the block-Krylov operator $\mathcal{K}_H$ for shared-latent compression, the diagonal operator Λ for stable-filter complexity, and the spectral projectors Π_k for factor recovery.

This perspective explains when seemingly similar architectures behave differently. In Theorem 1, a non-normal A separates one-step truncation from multi-horizon truncation, since shared-horizon singular directions need not align with one-step directions. In Theorem 2, pole magnitude separates same-budget diagonal SSM classes through the memory energy $\sum_j L_j(T)^2$, even when parameter count and product-norm budget are fixed. In Theorem 3, disjoint factor spectra together with diagonal latent dynamics force coordinate-level factor alignment, while the contour-resolvent extension in Theorem 4 controls approximate recovery through the defects η_{CE}, η_{EC}, and $\|\Delta\|_{\mathrm{op}}$.

Scope of the implications. These results provide exact guarantees within the stated linear settings. For nonlinear world models, they instead suggest quantities to examine before drawing conclusions

about latent sufficiency, memory, or factor alignment. In particular, the multi-horizon spectrum can guide compression in a fitted linear model, while factor interpretation requires separated spectral clusters and small encoder–decoder defects. The practical implication is therefore to check these conditions and task-relevant errors, rather than select a nonlinear architecture directly from the linear theorems.

Rollout and Control Metrics The saved visual tables in Appendix C provide a concrete example of the prediction–control mismatch. On Walker-Walk, SSM-Mamba achieves better visual MSE and SSIM ($0.0038/0.8238$), whereas RSSM-GRU obtains the higher return (41.950 versus 27.916). These single-seed observations record the ranking inversion, but do not establish its mechanism. Propositions 2 and 3 identify the additional reconstruction, decoder, reward, and task-metric quantities needed for downstream guarantees. Since these quantities were not estimated for the reported checkpoints, the tables motivate those bounds rather than validate them. Thus, world-model comparisons should examine both predictive representations and the errors relevant to the downstream task.

5.2 Limitations & Future Work

Our guarantees concern linear settings with theorem-specific restrictions. Under feedback policies, Theorem 1 requires the residualized operator $\mathcal{K}_{H,\mathrm{FWL}}$. Theorem 2 assumes fixed stable diagonal spectra and bounded input energy, excluding selective SSMs. Factor identifiability requires square maps, diagonalizable autonomous dynamics, and disjoint factor spectra; rectangular recovery needs additional structure. Moreover, Theorem 4 assumes controlled similarity and inverse defects, while Theorem 5 guarantees only local, gauge-constrained convergence with fixed poles. The downstream propositions likewise require factored, contractive, or Lipschitz structure.

Future work should estimate the spectral and downstream quantities on learned models, relate training to the recovery defects, and extend the analysis to noisy optimization and input-dependent dynamics. Beyond linear settings, a finite-dimensional invariant lift satisfying $\phi(T(s,a)) = A\phi(s) + Ba$ could make the results applicable under their remaining assumptions. Trajectory-wise Jacobian products offer a local Krylov analogue, but global guarantees require control of accumulated linearization error. Nonlinear factor identifiability would likewise need additional temporal or auxiliary-variable structure.

5.3 Conclusion

In summary, this paper develops a spectral theory of world models in linear settings. Theorem 1 identifies the shared-latent multi-horizon bottleneck through the block-Krylov operator and gives a perturbation condition for recovering the elbow from data. Theorem 2 gives an architecture-specific generalization bound for stable diagonal SSMs in terms of spectral memory energy. Theorem 3 shows that jointly trained diagonal SSMs recover factor-aligned latent coordinates at zero population loss under disjoint factor spectra. Theorem 4 gives a contour-resolvent stability theorem for approximate factor recovery, and Theorem 5 provides a local projected-gradient-flow companion near the gauge-fixed spectral solution. Together, the results and generated-system illustrations clarify how spectral structure governs compression, generalization, and identifiability in the stated linear settings, while the supporting propositions specify additional conditions needed to relate predictive representations to rollout and control guarantees.

Funding Statement & Competing Interests

The authors received no external funding specifically for this research. The authors declare no competing financial interests related to this work.

Acknowledgements

The authors gratefully acknowledge N2TP Technology Solutions JSC. for providing the computational resources that supported this research. We also extend our sincere appreciation to our collaborators, advisors, and colleagues for their valuable contributions.

References

Athanasios C Antoulas. *Approximation of large-scale dynamical systems*. SIAM, 2005.

Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1233–1243. PMLR, 26–28 Aug 2020.

Tri Dao and Albert Gu. Transformers are ssms: generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Deividas Eringis, John Leth, Zheng-Hua Tan, Rafal Wisniewski, and Mihály Petreczky. Pac-bayes generalisation bounds for dynamical systems including stable rnns. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

D. Hafner, J. Pasukonis, J. Ba, et al. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025a.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025b.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018. URL `http://jmlr.org/papers/v19/16-465.html`.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6705–6715, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 26–28 Aug 2020.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106:93–117, 2017. doi: 10.1007/s10994-016-5588-2.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 09–15 Jun 2019.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings*, page 3–17, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 978-3-319-46378-0.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

Milton L. Montero, Jeffrey Bowers, Rui Ponte Costa, Casimir JH Ludwig, and Gaurav Malhotra. Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American Control Conference (ACC)*, pages 5655–5661, 2019.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5610–5618. PMLR, 09–15 Jun 2019.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, 06–09 Jul 2018.

Namjoon Suh and Guang Cheng. A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models. *Annual Review of Statistics and Its Application*, 12: 177–207, 2025.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite, 2018.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Peter Van Overschee and Bart De Moor. *Subspace Identification for Linear Systems: Theory—Implementation—Applications*. Kluwer Academic Publishers, New York, NY, 1 edition, 1996. ISBN 978-1-4613-0465-4. doi: 10.1007/978-1-4613-0465-4.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. Storm: efficient stochastic transformer based world models for reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A Full Proofs

A.1 Proof of Theorem 1

Proof. The h-step state is

$$s_{t+h} = A^h s_t + \sum_{j=0}^{h-1} A^j B a_{t+h-1-j} + \sum_{j=0}^{h-1} A^j \xi_{t+h-1-j}.$$

Stacking $h = 1, \dots, H$ gives $Y_{t,H} = \mathcal{A}_H s_t + \mathcal{B}_H a_{t:t+H-1} + \eta_{t,H}$, where

$$\mathcal{A}_H = \begin{bmatrix} A \\ A^2 \\ \vdots \\ A^H \end{bmatrix},$$

$\mathcal{B}_H$ is the block lower-triangular action operator, and $\eta_{t,H}$ is the stacked process-noise vector with covariance Γ_H.

Because the action sequence is observed without latent compression and satisfies $\mathbb{E}[s_t a_{t:t+H-1}^\top] = 0$, the optimal V equals $\mathcal{B}_H$ in population risk and the residual state-prediction term is orthogonal to the action regressors. Thus

$$\inf_{E,U,V} \mathbb{E}\|Y_{t,H} - UEs_t - Va_{t:t+H-1}\|_2^2 = \inf_{E,U} \mathbb{E}\|(\mathcal{A}_H - UE)s_t + \eta_{t,H}\|_2^2.$$

By independence and zero mean,

$$\mathbb{E}\|(\mathcal{A}_H - UE)s_t + \eta_{t,H}\|_2^2 = \operatorname{tr}\left((\mathcal{A}_H - UE)\Sigma_s(\mathcal{A}_H - UE)^\top\right) + \operatorname{tr}(\Gamma_H).$$

Let $K = UE$. Since E has r rows, $\operatorname{rank}(K) \le r$. Conversely, every matrix $K \in \mathbb{R}^{HD\times D}$ with rank at most r can be factored as $K = UE$ with $U \in \mathbb{R}^{HD\times r}$ and $E \in \mathbb{R}^{r\times D}$. Therefore

$$\inf_{E,U} \operatorname{tr}\left((\mathcal{A}_H - UE)\Sigma_s(\mathcal{A}_H - UE)^\top\right) = \inf_{\operatorname{rank}(K)\le r} \|(\mathcal{A}_H - K)\Sigma_s^{1/2}\|_F^2.$$

Set $L = K\Sigma_s^{1/2}$. Since $\Sigma_s^{1/2}$ is invertible, $\operatorname{rank}(L) = \operatorname{rank}(K) \le r$, and every rank-$r$ L corresponds to $K = L\Sigma_s^{-1/2}$. Hence

$$\inf_{\operatorname{rank}(K)\le r} \|(\mathcal{A}_H - K)\Sigma_s^{1/2}\|_F^2 = \inf_{\operatorname{rank}(L)\le r} \|\mathcal{K}_H - L\|_F^2.$$

Eckart–Young–Mirsky gives $\inf_{\operatorname{rank}(L)\le r} \|\mathcal{K}_H - L\|_F^2 = \sum_{i>r} \sigma_i(\mathcal{K}_H)^2$. Adding $\operatorname{tr}(\Gamma_H)$ proves the risk formula.

The systematic term is zero exactly when all singular values after index r vanish, i.e. when $r \ge \operatorname{rank}(\mathcal{K}_H)$.

For perturbation, Weyl's inequality for singular values gives $|\sigma_i(\widehat{\mathcal{K}}_H) - \sigma_i(\mathcal{K}_H)| \le \|\widehat{\mathcal{K}}_H - \mathcal{K}_H\|_{\mathrm{op}} \le \varepsilon$. If $\sigma_r(\mathcal{K}_H) > \tau + \varepsilon$ and $\sigma_{r+1}(\mathcal{K}_H) < \tau - \varepsilon$, then

$$\sigma_i(\widehat{\mathcal{K}}_H) > \tau \quad (i \le r), \qquad \sigma_i(\widehat{\mathcal{K}}_H) < \tau \quad (i > r),$$

so thresholding empirical singular values at τ recovers exactly r singular directions.

The estimator used in the perturbation statement is the least-squares stacked horizon estimator:

$$(\widehat{\mathcal{A}}_H, \widehat{\mathcal{B}}_H) \in \arg\min_{M,N} \sum_{i=1}^{n} \|Y_{i,H} - Ms_i - Na_{i:i+H-1}\|_2^2,$$

with $\widehat{\mathcal{K}}_H = \widehat{\mathcal{A}}_H \widehat{\Sigma}_s^{1/2}$ and $\widehat{\Sigma}_s = n^{-1}\sum_i s_i s_i^\top$.

Under non-exogenous or feedback data, the action sequence $W_t = a_{t:t+H-1}$ can be statistically coupled with the future process noise $\eta_{t,H}$, so $\mathbb{E}[\eta_{t,H} W_t^\top] \ne 0$ in general and the optimum over V no longer simply deletes the action term in the shared-latent risk. By the Frisch–Waugh–Lovell

theorem, the population minimum of $\mathbb{E}\|Y_{t,H} - UEs_t - VW_t\|_2^2$ over V at fixed E, U equals the optimum of the partialled risk $\mathbb{E}\|\tilde{Y}_{t,H} - UE\tilde{s}_t\|_2^2$, where $\tilde{s}_t := s_t - \Pi_W s_t$, $\tilde{Y}_{t,H} := Y_{t,H} - \Pi_W Y_{t,H}$, and Π_W denotes the L_2 projection onto W_t. Decomposing

$$\tilde{Y}_{t,H} = M_H \tilde{s}_t + \zeta_{t,H}, \qquad M_H = \mathrm{Cov}(\tilde{Y}_{t,H}, \tilde{s}_t)\tilde{\Sigma}_s^{\dagger}, \qquad \mathbb{E}[\zeta_{t,H}\tilde{s}_t^{\top}] = 0,$$

gives

$$\mathbb{E}\|\tilde{Y}_{t,H} - UE\tilde{s}_t\|_2^2 = \mathrm{tr}\big((M_H - UE)\tilde{\Sigma}_s(M_H - UE)^{\top}\big) + \mathrm{tr}(\Omega_H),$$

with $\Omega_H = \mathrm{Cov}(\zeta_{t,H}) = \mathrm{Cov}(\tilde{Y}_{t,H}) - M_H\tilde{\Sigma}_s M_H^*$. Repeating the rank-$r$ Eckart–Young–Mirsky argument with $\mathcal{K}_{H,\mathrm{FWL}} = M_H\tilde{\Sigma}_s^{1/2}$ in place of $\mathcal{K}_H$ yields

$$\inf_{E,U,V} \mathbb{E}\|Y_{t,H} - UEs_t - VW_t\|_2^2 = \sum_{i>r} \sigma_i(\mathcal{K}_{H,\mathrm{FWL}})^2 + \mathrm{tr}(\Omega_H).$$

Under exogenous actions and process-noise orthogonality, one has $\Pi_W s_t = 0$, $\Pi_W Y_{t,H} = \mathcal{B}_H W_t$, $M_H = \mathcal{A}_H = [A; \dots; A^H]$, $\tilde{\Sigma}_s = \Sigma_s$, and $\Omega_H = \Gamma_H$, recovering the displayed Krylov formula.

For independent sub-Gaussian trajectory windows with actions in $\mathbb{R}^{d_a}$, write the full design vector as $x_i = (s_i^{\top}, a_{i:i+H-1}^{\top})^{\top} \in \mathbb{R}^{D+Hd_a}$. Assume the design covariance is well conditioned, $\lambda_{\min}(\mathbb{E}x_i x_i^{\top}) \geq \kappa_x^{-1}$, the stacked regression noise is sub-Gaussian with scale σ_H, and the finite-horizon operators and state covariance have bounded operator norm. Standard multivariate sub-Gaussian regression concentration gives

$$\|\widehat{\mathcal{A}}_H - \mathcal{A}_H\|_{\mathrm{op}} + \|\widehat{\mathcal{B}}_H - \mathcal{B}_H\|_{\mathrm{op}} + \|\widehat{\Sigma}_s^{1/2} - \Sigma_s^{1/2}\|_{\mathrm{op}} \leq C\,\kappa_x\sigma_H\sqrt{\frac{HD + D + Hd_a + \log(1/\delta)}{n}},$$

after absorbing fixed finite-horizon norm factors into the constant. Multiplying out

$$\widehat{\mathcal{K}}_H - \mathcal{K}_H = (\widehat{\mathcal{A}}_H - \mathcal{A}_H)\widehat{\Sigma}_s^{1/2} + \mathcal{A}_H(\widehat{\Sigma}_s^{1/2} - \Sigma_s^{1/2})$$

gives the stated shorthand rate

$$\|\widehat{\mathcal{K}}_H - \mathcal{K}_H\|_{\mathrm{op}} = O\left(L_H\sqrt{\frac{HD + D + Hd_a + \log(1/\delta)}{n}}\right),$$

where L_H collects κ_x, σ_H, state-covariance conditioning, and finite-horizon action/noise energy. This is the usual matrix-regression template; see [Vershynin, 2018, Ch. 4].

The general/feedback statement above uses the population projection Π_W. If Π_W is estimated, a held-out or cross-fitted residualization step can be used; an event $\|(\widehat{\Pi}_W - \Pi_W)s\|_{L_2} + \|(\widehat{\Pi}_W - \Pi_W)Y\|_{L_2} \leq \eta_{\mathrm{res}}$ adds an extra perturbation term of order $\|M_H\|_{\mathrm{op}}\eta_{\mathrm{res}}$ to $\|\widehat{\mathcal{K}}_{H,\mathrm{FWL}} - \mathcal{K}_{H,\mathrm{FWL}}\|_{\mathrm{op}}$. Combining the concentration event with the deterministic perturbation result gives sample elbow recovery. □

A.2 Proof of Theorem 2

Proof. For channel j, define the filtered feature $\psi_j(X) := \sum_{t=0}^{T-1} \lambda_j^t x_{T-1-t}$. By Cauchy–Schwarz,

$$\|\psi_j(X)\|_2 \leq \left(\sum_{t=0}^{T-1} |\lambda_j|^{2t}\right)^{1/2} \left(\sum_{t=0}^{T-1} \|x_{T-1-t}\|_2^2\right)^{1/2} \leq L_j(T)R_X.$$

Let $\sigma_1, \dots, \sigma_n$ be independent Rademacher variables. The empirical complexity is

$$\widehat{\mathfrak{R}}_n = \frac{1}{n}\mathbb{E}_{\sigma} \sup_{\sum_j c_j^2\|b_j\|^2 \leq A^2} \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{m} c_j b_j^{\top}\psi_j(X_i).$$

For fixed signs, write $g_j := \sum_{i=1}^{n} \sigma_i \psi_j(X_i)$. Then

$$\sup_{\sum_j c_j^2\|b_j\|^2 \leq A^2} \sum_j c_j b_j^{\top} g_j \leq A\left(\sum_{j=1}^{m} \|g_j\|_2^2\right)^{1/2},$$

because the quantities $a_j = c_j b_j$ range over vectors with $\sum_j \|a_j\|_2^2 \le A^2$.

Taking expectation and applying Jensen,

$$\mathbb{E}_\sigma \left(\sum_j \|g_j\|_2^2 \right)^{1/2} \le \left(\sum_j \mathbb{E}_\sigma \|g_j\|_2^2 \right)^{1/2}.$$

For each j, $\mathbb{E}_\sigma \|g_j\|_2^2 = \mathbb{E}_\sigma \|\sum_i \sigma_i \psi_j(X_i)\|_2^2 = \sum_i \|\psi_j(X_i)\|_2^2$, since the cross terms vanish by independence and zero mean of the Rademacher signs. Using the feature norm bound, $\sum_i \|\psi_j(X_i)\|_2^2 \le n R_X^2 L_j(T)^2$. Therefore

$$\widehat{\mathfrak{R}}_n \le \frac{A}{n} \left(\sum_j n R_X^2 L_j(T)^2 \right)^{1/2} = \frac{A R_X}{\sqrt{n}} \left(\sum_j L_j(T)^2 \right)^{1/2}.$$

For vector outputs with coordinates $k = 1, \dots, d_y$, let $\ell(y, \hat{y}) = d_y^{-1} \sum_{k=1}^{d_y} \phi(y_k, \hat{y}_k)$, where ϕ is bounded and 1-Lipschitz in its second argument, and let coordinate k have product-norm budget A_k. Apply the scalar calculation to each coordinate class. The empirical Rademacher complexity of the induced averaged loss class is subadditive over coordinates, hence

$$\widehat{\mathfrak{R}}_n^{\mathrm{vec}} \le \frac{R_X}{\sqrt{n}} \left(\sum_j L_j(T)^2 \right)^{1/2} \cdot \frac{1}{d_y} \sum_{k=1}^{d_y} A_k.$$

For a general L_ℓ-Lipschitz Euclidean vector loss, Maurer's [Maurer, 2016] vector-contraction inequality gives the same memory-energy term combined with L_ℓ and the chosen output-norm budget, up to a universal constant C that depends on the contraction convention. In both cases, the diagonal-SSM contribution remains the memory-energy term $\sum_j L_j(T)^2$; the only additional multiplicative factors are the output-norm budget and, for general vector losses, the universal constant.

For $|\lambda_j| < 1$, $L_j(T)^2 = (1 - |\lambda_j|^{2T})/(1 - |\lambda_j|^2)$, which is the stable memory energy of channel j.

For the β-mixing extension, partition the trajectory into 2μ consecutive blocks $B_1, \dots, B_{2\mu}$ of length a and retain one parity, for example $B_1, B_3, \dots, B_{2\mu-1}$. Let

$$\widehat{R}_{\mathrm{ret}}(f) := \frac{1}{\mu} \sum_{r=1}^{\mu} \frac{1}{a} \sum_{i \in B_{2r-1}} \ell(f(X_i), Y_i)$$

be the retained-block empirical risk and let $\widehat{\mathfrak{R}}_\mu^{\mathrm{block}}(\ell \circ \mathcal{F}_{\Lambda,A})$ denote the empirical Rademacher complexity over the μ retained blocks, defined by drawing one Rademacher sign per block and averaging the per-block within-block loss summary. The independent-block coupling of Mohri and Rostamizadeh [Mohri and Rostamizadeh, 2008] replaces the retained blocks by independent copies with total-variation failure at most $4(\mu - 1)\beta(a)$. If this quantity is at most $\delta/2$, then applying the independent-block Rademacher bound with confidence $\delta/2$ and adding bounded-loss concentration yields, with probability at least $1 - \delta$,

$$R(f) \le \widehat{R}_{\mathrm{ret}}(f) + 2\, \widehat{\mathfrak{R}}_\mu^{\mathrm{block}}(\ell \circ \mathcal{F}_{\Lambda,A}) + M \sqrt{\frac{\log(2/\delta)}{\mu}}.$$

The scalar Rademacher calculation above, applied with the per-block summarized features, gives

$$\widehat{\mathfrak{R}}_\mu^{\mathrm{block}}(\ell \circ \mathcal{F}_{\Lambda,A}) \lesssim L_\ell A R_X \left(\frac{\sum_j L_j(T)^2}{\mu} \right)^{1/2},$$

which is the displayed blocked generalization inequality. □

A.3 Proof of Theorem 3

Proof. Since A is diagonalizable and the spectra of the blocks A_k are pairwise disjoint, there exists for each factor k a polynomial p_k such that

$$p_k(z) = 1 \quad \text{for } z \in \mathrm{spec}(A_k), \qquad p_k(z) = 0 \quad \text{for } z \in \mathrm{spec}(A_\ell),\ \ell \ne k,$$

by Lagrange interpolation on the distinct eigenvalues of A. Since q is the degree of the minimal polynomial, p_k can be represented modulo the minimal polynomial with degree at most $q-1$. Therefore $p_k(A) = \Pi_k$.

Assume first that $\mathcal{L}(C, \Lambda, E) = 0$. Then for every $h = 0, \ldots, q_+ - 1$, $C\Lambda^h E = A^h$. Because $q_+ \geq 2$, the lags $h = 0$ and $h = 1$ are always present, so the identities $CE = I$ and $C\Lambda E = A$ hold. The $h = 0$ identity makes C and E square invertible, and the $h = 1$ identity then gives $\Lambda = EAE^{-1}$, so Λ is similar to A in the square case. Without enlarging q to q_+, the case $q = 1$ would only enforce $CE = I$, leaving Λ unconstrained; the use of $q_+ = \max\{q, 2\}$ is precisely what closes this edge case. Higher lags up to $q-1$ are then redundant under the $h = 0, 1$ identities, since $E = C^{-1}$ in the square case forces $\Lambda = C^{-1}AC$ and hence $C\Lambda^h E = C(C^{-1}AC)^h C^{-1} = A^h$ for every $h \geq 0$.

Let $r(x) = \sum_{h=0}^{q-1} r_h x^h$ be any polynomial of degree at most $q-1$. By linearity,

$$Cr(\Lambda)E = C\left(\sum_{h=0}^{q-1} r_h \Lambda^h\right) E = \sum_{h=0}^{q-1} r_h C\Lambda^h E = \sum_{h=0}^{q-1} r_h A^h = r(A).$$

Choosing $r = p_k$ gives $Cp_k(\Lambda)E = p_k(A) = \Pi_k$.

It remains to identify $p_k(\Lambda)$ as a coordinate selector. Since Λ is diagonal, $p_k(\Lambda)$ is diagonal with entries $p_k(\lambda_i)$. The additional content here is that disjoint factor spectra force a partition of the learned diagonal coordinates by factor, and the multi-lag formulation gives the approximate polynomial-projector statement below. Because p_k is one on the eigenvalues belonging to factor k and zero on the eigenvalues belonging to other factors, $p_k(\Lambda)$ is a diagonal projector onto the latent coordinates whose poles belong to factor k. Denote that coordinate set by I_k. Then $p_k(\Lambda) = P_{I_k}$ and $CP_{I_k}E = \Pi_k$. Since $CE = I$ in the square case, $E = C^{-1}$. Multiplying the last identity on the left by E yields $P_{I_k}E = E\Pi_k$, which says exactly that the latent coordinates indexed by I_k depend only on factor subspace U_k.

The sets I_k are disjoint and cover all coordinates associated with the spectrum of A, so the latent representation is factor-aligned at the level of factor coordinate blocks. To pin down the remaining non-identifiability class, fix a zero-loss reference (C_0, Λ, E_0) with $C_0E_0 = I$ and $C_0\Lambda E_0 = A$, and consider $C = C_0R$, $E = R^{-1}E_0$ for invertible R. Then $CE = I$ automatically, but

$$C\Lambda E = C_0 R\Lambda R^{-1}E_0 = A \iff R\Lambda R^{-1} = \Lambda \iff R\Lambda = \Lambda R.$$

Hence, after fixing the diagonal ordering of Λ, the allowed reparametrizations are exactly the invertible commutant of Λ: invertible mixing is allowed only inside repeated-pole eigenspaces. Permutation of two distinct-pole coordinates does not commute with Λ and is therefore not in the commutant under fixed diagonal ordering; for example, with $\Lambda = \mathrm{diag}(1, 2)$ and the swap $P = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, $P\Lambda - \Lambda P \neq 0$. An invertible R that mixes coordinates with distinct learned poles violates $R\Lambda = \Lambda R$ and breaks the $h = 1$ identity, even when it preserves the factor projectors $CP_{I_k}E$. Under simple spectra and fixed pole ordering, the commutant consists exactly of nonzero diagonal matrices, so the continuous ambiguity reduces to nonzero coordinate rescaling. The usual coordinate-permutation ambiguity reappears only after additionally quotienting by simultaneous relabeling of latent coordinates and the diagonal entries of Λ.

For the approximate statement, suppose $\max_{0 \leq h \leq q_+ - 1} \|C\Lambda^h E - A^h\|_F \leq \varepsilon$. Write $p_k(x) = \sum_{h=0}^{q-1} a_{k,h} x^h$. Then

$$Cp_k(\Lambda)E - p_k(A) = \sum_{h=0}^{q-1} a_{k,h}(C\Lambda^h E - A^h).$$

By the triangle inequality,

$$\|Cp_k(\Lambda)E - \Pi_k\|_F = \|Cp_k(\Lambda)E - p_k(A)\|_F \leq \sum_{h=0}^{q-1} |a_{k,h}| \, \|C\Lambda^h E - A^h\|_F \leq \varepsilon \sum_{h=0}^{q-1} |a_{k,h}|.$$

The coefficient sum $\sum_h |a_{k,h}|$ is a spectral-separation constant: it grows when the factor spectra approach one another because the interpolation polynomial must change rapidly between nearby spectral sets. □

A.4 Proof of Theorem 4

Proof. Define $\Delta := A - C\Lambda E$, $\eta_{CE} := \|CE - I\|_{\mathrm{op}}$, and $\eta_{EC} := \|EC - I\|_{\mathrm{op}}$. For each factor k, let Γ_k be a positively oriented simple closed contour with $\Gamma_k \subset \rho(A) \cap \rho(\Lambda)$, enclosing $\mathrm{spec}(A_k)$ and no other factor block of $\mathrm{spec}(A)$, and crossing no learned pole. Define $I_k := \{i : \lambda_i \text{ lies in the interior of } \Gamma_k\}$. Because $\Gamma_k \subset \rho(A)$ encloses exactly the spectrum of factor A_k and no other factor spectrum, the Riesz projector for factor k is

$$\Pi_k = \frac{1}{2\pi i} \oint_{\Gamma_k} (zI - A)^{-1}\, dz.$$

Because Λ is diagonal, $\Gamma_k \subset \rho(\Lambda)$, and Γ_k encloses exactly the learned poles indexed by I_k,

$$P_{I_k} = \frac{1}{2\pi i} \oint_{\Gamma_k} (zI - \Lambda)^{-1}\, dz.$$

Multiplying by C on the left and E on the right yields

$$CP_{I_k}E = \frac{1}{2\pi i} \oint_{\Gamma_k} C(zI - \Lambda)^{-1}E\, dz.$$

Write $R_\Lambda(z) := (zI - \Lambda)^{-1}$ and $M(z) := CR_\Lambda(z)E$. Since $zR_\Lambda(z) - \Lambda R_\Lambda(z) = I$ and $A = C\Lambda E + \Delta$,

$$\begin{aligned}(zI - A)M(z) &= zCR_\Lambda(z)E - (C\Lambda E + \Delta)CR_\Lambda(z)E \\ &= C\left(zR_\Lambda(z) - \Lambda ECR_\Lambda(z)\right)E - \Delta CR_\Lambda(z)E \\ &= CE - C\Lambda(EC - I)R_\Lambda(z)E - \Delta CR_\Lambda(z)E.\end{aligned}$$

Therefore

$$(zI - A)^{-1} - M(z) = (zI - A)^{-1}\left[I - CE + C\Lambda(EC - I)R_\Lambda(z)E + \Delta CR_\Lambda(z)E\right].$$

Integrating over the contour gives

$$\Pi_k - CP_{I_k}E = \frac{1}{2\pi i} \oint_{\Gamma_k} \left((zI - A)^{-1} - CR_\Lambda(z)E\right) dz.$$

Taking operator norms,

$$\|\Pi_k - CP_{I_k}E\|_{\mathrm{op}} \le \frac{\mathrm{len}(\Gamma_k)}{2\pi}\, \mathfrak{R}_{A,k} \left[\eta_{CE} + \frac{\|C\|_{\mathrm{op}}\|E\|_{\mathrm{op}}}{d_{\Lambda,k}} \left(\|\Delta\|_{\mathrm{op}} + \|\Lambda\|_{\mathrm{op}}\eta_{EC}\right)\right],$$

because $\|R_\Lambda(z)\|_{\mathrm{op}} \le d_{\Lambda,k}^{-1}$ for $z \in \Gamma_k$.

If $A = S\Lambda_\star S^{-1}$ is diagonalizable, then $(zI - A)^{-1} = S(zI - \Lambda_\star)^{-1}S^{-1}$, so

$$\|(zI - A)^{-1}\|_{\mathrm{op}} \le \|S\|_{\mathrm{op}}\|S^{-1}\|_{\mathrm{op}}\|(zI - \Lambda_\star)^{-1}\|_{\mathrm{op}} \le \frac{\kappa(S)}{d_{A,k}},$$

for $d_{A,k} := \min_{z\in\Gamma_k} \mathrm{dist}(z, \mathrm{spec}(A))$. Taking the supremum over Γ_k yields the displayed resolvent envelope. If $E = C^{-1}$, then $CE = EC = I$, so $\eta_{CE} = \eta_{EC} = 0$ and the exact-inverse bound is recovered. □

A.5 Proof of Theorem 5

Proof. Define the residual map $F : \mathcal{G} \to (\mathbb{C}^{D\times D})^{q_+}$ by

$$F_h(X, Y) = (I + X)\Lambda_\star^h(I + Y) - \Lambda_\star^h, \qquad h = 0, \dots, q_+ - 1,$$

with $q_+ = \max\{q, 2\}$ as in Assumption 3 and Theorem 3, so the lags $h = 0, 1$ are always present. All gradients below are real gradients under $\langle U, V\rangle_{\mathbb{R}} = \mathrm{Re}\,\mathrm{tr}(U^*V)$. Then $\Phi(X, Y) = \frac{1}{2}\|F(X, Y)\|_F^2$. Its linearization at the origin is $\mathcal{J}(X, Y)_h = X\Lambda_\star^h + \Lambda_\star^h Y$.

We first show that $\mathcal{J}$ is injective on $\mathcal{G}$. Suppose $\mathcal{J}(X, Y) = 0$. The $h = 0$ equation gives $X + Y = 0$. The $h = 1$ equation gives $X\Lambda_\star + \Lambda_\star Y = 0$. Writing $\Lambda_\star = \mathrm{diag}(\lambda_1, \dots, \lambda_D)$ and substituting $Y = -X$, the (i, j) entry of the $h = 1$ equation is $X_{ij}(\lambda_j - \lambda_i) = 0$. Because the spectrum is simple,

$\lambda_i \neq \lambda_j$ for $i \neq j$, so all off-diagonal entries of X vanish. For diagonal entries, $Y = -X$ and the gauge condition $\mathrm{diag}(X) = \mathrm{diag}(Y)$ imply $X_{ii} = Y_{ii} = 0$. Hence $X = Y = 0$, proving injectivity.

Because $\mathcal{G}$ is finite-dimensional and $\mathcal{J}$ is injective, its minimum singular value on the unit sphere of $\mathcal{G}$ is positive. Denote this value by $\sigma_{\mathcal{G}} > 0$:

$$\|\mathcal{J}(X,Y)\|_F \geq \sigma_{\mathcal{G}} \|(X,Y)\|_F \qquad ((X,Y) \in \mathcal{G}).$$

The gap dependence is explicit from the first two lags. For $D > 1$, let $\Delta := \min_{i \neq j} |\lambda_i - \lambda_j|$ and $M := \max_i |\lambda_i|$. For each off-diagonal pair (i,j), the two-lag map $(X_{ij}, Y_{ij}) \mapsto (X_{ij}+Y_{ij}, X_{ij}\lambda_j + \lambda_i Y_{ij})$ has determinant $\lambda_i - \lambda_j$ and operator norm at most $(2 + |\lambda_i|^2 + |\lambda_j|^2)^{1/2}$, hence its smallest singular value is at least $|\lambda_i - \lambda_j|/(2 + |\lambda_i|^2 + |\lambda_j|^2)^{1/2}$. For diagonal entries, the gauge condition gives a squared singular value of 2 from the $h = 0$ term. Therefore

$$\sigma_{\mathcal{G}}^2 \geq \min\left\{2, \frac{\Delta^2}{2 + 2M^2}\right\}.$$

For $D = 1$, $q = 1$ but $q_+ = 2$, so the loss still includes the $h = 1$ term and the diagonal bound combined with the $h = 1$ identity gives $\sigma_{\mathcal{G}}^2 \geq 2$.

The residual decomposes as $F_h(X,Y) = \mathcal{J}(X,Y)_h + X\Lambda_\star^h Y$. Therefore $DF(X,Y) = \mathcal{J} + \mathcal{R}(X,Y)$, where

$$\|\mathcal{R}(X,Y)\|_{\mathrm{op}} \leq c_{\Lambda,q_+} \|(X,Y)\|_F, \qquad c_{\Lambda,q_+} := \sqrt{2 \sum_{h=0}^{q_+-1} \|\Lambda_\star^h\|_{\mathrm{op}}^2}.$$

Indeed, for each lag h the derivative of the quadratic term $(X,Y) \mapsto X\Lambda_\star^h Y$ in direction $(\Delta X, \Delta Y)$ is

$$\Delta X \Lambda_\star^h Y + X \Lambda_\star^h \Delta Y,$$

whose Frobenius norm is bounded by $\sqrt{2}\|\Lambda_\star^h\|_{\mathrm{op}} \|(X,Y)\|_F \|(\Delta X, \Delta Y)\|_F$. Summing over h yields the displayed constant. Choose $r_0 > 0$ small enough that $\|\mathcal{R}(X,Y)\|_{\mathrm{op}} \leq \sigma_{\mathcal{G}}/2$ whenever $\|(X,Y)\|_F \leq r_0$.

Write $z = (X,Y)$ with the Frobenius product norm. Since $F(z) = \mathcal{J}z + O(\|z\|_F^2)$, shrinking r_0 if necessary gives constants $0 < c_1 \leq c_2 < \infty$ such that $c_1\|z\|_F^2 \leq \Phi(z) \leq c_2\|z\|_F^2$ throughout the radius-r_0 ball in $\mathcal{G}$. The projected gradient has expansion $\mathrm{grad}_{\mathcal{G}}\, \Phi(z) = \mathcal{J}^*\mathcal{J}z + O(\|z\|_F^2)$. Since $\mathcal{J}^*\mathcal{J} \succeq \sigma_{\mathcal{G}}^2 I$ on $\mathcal{G}$, another shrinkage of r_0 gives

$$\langle \mathrm{grad}_{\mathcal{G}}\, \Phi(z), z \rangle \geq \frac{\sigma_{\mathcal{G}}^2}{2} \|z\|_F^2.$$

By Cauchy–Schwarz, $\|\,\mathrm{grad}_{\mathcal{G}}\, \Phi(z)\|_F \geq \sigma_{\mathcal{G}}^2 \|z\|_F/2$. Combining this with $\Phi(z) \leq c_2\|z\|_F^2$ gives a local Polyak–Lojasiewicz inequality $\|\,\mathrm{grad}_{\mathcal{G}}\, \Phi(z)\|_F^2 \geq \mu\, \Phi(z)$ for some $\mu > 0$ depending on $\sigma_{\mathcal{G}}$, $\Lambda_\star$, and q_+.

Along projected gradient flow,

$$\frac{d}{dt}\Phi(X(t),Y(t)) = -\|\,\mathrm{grad}_{\mathcal{G}}\, \Phi(X(t),Y(t))\|_F^2 \leq -\mu\Phi(X(t),Y(t)),$$

as long as the trajectory remains in the radius-r_0 neighborhood. Taking the initial radius smaller if needed makes the initial sublevel set contained in that neighborhood; since Φ decreases along the flow, the trajectory remains there for all time. Gronwall's inequality gives $\Phi(X(t),Y(t)) \leq e^{-\mu t}\Phi(X(0),Y(0))$.

The lower quadratic bound $c_1\|z(t)\|_F^2 \leq \Phi(z(t))$ implies $z(t) \to 0$ in the gauge slice. Thus $C(t) \to S$ and $E(t) \to S^{-1}$ in spectral coordinates. This limiting point has zero multi-lag loss and satisfies the hypotheses of Theorem 3, so the limiting encoder and decoder are factor-aligned via the spectral projectors. $\square$

A.6 Proof of Proposition 1

Proof. By the factored transition and factored predictor assumptions, $T(s,a) = (T_1(s^{(1)},a),\dots,T_K(s^{(K)},a))$ and $h(s,a) = (h_1(s^{(1)},a),\dots,h_K(s^{(K)},a))$. The orthogonal squared-loss decomposition in Assumption 4 gives, for every test point (s,a),

$$\|h(s,a) - T(s,a)\|_2^2 = \sum_{k=1}^{K} \|h_k(s^{(k)},a) - T_k(s^{(k)},a)\|_2^2.$$

If (s,a) lies in the joint test support whose factor marginals are Q_k, then each $(s^{(k)},a)$ lies in $\operatorname{supp} Q_k$. By the support-uniform assumption, $\|h_k(s^{(k)},a) - T_k(s^{(k)},a)\|_2^2 \le \epsilon_k^2$ for every k. Hence pointwise on the support of Q, $\|h(s,a) - T(s,a)\|_2^2 \le \sum_{k=1}^{K} \epsilon_k^2$. Taking expectation under Q preserves the inequality:

$$\mathbb{E}_Q\|h(s,a) - T(s,a)\|_2^2 \le \mathbb{E}_Q \sum_{k=1}^{K} \epsilon_k^2 = \sum_{k=1}^{K} \epsilon_k^2.$$

This proves the claim. □

A.7 Proof of Proposition 2

Proof. Let $z_h^\star := \mathcal{E}(s_{t+h})$ be the encoded ground-truth trajectory and let $e_h := \hat{z}_h - z_h^\star$ be the latent rollout error. The rollout is initialized from the ground-truth encoding, so $e_0 = 0$.

For one step, $e_{h+1} = f_\theta(\hat{z}_h, a_{t+h}) - z_{h+1}^\star$. Add and subtract $f_\theta(z_h^\star, a_{t+h})$:

$$e_{h+1} = \big(f_\theta(\hat{z}_h, a_{t+h}) - f_\theta(z_h^\star, a_{t+h})\big) + \big(f_\theta(z_h^\star, a_{t+h}) - z_{h+1}^\star\big).$$

By ρ-contractivity and the definition of one-step latent error ϵ, $\|e_{h+1}\|_2 \le \rho\|e_h\|_2 + \epsilon$. Unrolling from $e_0 = 0$ gives

$$\|e_H\|_2 \le \epsilon \sum_{j=0}^{H-1} \rho^j = \epsilon \frac{1-\rho^H}{1-\rho}.$$

Now add and subtract $\mathcal{D}(z_H^\star) = \mathcal{D}(\mathcal{E}(s_{t+H}))$:

$$\|\mathcal{D}(\hat{z}_H) - s_{t+H}\|_2 \le \|\mathcal{D}(\hat{z}_H) - \mathcal{D}(z_H^\star)\|_2 + \|\mathcal{D}(\mathcal{E}(s_{t+H})) - s_{t+H}\|_2.$$

The decoder is L_D-Lipschitz, so the first term is at most $L_D\|e_H\|_2$. The reconstruction-error assumption bounds the second term by α. Therefore

$$\|\mathcal{D}(\hat{z}_H) - s_{t+H}\|_2 \le L_D \epsilon \frac{1-\rho^H}{1-\rho} + \alpha.$$

This is the desired envelope. □

A.8 Proof of Proposition 3

Proof. Let T^π and $\hat{T}^\pi$ denote the Bellman operators under the true and learned models for the fixed policy π. For any state s and action $a = \pi(s)$ on the relevant support, the reward error satisfies $|r(s,a) - \hat{r}(s,a)| \le \eta_r$. The transition error satisfies $d_{\text{task}}(T(s,a), \hat{T}(s,a)) \le \eta_T$. Since the true value function V^π is L_V-Lipschitz in d_{task}, $|V^\pi(T(s,a)) - V^\pi(\hat{T}(s,a))| \le L_V\eta_T$. Thus the one-step Bellman residual of evaluating the learned model against the true value function is bounded by

$$|(T^\pi V^\pi)(s) - (\hat{T}^\pi V^\pi)(s)| \le \eta_r + \gamma L_V \eta_T.$$

The Bellman operator for a fixed policy is a γ-contraction in sup norm, so

$$\|V^\pi - \hat{V}^\pi\|_\infty = \|T^\pi V^\pi - \hat{T}^\pi \hat{V}^\pi\|_\infty \le \|T^\pi V^\pi - \hat{T}^\pi V^\pi\|_\infty + \|\hat{T}^\pi V^\pi - \hat{T}^\pi \hat{V}^\pi\|_\infty.$$

The first term is bounded by $\eta_r + \gamma L_V \eta_T$, while the second is at most $\gamma \|V^\pi - \hat{V}^\pi\|_\infty$ by γ-contraction of $\hat{T}^\pi$. Therefore

$$\|V^\pi - \hat{V}^\pi\|_\infty \le \eta_r + \gamma L_V \eta_T + \gamma \|V^\pi - \hat{V}^\pi\|_\infty,$$

which rearranges to

$$\|V^\pi - \hat{V}^\pi\|_\infty \le \frac{\eta_r + \gamma L_V \eta_T}{1-\gamma}.$$

The performance difference $|J(\pi) - \hat{J}(\pi)|$ is bounded by the same sup-norm value difference at the initial distribution, so

$$|J(\pi) - \hat{J}(\pi)| \le \frac{\eta_r + \gamma L_V \eta_T}{1-\gamma}.$$

This proves the claim. □

B Reproducibility Details for Numerical Illustrations

The generated-system illustrations use fixed random seeds and fully specified finite-dimensional systems. The same data-generation protocol is used for all reported quantities: sample the system or trajectory windows, estimate the theorem-specific operator when required, and report the corresponding singular values, memory energies, projector errors, rollout envelopes, or value residuals. For Theorem 4, the deterministic row is a contour-separation illustration with a conservative analytic resolvent bound.

Table 2: Reproducibility details for generated-system illustrations.

Result	Seed	Construction
Theorem 1	123	$D = 6$, action dimension 2, rank 3, horizon $H = 4$, $n = 8000$, Gaussian actions, least-squares $\widehat{\mathcal{A}}_H, \widehat{\mathcal{B}}_H$
Theorem 2	456	$n = 512$, sequence length 40, input dimension 5, $m = 8$, 400 Rademacher sign draws; fast poles in $[0.15, 0.55]$, slow poles in $[0.72, 0.96]$
Theorem 3	789	Three disjoint factor spectra $[0.12, 0.28]$, $[0.56, 0.73]$, $[1.11, 1.34]$
Theorem 4	deterministic	4×4 diagonalizable example with two separated factor contours, perturbed learned poles, one-step similarity defect $\|\Delta\|_{\mathrm{op}} = 3.83 \times 10^{-2}$, max projector error 3.00×10^{-2}, with the resolvent bound reported as a conservative analytic certificate
Theorem 5	20260427	$\Lambda_\star = \mathrm{diag}(0.2, 0.55, 0.9)$, local projected gradient descent in the gauge slice, initial loss 6.26×10^{-3}, final loss 1.86×10^{-12}
Proposition 1	321	Generated factored errors with $\epsilon = (0.08, 0.12, 0.05)$ and 4096 samples
Proposition 2	654	$\rho = 0.62$, $\epsilon = 0.04$, $\alpha = 0.015$, $L_D = 1.8$, horizons $1, 3, 5, 10, 25$
Proposition 3	deterministic	$\gamma = 0.95$, $\eta_r = 0.02$, $\eta_T = 0.05$, $L_V = 3.0$

C Saved Visual Observations

The following figures and tables are saved motivating observations. They illustrate the kinds of architecture-ranking inversions and latent-dimension effects that Propositions 2 and 3 are designed to interpret once the missing quantities are estimated. The compared architectures are RSSM-GRU [Hafner et al., 2025a], Transformer-STORM [Zhang et al., 2023], and SSM-Mamba [Gu and Dao, 2024], evaluated on DMControl tasks [Tassa et al., 2018]. These observations document differences between evaluation metrics, but the checkpoint-level quantities $\alpha, \epsilon, \rho, L_D, \eta_r, \eta_T$ were not estimated. The tables therefore motivate Propositions 2 and 3, rather than validate their bounds or explain the observed rankings. In particular, selective Mamba is outside the fixed-spectrum setting of Theorem 2.

C.1 Architecture-Level Visual Prediction

Figure 1 summarizes pixel-level prediction errors across several DMControl tasks. The comparison is not intended as a benchmark of the three architectures, but as a motivating observation: the architecture with the lowest pixel MSE depends strongly on the task. For example, SSM-Mamba has the lowest visual MSE on several tasks, while other architectures are closer or better on others. This supports the paper's broader point that world-model quality should not be treated as a single architecture-level property independent of the task, observation map, and downstream metric.

This observation is directly connected to Proposition 3. Pixel MSE is an observation-space metric, whereas control depends on reward error and task-relevant transition error in the value metric. Thus, even when one architecture gives better visual prediction, it need not imply better policy evaluation or control performance unless the visual metric controls the quantities η_r and η_T appearing in Proposition 3.

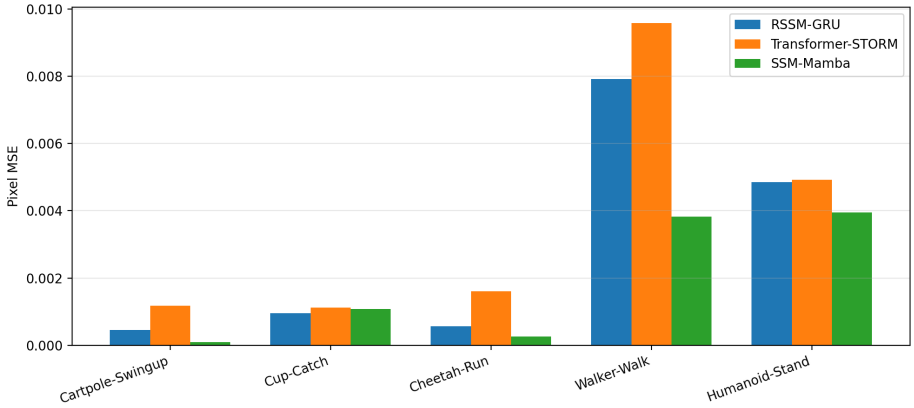


Figure 1: Architecture comparison across DMControl tasks. The lowest pixel-MSE model varies by task, motivating the distinction between observation-space prediction quality and control-relevant prediction in Proposition 3.

C.2 Cheetah Rollout Profile

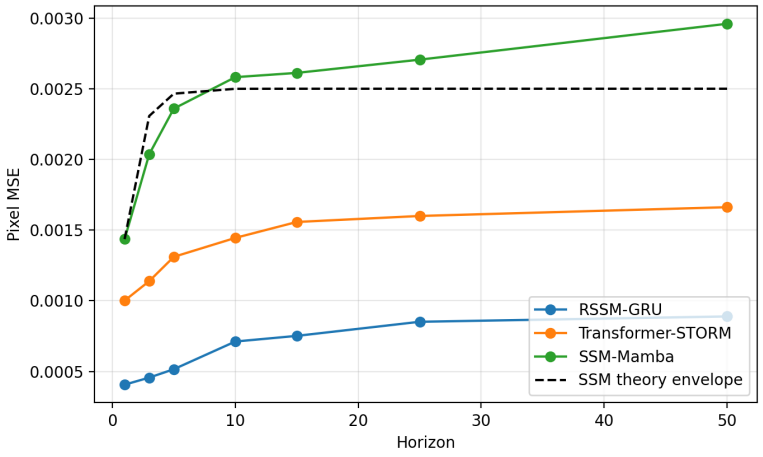


Figure 2: Cheetah rollout profile. RSSM-GRU has the lowest rollout MSE at every evaluated horizon. The displayed SSM envelope is illustrative; it is not a certified bound from Proposition 2, since the required checkpoint-level constants were not estimated.

Table 3: Cheetah rollout MSE profile.

Horizon	RSSM-GRU	Transformer-STORM	SSM-Mamba
1	**0.000406**	0.001001	0.001437
5	**0.000515**	0.001311	0.002361
10	**0.000711**	0.001445	0.002583
25	**0.000851**	0.001601	0.002707
50	**0.000889**	0.001663	0.002961

RSSM-GRU has the lowest rollout MSE at every horizon. Proposition 2 identifies the missing explanatory envelope $\alpha_m + L_{D,m}\epsilon_m(1-\rho_m^H)/(1-\rho_m)$, which depends on reconstruction error, decoder geometry, one-step latent error, and contraction.

C.3 Cheetah Visual Latent Sweep

Figure 3 records a latent-dimension sweep for the Cheetah visual prediction setting. The trend is not monotone across the two visual metrics: larger latent dimension can reduce one metric while worsening another. This suggests that increasing latent capacity alone is not a reliable proxy for improved world-model quality, even before considering downstream control.The sweep is consistent with the spectral viewpoint developed in the paper. The relevant bottleneck is not merely the nominal latent dimension, but whether the learned representation preserves the directions that matter for the prediction horizon, observation decoder, and task metric. In the linear theory, Theorem 1 makes this precise through the shared-horizon Krylov spectrum, while Propositions 2 and 3 explain why observation-space and control-relevant rankings can diverge.

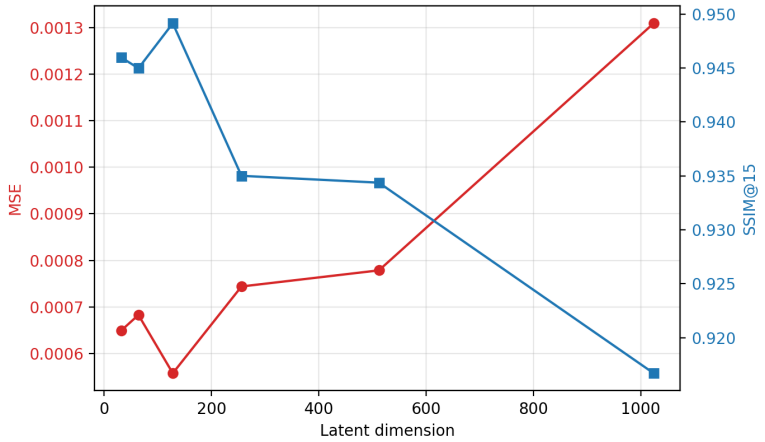


Figure 3: Cheetah visual latent-dimension sweep. Increasing latent dimension does not monotonically improve both MSE and SSIM, motivating the distinction between latent capacity, observation-space metrics, and downstream task-relevant quality.

C.4 Prediction-Control Mismatch

Table 4: Visual prediction and control-return mismatch on DMControl tasks.

Task	Model	Visual MSE	SSIM@15	Episode return
Walker-Walk	RSSM-GRU	0.0079	0.7468	**41.950**
Walker-Walk	Transformer-STORM	0.0096	0.7151	25.173
Walker-Walk	SSM-Mamba	**0.0038**	**0.8238**	27.916
Humanoid-Stand	RSSM-GRU	0.0049	0.7536	6.076
Humanoid-Stand	Transformer-STORM	0.0049	0.7449	**8.614**
Humanoid-Stand	SSM-Mamba	**0.0039**	**0.7793**	6.876

The ranking mismatch is direct: the best visual predictor is not the best controller. Proposition 3 identifies the missing explanatory quantities as reward error η_r and task-relevant transition error η_T in the value metric.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's three main theoretical contributions—the Krylov bottleneck, stable diagonal-SSM memory-energy bound, and spectral factor-identifiability results—and restrict the claims to linear settings. The scope and limitations are further clarified in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5.2 discusses the main limitations, including linear dynamics, exogenous-action assumptions, fixed diagonal spectra, square encoder/decoder maps, local optimization guarantees, and restrictive contractive/factored/Lipschitz assumptions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are centralized in Section 3.1 and referenced in each theorem or proposition. Proof sketches are given in the main text, and full proofs for Theorems 1–5 and Propositions 1–3 are provided in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The numerical illustrations are generated finite-dimensional systems, and Appendix B reports the random seeds, dimensions, horizons, sample sizes, spectra, and deterministic settings needed to reproduce the reported quantities.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper provides detailed reproducibility information for generated systems in Appendix B, but the current submission does not include an open code release. The experiments use synthetic/generated systems rather than a newly introduced dataset.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 summarizes the numerical settings, and Appendix B gives the seeds, system dimensions, sample sizes, pole ranges, horizons, and optimization or estimation procedures used for each generated-system illustration.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The numerical results are intended as generated-system illustrations of deterministic theorem-predicted quantities rather than large-scale empirical comparisons, and the current paper reports single-seed or deterministic outcomes without error bars. This limitation is partly mitigated by fully specified seeds and constructions in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper specifies the generated systems and random seeds, but it does not currently report CPU/GPU type, memory usage, wall-clock runtime, or total compute. The experiments are small finite-dimensional generated-system illustrations, but compute details should be added for completeness.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and uses generated systems plus saved aggregate visual observations. It does not involve human subjects, private data, deployed systems, or high-risk model releases.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper is primarily foundational theory for linear world models and does not currently include a dedicated broader-impact discussion. The likely impacts are indirect, through improved understanding and evaluation of model-based reinforcement learning systems.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pretrained models, image generators, scraped datasets, or other high-risk assets requiring safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites existing works and environments such as DMControl and prior world-model architectures, but the current draft does not explicitly list asset licenses or terms of use. If code or data from these assets are used directly, their licenses should be added.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset, benchmark, model checkpoint, or software package. The generated numerical examples are described for reproducibility in Appendix B.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or human-subject experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject research, so IRB or equivalent approval is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core method, theory, proofs, or numerical experiments. Any LLM use for writing, editing, or formatting does not affect the scientific methodology or originality of the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.