STRIDE: STATE-TRANSITION REPRESENTATION VIA INCREMENT DYNAMICS AND EVOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce STRIDE (State-Transition Representation via Increment Dynamics and Evolution), which defines states through derivative fingerprints and learns local functions for state transitions (qpairs), recasting continuous forecasting as transition prediction. Trailing convolution windows estimate local joint-transition frequencies, whose lagged differences form a high-dimensional increment trajectory. Proper orthogonal decomposition (POD) gives coordinate paths, jointly forecast by sparse dynamics with memory. Recombining their forecasts along the POD directions gives future increments; history-anchored inverse differencing recovers frequency windows. Learned causal deconvolution gives joint-transition distributions; sampled state paths select local functions to generate continuous forecasts. Three-seed experiments compare STRIDE against fourteen baselines across nine benchmark families. Five independent Markov and hidden-state baselines cover all 230 tasks. Against these comparators, STRIDE wins 139–158 of 220 complete whole-horizon pairs and achieves system-weighted late-Energy win fractions of 73.6–78.5% on 190 multi-step pairs. Against DLinear, its system-weighted late-Energy win fraction is 77.3% across 72 paired multi-step tasks. Matched controls examine the contribution of the intermediate representation. On 64 independently initialized Aizawa trajectories, STRIDE lowers late Energy by 51.5%, 55.6%, 52.0%, and 24.2% against No-D, same-q Markov, history-direct, and duration controls; all four prespecified comparisons pass Holm correction. These results connect transition-statistic prediction to continuous probabilistic forecasting, with substantial long-horizon gains in the matched Aizawa study.

1 INTRODUCTION

STRIDE models complex dynamics through state transitions and local continuous functions. A *derivative fingerprint* combines first and second derivatives with higher derivatives or observation levels. Clustering defines states; a local function models continuous evolution during each ordered state transition (qpair). Convolution and lagged differencing trace local transition-frequency changes as a high-dimensional trajectory for decomposition and coupled forecasting. From the predicted trajectory, inverse differencing recovers future frequency windows; causal deconvolution yields joint-transition distributions. Sampled state pairs then select local functions to generate future observations.

Specifically, local polynomials fitted to past observations give the derivative estimates. Training fingerprints define states q; validation selects the fingerprint and state count. To predict their transitions, STRIDE models how their local joint frequencies change. A trailing convolution of one-hot state-pair tables estimates local frequencies. Lagged differences, vectorized over all state pairs, form the increment trajectory.

Proper orthogonal decomposition (POD) projects this increment trajectory onto fixed directions, giving one subtrajectory per direction (Figure 1). A sparse coupled model simultaneously computes each subtrajectory's next point from preceding coordinates of *all* retained subtrajectories, their finite histories, and accumulated memories. Repeated updates extend the subtrajectories; recombination gives the future transition-frequency increment trajectory. Inverse differencing combines each predicted increment with its observed or previously reconstructed lagged frequency table to recover the next window. Learned causal deconvolution then uses these windows and historical boundary tran-

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

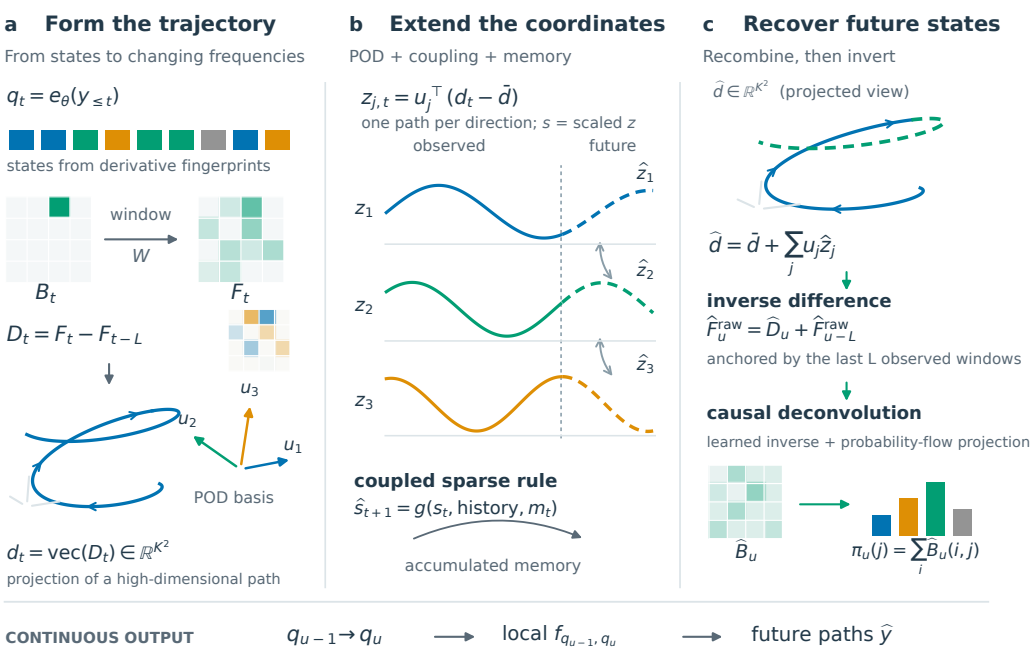


Figure 1: **Form a trajectory, extend its coordinates, recover future states.** (a) Derivative fingerprints define states; window convolution and differencing produce transition-frequency increments D. The path in $\mathbb{R}^{K^2}$ and three POD directions are shown in projection. (b) Each direction gives a scalar subtrajectory. All next coordinates are jointly advanced from the current coordinate vector and its histories; arrows illustrate coupling and dashed curves predictions. (c) Recombination gives future increments $\widehat{d}$; observed historical boundaries initialize inverse differencing and causal deconvolution to recover windows $\widehat{F}^{\text{raw}}$ and flow-consistent tables $\widehat{B}$. Their marginals give state distributions; sampled pairs select local continuous functions. Four states and three directions are illustrative.

sitions to recover joint-transition distributions. Their marginals give future-state distributions, and row conditioning gives the transition probabilities used to sample state paths. Each sampled state pair selects its local function to extend the continuous trajectory.

The state encoder, coordinate dynamics, inverse readout, and local output functions are fitted in stages. This modular construction allows the transition forecaster to be replaced while retaining the same state partition and continuous predictors. Observed anchors carry the frequency levels, while the sparse coordinate model advances their increments.

Our evaluation separates the performance of the complete forecasting system from the contribution of its intermediate representation. Three-seed comparisons against fourteen baselines span nine benchmark families. Markov–Gaussian, first- and third-order Gaussian HMMs, ARHMM, and continuous Gaussian Markov each cover the full 230-task inventory and yield 220 complete pairs with STRIDE. These independent forecasters learn their own dynamics and observation models. Matched controls examine trajectory dynamics, transition decoding, inversion, and continuous output. Finally, a separately locked Aizawa study tests long-horizon forecasts on 64 independently initialized trajectories. Our forecasting chain constructs transition-frequency increment trajectories from observations, decomposes and predicts them, and recovers future states and continuous paths. The evaluation separates system-level performance from conditional module contributions.

Related work. POD and sparse identification provide coordinates and parsimonious dynamical models (Holmes et al., 2012; Brunton et al., 2016). Delay features support compact nonlinear forecasting (Gauthier et al., 2021), while data-driven Mori–Zwanzig models separate current action from memory (Lin et al., 2021). SINDy autoencoders jointly learn coordinates and sparse equations

(Champion et al., 2019); SINDy-SHRED couples recurrent sensing and decoding to sparse latent dynamics (Gao et al., 2026). STRIDE fixes an empirical transition-statistics lift, uses POD coordinates, and places its learned inverse readout after the sparse dynamics. It uses no recurrent SHRED encoder or joint end-to-end training.

Markov and hidden Markov models describe future state uncertainty through transition probabilities and, where appropriate, observation distributions (Rabiner, 1989). STRIDE predicts future transition probabilities by forecasting an evolving trajectory of empirical frequency increments. In observation space, DeepAR learns recurrent probabilistic forecasts, DLinear uses linear trend–seasonal decomposition, TimePrism models weighted forecast scenarios, and context parroting reuses continuations of similar historical patterns (Salinas et al., 2020; Zeng et al., 2023; Dai et al., 2026; Zhang & Gilpin, 2026). Koopman forecasters model transformed coordinates, including adaptation to distribution shifts and probabilistic prediction (Wang et al., 2023; Zheng et al., 2025). STRIDE instead models explicit joint-transition statistics and reconstructs discrete paths through a probability-constrained readout. The empirical question is when coupled forecasting of changing transition statistics and causal inversion improve on direct observation-space forecasts and simpler state-space dynamics.

2 From state transitions to a forecastable trajectory and back

Let $y_t \in \mathbb{R}^d$ denote observations and $\mathcal{H}_t$ the information available at forecast origin t. The objective is a distribution over $y_{t+1:t+H}$, evaluated as an ensemble of complete paths. Every fitted transformation below uses training data; validation selects configurations and checkpoints. Future observations appear only as training targets or evaluation references.

The state partition supports two linked operations: local functions model continuous evolution conditional on a state pair, while forecasts of transition statistics provide the probabilities used to select those functions.

2.1 Define states and form the transition-frequency trajectory

The state q_t is a cluster label of a *derivative fingerprint*. A local polynomial fitted to the last $J = c_q$ observations gives derivative estimates $\widehat{y}_t^{(k)}$ at their final endpoint. The candidate fingerprints are $(\widehat{y}_t^{(1)}, \ldots, \widehat{y}_t^{(4)})$ or $(y_t, \widehat{y}_t^{(1)}, \widehat{y}_t^{(2)})$, using the dataset's observation adapter. Training medians and interquartile ranges normalize the blocks, successive blocks receive weights $1, 0.45, 0.45^2, \ldots$, and training-only K-means defines K states. Nearest-center assignment then gives $q_t = e_\theta(y_{t-J+1:t})$. Training and forecast-origin encoding use derivatives estimated from observed prefixes. Stencils, short-context restrictions, and high-dimensional adapters are specified in Appendix A. Define the observed joint-transition table and its windowed frequency by

$$B_t(i,j) = \mathbf{1}\{q_{t-1} = i, q_t = j\}, \qquad F_t = \frac{1}{W}\sum_{k=0}^{W-1} B_{t-k}, \qquad D_t = F_t - F_{t-L}. \tag{1}$$

The causal uniform convolution estimates local joint frequencies; lag differencing measures their change. Here W counts transitions, requiring $W+1$ states, and L is a difference lag. D_t is a finite increment, without division by L or physical time. F_t is an empirical joint-frequency table: its entries sum to one, while those of D_t sum to zero. Flattening D_t into $d_t = \mathrm{vec}(D_t) \in \mathbb{R}^{K^2}$ gives the modeled trajectory.

2.2 Resolve the trajectory into directional coordinate paths

POD is fitted to centered training vectors d_t. Let $U_r = [u_1, \ldots, u_r] \in \mathbb{R}^{K^2 \times r}$ contain its leading orthonormal directions and $\bar{d}$ the training mean. Projection gives scalar paths $z_{j,t} = u_j^\top (d_t - \bar{d})$: each tracks the strength of one pattern of changing transition mass. The retained trajectory is $\bar{d} + \sum_{j=1}^r u_j z_{j,t}$. For fitting, we standardize these coordinates using training locations b and scales $a \in \mathbb{R}^r_+$:

$$s_t = \mathrm{diag}(a)^{-1}[U_r^\top (d_t - \bar{d}) - b], \qquad \widehat{d}_t = \bar{d} + U_r[b + \mathrm{diag}(a)\widehat{s}_t]. \tag{2}$$

The resulting r scalar paths share a time axis and jointly reconstruct the retained high-dimensional trajectory. Since training changes are zero-sum, the nonzero POD rank is at most $K^2 - 1$. Ranks retaining 90%, 95%, or 99% of the covariance variance of a training sample form dataset-specific candidates. For each candidate, the basis is refitted from all eligible training changes; validation D forecasts over $\min(H, 16)$ steps select rank and history length, followed by full-chain validation Energy. The retained directions are the leading training-covariance eigenvectors. Exchange selects rank one and StocBench rank 15; the actual retained variance is reported separately from the screening threshold (Appendix B).

2.3 Advance the coordinate paths through coupling and memory

Following sparse identification of nonlinear dynamics (SINDy; Brunton et al., 2016), we fit a discrete next-coordinate map with an explicit feature library:

$$\begin{aligned} \widehat{s}_{t+1} &= (\Theta_t \Xi)^\top + \varepsilon_{t+1}, \\ \Theta_t &= \big[1, s_t^\top, \sin(\pi s_t)^\top, \cos(\pi s_t)^\top, \\ &\qquad \{(s_{t-\ell} - s_t)^\top\}_{\ell=1}^{p-1}, \{(m_t^{(\rho,j)} - s_t)^\top\}_{\rho\in\mathcal{R},\ j\in\{1,2\}}\big]. \end{aligned} \tag{3}$$

Here $\mathcal{R} = \{0.6, 0.9\}$. For coordinate j, the prediction is $\widehat{s}_{j,t+1} = \Theta_t \Xi_{:,j} + \varepsilon_{j,t+1}$. Its feature row contains the current, lagged, and accumulated-memory coordinates of *all* retained directions. With three paths, for example, the next coordinate of path 1 can depend on paths 1, 2, and 3 and their histories; the other outputs are fitted in the same way. Coupling therefore occurs through shared cross-direction inputs, while the sine and cosine functions act elementwise. The implemented library has no cross-coordinate polynomial products. Memory evolves as

$$m_t^{(\rho,1)} = \rho m_{t-1}^{(\rho,1)} + (1-\rho) s_t, \qquad m_t^{(\rho,2)} = \rho m_{t-1}^{(\rho,2)} + (1-\rho) m_t^{(\rho,1)}. \tag{4}$$

The second stage uses the newly updated first stage. These fixed recursions retain past inputs through a compact state; validation chooses p, and sparse fitting chooses feature coefficients. The decay constants themselves are not learned.

We normalize feature columns by training root-mean-square values, fit ridge regression, and perform three sequential threshold-and-refit passes. Training residuals supply the innovation covariance. At each forecast step, Equation (3) evaluates all next coordinates from the same current histories and memories. Their histories and memories are then updated together for the next step. Equation (2) recombines the resulting paths through the fixed POD directions to produce $\widehat{d}_{t+h}$ and its reshaped transition-frequency change table $\widehat{D}_{t+h}$. The reusable components are these coordinate rules and memory updates.

2.4 Invert the trajectory into windows and future state distributions

Inverse differencing recovers future windows. For each recombined D trajectory, the inverse of the lag difference uses the last L observed frequency tables as boundary conditions:

$$\widehat{F}_{t+h}^{\mathrm{raw}} = \widehat{D}_{t+h} + \begin{cases} F_{t+h-L}, & h \le L, \\ \widehat{F}_{t+h-L}^{\mathrm{raw}}, & h > L. \end{cases} \tag{5}$$

Causal deconvolution recovers the current transition table. With exact frequencies and boundary transitions, the convolution has the recursive inverse $B_u = WF_u - \sum_{k=1}^{W-1} B_{u-k}$. Applied to predictions, this recursion carries previous reconstruction errors into subsequent steps. STRIDE therefore keeps Equation (5) fixed and learns the following window-to-transition operation, using forecast windows and observed boundary information.

At each future step, a small causal temporal convolutional network receives the ensemble mean and standard deviation of $\widehat{F}^{\mathrm{raw}}$, the observed $W - 1$ boundary transitions, L observed frequency anchors, the last observed state q_t, and horizon features. It has width 16 and three residual blocks with dilations 1, 2, and 4. It is trained on full-horizon joint-transition Brier loss, with all upstream components fixed. At inference, these forecast summaries and historical boundaries determine the transition-table sequence used for path sampling.

Probability flow gives the next state distribution. Let $A_h(i,:)$ be a row after global softmax over K^2 entries, and let π_{h-1} be the previous state's marginal, initialized by $\pi_0 = \mathrm{onehot}(q_t)$. The readout projects each row onto a nonnegative simplex with mass $\pi_{h-1}(i)$:

$$\widehat{B}_h(i,:) = \mathrm{Proj}_{\Delta(\pi_{h-1}(i))} A_h(i,:), \qquad \pi_h(j) = \sum_i \widehat{B}_h(i,j). \tag{6}$$

This enforces nonnegativity, total mass one, and agreement between adjacent source and destination marginals. For positive source mass, set $T_h(i,j) = \widehat{B}_h(i,j)/\pi_{h-1}(i)$; zero-mass rows can be assigned any probability vector because they are unreachable. Sampling from these matrices produces a history-conditioned, time-inhomogeneous Markov path. Rich history enters the predicted matrix sequence; the sampler is first-order conditional on that sequence.

Proposition 1 (Probability-flow consistency)**.** *If Equation (6) holds and $q_t \sim \pi_0$, then sampling with $T_1, \ldots, T_H$ gives* $\Pr(q_{t+h-1} = i, q_{t+h} = j \mid \mathcal{H}_t, \widehat{B}_{1:H}) = \widehat{B}_h(i,j)$.

Induction on the propagated marginals gives the stated joint-transition probabilities at every forecast step; Appendix A provides the proof.

2.5 Route local continuous functions with the predicted states

For each sampled transition $q_{u-1} = i, q_u = j$, a local model predicts a residual around constant-velocity extrapolation. In training-standardized observation coordinates x, its form is

$$\widehat{x}_u = 2\widehat{x}_{u-1} - \widehat{x}_{u-2} + (\beta_0 + \beta_{ij})^\top \phi(\widehat{x}_{u-1}, \widehat{x}_{u-1} - \widehat{x}_{u-2}) + \eta_{u,ij}. \tag{7}$$

The polynomial library has validation-selected degree 1, 2, or 3. A shared ridge fit β_0 is followed by transition-specific residual fits β_{ij}; training residuals determine Gaussian innovation covariances. Equation (7) recursively generates continuous forecasts from sampled state pairs and preceding observations. Appendix A gives feature scaling, shrinkage, and dimensionality caps.

Training proceeds in stages: fit the encoder and POD on training data; select sparse dynamics using validation forecasts; train the inverse readout against observed training transitions; fit and validate the local output model. Validation selects the final configuration using the prescribed forecast scores. This staged fitting permits controlled replacement of a single module while keeping the state partition and local functions fixed.

3 Experiments

Our evaluation separates three questions: the performance of the complete forecaster, the contribution of predicted transition statistics with the encoder and local predictors held fixed, and the role of individual stages under matched controls. Independent baselines assess complete forecasts; the stage-specific tests and long-horizon study examine the intermediate representation.

3.1 Protocol and comparison scope

Tasks and independent baselines. The benchmark spans nine families: dysts, MDBench, FinStressTS, ProbTS, MDGen-pilot, StocBench, GIFT-Eval, fev, and BOOM. Of 240 planned tasks, 230 reach evaluation and 220 have fully scorable STRIDE forecasts across three fitting seeds. Eligibility, failures, and comparator-specific coverage are recorded in Appendices B–C. Fourteen baselines cover clustered Markov–Gaussian, continuous Gaussian Markov, Gaussian HMM, third-order HMM, ARHMM, direct SINDy, seasonal naive, Holt, VAR, NVAR, context parroting, DLinear, DeepAR, and TimePrism. These are complete observation-space forecasters with their own predictive outputs; the state models use their own states or continuous dynamics and emissions. They do not use STRIDE's encoder, inverse decoder, or local functions. The same-q Markov *module control* retains STRIDE's encoder and local functions (Appendix B.1).

All comparisons share the prescribed splits, contexts, horizons, test origins, and seeds 1729, 2718, and 3141. Selection uses validation data. Evaluation uses up to 256 origins and 32 paths for probabilistic forecasts; deterministic predictions are point masses. Normalization, clustering, and POD use training data only. The five full-inventory state-space baselines each yield 220/230 complete pairs. Appendix B.1 specifies the selection grids and links the frozen optimization protocol.

Scores and statistical units. The primary score is Euclidean path Energy in training-standardized coordinates (Gneiting & Raftery, 2007; Shahroudi et al., 2024; Ashok et al., 2024).

For $S > 1$ paths and M observed target coordinates,

$$\widehat{\mathrm{ES}} = \frac{1}{S\sqrt{M}} \sum_{s=1}^{S} \|x_s - y\|_2 - \frac{1}{2S(S-1)\sqrt{M}} \sum_{s \neq r} \|x_s - x_r\|_2. \tag{8}$$

For a point mass, the second term is zero. Intermediate endpoints are D MSE, checked-window F MSE, joint-pair Brier, and state Brier; continuous diagnostics add fair CRPS, variograms, and coverage (Ferro, 2014; Scheuerer & Hamill, 2015). Task scores average matched forecast origins within each seed, then the three seeds equally. Family means average task scores within each underlying system/source, then weight these groups equally. Sampling, noise, and horizon variants are not independent datasets. Paired scores retain the full coverage inventory; raw errors are not averaged across unrelated families. Group-bootstrap intervals are descriptive. Stratified analyses reuse fixed test records, whereas the independent Aizawa replication locks fitting records before generating new test trajectories.

3.2 Whole-horizon and late-horizon benchmark performance

Figure 2 compares STRIDE against all fourteen baseline forecasters using whole-path and final-half Energy. Both panels use identical complete paired tasks for each method. The 200 multi-step tasks give 1,903 planned comparisons and 1,777 complete pairs; the 30 single-step tasks remain in the whole-horizon tables. Late Energy jointly scores the subpath $\{\lfloor H/2 \rfloor + 1, \ldots, H\}$ with equal weights on its observed time–coordinate entries. Each task's three-seed mean determines a win indicator. We average these indicators within each underlying system/source, then weight groups equally. Counts are unweighted task wins; their ratio to paired tasks need not equal the plotted fraction.

On 190 common multi-step tasks, whole-/late-horizon wins are 156/146 against Markov–Gaussian, 153/143 against Gaussian HMM, 139/137 against ARHMM, and 139/134 against continuous Gaussian Markov. Their late system-weighted win fractions are 78.5%, 76.4%, 73.6%, and 74.6%, respectively. It wins 44/72 DLinear comparisons at both endpoints (77.7% and 77.3% weighted). Against third-order HMM, the corresponding counts are 145/190 and 140/190 (76.3% and 76.0% weighted). Appendix D reports the complete all-horizon comparisons, family scores, and third-order HMM results.

These results show strong competitiveness across the benchmark. On dysts, STRIDE achieves Energy 0.3838 versus 0.6202 for Markov–Gaussian and 0.6283 for Gaussian HMM on 21 paired conditions from eight systems. It improves over clustered Markov–Gaussian by 18.0% on ProbTS and VAR by 10.0% on GIFT-Eval. On Weather, it has lower Energy than DLinear and TimePrism at all four horizons, by 4.70–11.48% and 14.23–44.48%. It remains competitive against NVAR and TimePrism, with system-weighted late-horizon wins of 51.6% and 52.5%. On MDBench, STRIDE wins 75 of 101 paired conditions against VAR across 66 system groups; its mean loss is concentrated in a small number of systems, examined individually in Appendix D. Appendix D presents the complete family effects, late scores, and error-growth slopes on their comparator-specific cohorts.

3.3 Representation and trajectory evolution

Measured transition activity and window choice. The (q, W, L) configuration determines the measured trajectory. A validation-only window/lag study varies W and L with the encoder, state count, rank, history setting, and local functions fixed (Appendix E.2). Separately, true-data activity describes how often these windows change: $D = 0$ occupies 98.81% of Exchange, 80.67–85.41% of Weather, and 85.43% of Mackey–Glass's distinct evaluated targets. High quiet fractions also occur in training. Weather nevertheless has 95th-percentile $\frac{1}{2}\|D\|_1$ of 0.33–0.40, while Aizawa and StocBench provide frequently changing targets. Activity profiles and observed pair counts appear together in Appendix E.3. Coupling and memory are assessed on active, multidirectional representations; quiet regimes are characterized separately through activity and conditional-score measurements. At the common 80% lead, quiet-stratum single-step Energy is equivalent within $\pm 1\%$ on Exchange and $\pm 3\%$ on Weather $H = 336$ to matched context-only decoding under both block lengths (Table 22), complementing the active-node tests below.

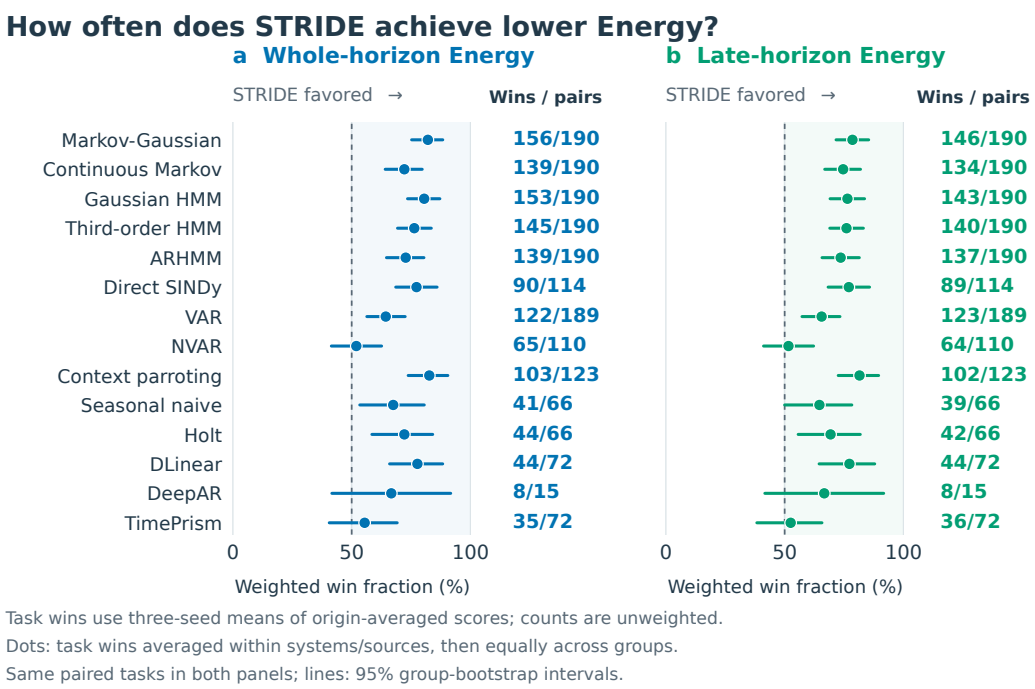


Figure 2: **Whole- and late-horizon wins against independent forecasters.** (a) Whole-path Energy; (b) final-half path Energy, on identical comparator-specific multi-step cohorts. Bold counts give unweighted task wins / paired tasks. Points average task win indicators within each system/source, then equally across groups; lines give descriptive 95% group-bootstrap intervals. The dashed line marks 50%. The five full-inventory state-space baselines each have 190 paired tasks. Table 11 gives the paired/planned inventory.

Coupling and memory in the directional dynamics. Known-operator controls isolate the POD/sparse-dynamics module. They generate $d_t = Vz_t$ as 16-component zero-sum vectors with current-action, two-step, or distributed-memory dynamics. On held-out two-step trajectories, STRIDE's one-step action error is 0.0784 versus 0.1950 without coupling and 0.3049 without memory; autonomous impulse errors are 0.1879, 0.3588, and 0.3620. Distributed-memory action errors are 0.1005, 0.1918, and 0.1545. All three generators and matched controls appear in Figure 8 and Table 20. On observed transitions, removing memory raises D MSE from 0.005016 to 0.079724 on Aizawa and from 0.004817 to 0.007411 on StocBench (Table 15).

Forecasting observed increment trajectories. At the common lead $h = \lceil 0.8H \rceil$, StocBench's active-node D MSE is 20.20% below zero change and 71.52% below persistence; Hospital $H = 12$ improves by 59.29% over persistence. Active nodes satisfy $\frac{1}{2}\|D_{t+h}\|_1 > 10^{-12}$. These comparisons pass the original simultaneous-band and sign-flip criteria (Appendix E.5), directly measuring the ability to forecast changing transition statistics. A matched direct-F control further supports increment modeling: on 64 independently initialized Aizawa trajectories, recovering F from predicted D lowers MSE by 76.05% over leads 513–1024 (Appendix E.6).

3.4 Causal inversion and local continuous output

No-D (context-only) decoding fits a separate decoder with the forecast-frequency mean and standard deviation masked during both training and evaluation. The architecture, upstream predictions, historical inputs, and local functions remain fixed. On Aizawa $H = 64$, using these frequency inputs lowers active-node joint-pair Brier by 89.39% at the common 80% lead, with corrected support over multiple leads (Figure 3).

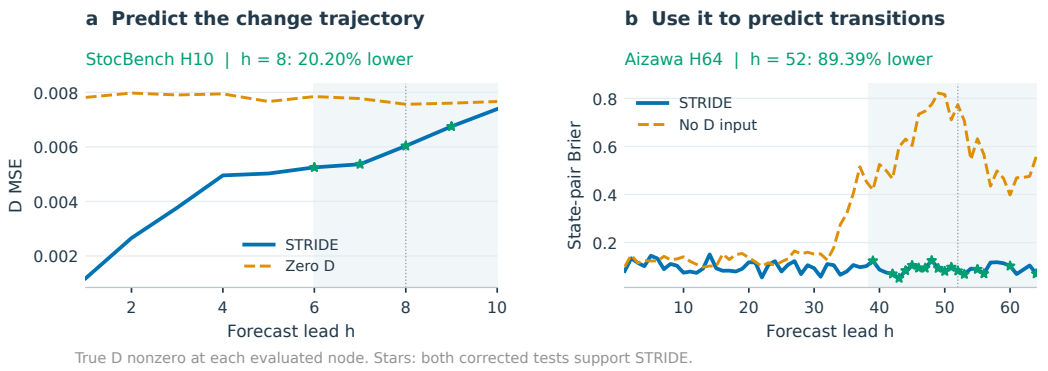


Figure 3: **Increment prediction and its use in transition decoding.** Each lead averages origins with true $D \neq 0$, across three fits. Left: StocBench D prediction versus zero change. Right: Aizawa joint-transition prediction versus matched No-D decoding. Stars require both the simultaneous band and block sign-flip test to pass the original full-family correction. The common 80% lead is marked; shading denotes the analyzed final 40%, not uncertainty.

With the entire predicted D ensemble and inverse-difference recurrence fixed, learned deconvolution also lowers StocBench Energy from 0.730442 for recursive inversion to 0.713792, a 2.28% reduction; the recording-weighted hard-minus-learned interval is $[0.01507, 0.01771]$ (Figure 10).

Oracle-input diagnostics then separate the inverse readout from upstream forecasting: refreshed true history and true D/F give Aizawa argmax accuracies 95.21% at $H = 64$ and 92.02% at $H = 256$. Exact algebraic inversion under these boundaries has maximum table error below 1.6×10^{-14}. These are diagnostic inputs, not forecast baselines.

Local continuous prediction. Teacher-forced local prediction with true state pairs and zero innovation gives Aizawa test RMSE 1.11×10^{-4} versus 1.58×10^{-4} for the shared function. Appendix F groups these tests with the continuous-path distribution scores.

3.5 Matched long-horizon contribution

The final case study tests whether the frequency inputs improve continuous prediction as the horizon grows. At Aizawa $H = 256$, matched long-horizon fits yield late Energy 0.090716 versus 0.221461 for No D, a 59.0% reduction. This original five-block comparison is descriptive. A separate replication locks all fitting choices before generating 64 independently initialized trajectories of the same system. Late Energy is 0.090483 for STRIDE, versus 0.186657 for No-D, 0.203570 for Same-q Markov, 0.188623 for history direct, and 0.119340 for duration. The respective reductions are 51.5%, 55.6%, 52.0%, and 24.2%, as summarized in the abstract. All four prespecified contrasts pass Holm correction ($p = 0.0004$). Appendix G places both studies together with their distinct sampling units, matched fitting protocols, and error-growth curves. The trajectories provide replication within one fixed system and initial-condition distribution.

4 Discussion and conclusion

STRIDE connects states defined by local derivative fingerprints to continuous forecasting through transition-frequency prediction. Coupled coordinate dynamics extend the frequency-increment trajectory; inverse differencing recovers future frequency windows, and a learned causal readout yields transition probabilities. Sampled state pairs select the local functions that generate future observations. The experiments support this intermediate representation in the evaluated settings, with substantial matched long-horizon gains on Aizawa. Quiet-target score equivalence and gains on specified active cases characterize its conditional contributions. The result is a concrete, testable connection between transition-statistic prediction and continuous probabilistic forecasting.

Reproducibility Statement

The original benchmark comparisons use saved predictions or frozen model checkpoints. Appendices A–C specify the implementation, dataset and baseline protocols, and statistical units. Appendices D, E, and F organize the complete-forecast comparisons and stage-specific diagnostics. The source package regenerates tables and figures from supplied evidence and includes the frozen benchmark reference implementation, evaluated-task registry, and environment inventory listed in Appendix H. Raw datasets, fitted archives, and study-specific follow-up scripts remain separate. Appendix G.2 specifies the independent-initial-condition Aizawa study, matched readout budgets, pre-generation fitting lock, and trajectory-level inference. Supplied aggregate records support its reported tables.

Ethics Statement

The study uses existing benchmark data and numerical simulations. It does not introduce human-subject experiments or a deployed decision system. Financial series are used solely for forecasting evaluation. The demonstrated undercoverage and task-specific instability should be considered before using the predictions in consequential applications.

AI Use Statement

Generative AI tools assisted with literature retrieval, figure preparation, and manuscript drafting and editing. The authors take responsibility for the final manuscript, including its references, figures, and scientific claims.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. arXiv preprint arXiv:2410.10393, 2024. URL https://arxiv.org/abs/2410.10393.

Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Nicolas Chapados, and Alexandre Drouin. TACTiS-2: Better, faster, simpler attentional copulas for multivariate time series. In *International Conference on Learning Representations*, pp. 22897–22924, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/63796148c99205adb0fcac069cc714d4-Paper-Conference.pdf.

Amirmohammad Ziaei Bideh, Aleksandra Georgievska, and Jonathan Gryak. MDBench: Benchmarking data-driven methods for model discovery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(24):19746–19754, 2026. doi: 10.1609/aaai.v40i24.39056.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45):22445–22451, 2019. doi: 10.1073/pnas.1906995116.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This time is different: An observability perspective on time series foundation models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50907–50951, 2025. doi: 10.52202/085713-1698.

Xilin Dai, Zhijian Xu, Wanxu Cai, and Qiang Xu. From samples to scenarios: A new paradigm for probabilistic forecasting. In *International Conference on Learning Representations*, pp. 140721–140750, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e44337573fcac83f219c8effa4ebf90d-Paper-Conference.pdf.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. doi: 10.1002/qj.2270.

Mars Liyao Gao, Jan P. Williams, and J. Nathan Kutz. Sparse identification of nonlinear dynamics and Koopman operators with Shallow Recurrent Decoder Networks. *Proceedings of the National Academy of Sciences*, 123(16):e2508144123, 2026. doi: 10.1073/pnas.2508144123.

Daniel J. Gauthier, Erik Bollt, Aaron Griffith, and Wendson A. S. Barbosa. Next generation reservoir computing. *Nature Communications*, 12(1):5564, 2021. doi: 10.1038/s41467-021-25801-2.

William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/ec5decca5ed3d6b8079e2e7e7bacc9f2-Paper-round2.pdf.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Philip Holmes, John L. Lumley, Gahl Berkooz, and Clarence W. Rowley. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press, 2nd edition, 2012. doi: 10.1017/cbo9780511919701.

Bowen Jing, Hannes Stärk, Tommi Jaakkola, and Bonnie Berger. Generative modeling of molecular dynamics trajectories. In *Advances in Neural Information Processing Systems*, volume 37, pp. 40534–40564, 2024. doi: 10.52202/079017-1282.

Yen Ting Lin, Yifeng Tian, Daniel Livescu, and Marian Anghel. Data-driven learning for the Mori–Zwanzig formalism: A generalization of the Koopman learning framework. *SIAM Journal on Applied Dynamical Systems*, 20(4):2558–2601, 2021. doi: 10.1137/21m1401759.

Sebastian Pfister, Benjamin Holzschuh, and Nils Thuerey. StocBench: A benchmark for generative modeling of stochastic dynamics. arXiv preprint arXiv:2608.22309, 2026. URL https://arxiv.org/abs/2608.22309.

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. doi: 10.1175/mwr-d-14-00269.1.

Novin Shahroudi, Mihkel Lepson, and Meelis Kull. Evaluation of trajectory distribution predictions with energy score. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44322–44341, 2024. URL https://proceedings.mlr.press/v235/shahroudi24a.html.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. arXiv preprint arXiv:2509.26468v1, 2025. URL https://arxiv.org/abs/2509.26468v1.

Jiaze Sun, Kelvin J.L. Koa, Ruiyang Ni, Yize Liu, Haonan Chen, and Ke-Wei Huang. FinStressTS: A Parametric Synthetic Benchmark for Time-Series Forecasting in Finance. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 9836–9847, 2026. doi: 10.1145/3770855.3817578.

Rui Wang, Yihe Dong, Sercan Ö. Arik, and Rose Yu. Koopman neural operator forecaster for time-series with temporal distributional shifts. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kUmdmHxK5N.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Jiawen Zhang, Xumeng Wen, Zhenwei Zhang, Shun Zheng, Jia Li, and Jiang Bian. ProbTS: Benchmarking point and distributional forecasting across diverse prediction horizons. In *Advances in Neural Information Processing Systems*, volume 37, pp. 48045–48082, 2024. doi: 10.52202/079017-1523.

Yuanzhao Zhang and William Gilpin. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning. In *International Conference on Learning Representations*, pp. 58314–58340, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5fbdcd4d2190682b913c6b39b58e95f0-Paper-Conference.pdf.

Ronghua Zheng, Hanru Bai, and Weiyang Ding. KooNPro: A variance-aware Koopman probabilistic model enhanced by neural process for time series forecasting. In *International Conference on Learning Representations*, pp. 29457–29481, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4846257e355f6923fc2a1fbe35099e91-Paper-Conference.pdf.

The appendices follow the forecasting chain and its evaluation: implementation and mathematical properties (Appendix A), data and comparison protocols (Appendix B), statistical estimands (Appendix C), end-to-end benchmarks (Appendix D), representation and trajectory evolution (Appendix E), inverse decoding and local output (Appendix F), and matched long-horizon controls (Appendix G).

A Implementation and mathematical properties

A.1 Causal encoding and index alignment

Let $J = c_q$ be the observation-window length. The encoder fits a local polynomial to the J observations ending at the current time. For normalized sample positions $u_j = j/(J-1)$, $j = 1-J, \ldots, 0$, and Vandermonde matrix $V_{jk} = u_j^k$, the derivative stencil for order k is

$$c_k^\top = \frac{k!}{(J-1)^k}(V^\dagger)_{k,:}. \tag{9}$$

The two feature families are four derivative blocks (`jet4`, polynomial degree four) and the current level with first and second derivatives (`level_jet2`, degree two). Derivative features subtract the current endpoint before applying the stencil. Coordinatewise training medians and interquartile ranges normalize features, and successive blocks receive weights $1, 0.45, 0.45^2, \ldots$. K-means uses five initializations and at most 30,000 sampled training windows. Prediction is nearest-center assignment under the resulting frozen transformation. No centered stencil uses future observations.

The default J is 9 when context $C \geq 32$, 5 when $16 \leq C < 32$, and 3 otherwise; feature families are restricted when a stencil cannot support the requested order. For $d > 64$, the encoder summarizes observations using the leading 16 signed coordinates, 16 coordinate-band means, and 16 log mean squared amplitudes. These summaries cover all coordinate bands but are lossy, with no fixed retained proportion of the original field information.

For a zero-indexed encoded sequence, the first pair ends at index 1, the first W-pair frequency ends at index W, and the first change ends at $W + L$. On an eligible context of C raw observations,

$$N_q = C - J + 1, \qquad N_B = N_q - 1, \qquad N_F = N_q - W, \qquad N_D = N_q - W - L. \tag{10}$$

The history length must satisfy $1 \leq p \leq N_D$. The saved encoder window is task-specific: Hospital admissions $H = 13$ uses $J = 3$ with $C = 12, W = 1, L = 4$, giving $N_q = 10$, $N_B = N_F = 9$, and $N_D = 5$. Its selected $p = 5$ therefore uses five observed change nodes. In raw indices $0, \ldots, 11$, encoded states end at $2, \ldots, 11$, frequency windows at $3, \ldots, 11$, and changes at $7, \ldots, 11$. Table 17 reports each diagnostic task's saved J, available N_D, and selected p. Independent recordings are never concatenated to create artificial boundary transitions. Missing inputs are forward-filled using available history; missing targets remain masked in evaluation.

A.2 POD, sparse fitting, and memory initialization

Transition POD uses the streamed covariance of eligible training changes. Eigenvector signs follow a deterministic convention. The training mean is corrected for any numerical all-ones component, and coordinate standard deviations have floor 10^{-5}. Spatial POD used by field adapters is a separate transformation fitted before state encoding. A spatial rank and a transition-POD rank must not be confused.

For the design matrix A of Equation (3), define feature scales $c_j = (n^{-1}\sum_i A_{ij}^2)^{1/2}$, replacing near-zero scales and the intercept scale by one. With $X = A\,\mathrm{diag}(c)^{-1}$, the initial fit is

$$\widetilde{\Xi} = (X^\top X/n + \lambda R)^{-1} X^\top Y/n, \qquad \Xi = \mathrm{diag}(c)^{-1}\widetilde{\Xi}, \tag{11}$$

where Y contains the next observed coordinates, $\lambda = 0.01$, and R leaves the intercept unpenalized. Three sequential threshold/refit passes remove normalized coefficients smaller than 0.02 and refit on the retained support. The intercept is retained. Fits use up to 16,000 true temporal transitions, with their correct lag and memory values; the POD covariance itself uses all eligible training changes. The residual covariance is represented by an eigendecomposition with negative numerical eigenvalues clipped to zero. Independent Gaussian innovations are drawn at each rollout step with this full cross-coordinate covariance.

All four memory channels start at the first latent coordinate of an independent sequence. Their states persist through processing chunks and reset only at sequence boundaries. At an evaluation origin, initialization uses the latent prefix available inside the observation context. The finite history length is validation-selected from the deduplicated set $\{1, \min(8, N_D), \min(17, N_D)\}$; the decay rates 0.6 and 0.9 are fixed. The model does not perform continuous delay optimization. Direction uncoupling restricts each output to its own coordinate's features, while retaining the same residual-covariance estimation procedure.

A.3 Inverse readout architecture and objective

The dynamic channels are the per-step mean and standard deviation of the integrated frequency ensemble. Static boundary channels consist of the $W-1$ observed pair tables, the L frequency anchors, and current one-hot state. Two time features are h/H and $\min(h, W)/W$. The total input width is

$$c_{\text{in}} = (W + L + 1)K^2 + K + 2. \tag{12}$$

Feature normalization is fitted on decoder training inputs, with standard-deviation floor 0.05. A pointwise input convolution maps to 16 channels. Three residual blocks each contain two length-3 left-padded convolutions, using dilations 1, 2, and 4 respectively, with ReLU between convolutions and after addition. A pointwise output convolution gives K^2 logits; global softmax precedes the row-simplex projection in Equation (6). The convolutional receptive field is 29 forecast nodes, and marginal propagation additionally links successive steps. The trainable parameter count is

$$16c_{\text{in}} + 4720 + 17K^2. \tag{13}$$

For $K = 4, W = 7, L = 4$ this is 8,160 parameters; other configurations change the input and output widths.

For N training windows, the supervised objective is

$$\mathcal{L}_B = \frac{1}{NH} \sum_{n=1}^{N} \sum_{h=1}^{H} \left\| \widehat{B}_{n,h} - e_{q_{n,t+h-1}} e_{q_{n,t+h}}^{\top} \right\|_F^2. \tag{14}$$

Future training labels serve as targets. The decoder uses up to 256 training windows, 32 validation windows, 16 upstream draws per training window, and 32 draws at deployment. It receives 400 Adam updates, batch size 8, learning rate 0.002, and weight decay 10^{-4}. Validation Brier loss selects checkpoints every 50 updates. Documented memory-budget reductions may lower training/validation window counts on unusually wide tasks; the draw counts remain fixed. Upstream weights remain fixed during decoder training, and the local continuous writer is fitted separately.

The frequency prediction scored as F MSE is $\widehat{F}^{\text{check}}$, obtained by convolving the decoded joint tables with their observed boundary. By Proposition 1, this table gives the conditional expectation of the sampled paths' empirical window frequencies.

A.4 Local output functions

Write standardized observations as x_t, and let

$$v_t = [x_t^{\top}, ((x_t - x_{t-1})/0.2)^{\top}]^{\top}. \tag{15}$$

The constant 0.2 serves as a feature scale. $\varphi_m(v_t)$ contains every monomial through degree m, including the intercept, so its size is $f = \binom{2d+m}{m}$. For $d \le 32$, degrees 1, 2, and 3 are considered only when $f \le 512$. For higher dimensions only degree 1 is considered, with budget $\max(512, 2d+1)$. An over-budget degree is rejected rather than truncating its feature list.

Training targets are residual accelerations $r_t = x_{t+1} - 2x_t + x_{t-1}$. Polynomial columns receive RMS normalization with floor 0.01. A shared ridge fit uses regularization 10^{-4} and an intercept penalty 0.01 times the other penalties. For each pair $e = (i, j)$, a second ridge fit models the shared fit's residual, with additive normal-equation regularization $300I$. Pairs with fewer than $\max(40, \lfloor f/2 \rfloor)$ observations use zero local correction. At most 16,000 training rows are sampled per fitting seed.

Pairwise residual covariance is used when at least ten residuals are available; otherwise the global covariance is used. Pair covariance combines 85% of its empirical estimate and 15% of the global estimate, with a small isotropic numerical floor. The innovation multiplier is selected from $\{0, 0.5, 1\}$

jointly with the eligible polynomial degree using full-chain validation Energy. This is shared regression followed by local residual fitting, not an end-to-end mixture-of-experts objective.

A.5 Forecast algorithm

Algorithm 1 STRIDE inference with frozen parameters

Require: Observed context $y_{t-C+1:t}$, horizon H, ensemble size S, frozen encoder, POD, sparse map, readout, and local functions.

Encode the context causally to obtain q; compute observed B, F, D with Equation (1).
Project D to standardized coordinates; initialize and replicate finite-lag and memory states from the observed prefix for all S paths.
for $s = 1, \dots, S$ **do**
  for $h = 1, \dots, H$ **do**
    Predict the next coordinate with Equation (3); update memory with Equation (4).
    Reconstruct $\widehat{D}^{(s)}_{t+h}$ and integrate $\widehat{F}^{\mathrm{raw},(s)}_{t+h}$ using observed anchors or earlier forecasts.
  end for
end for
Form mean/std forecast features and observed boundary features; apply the causal inverse readout.
Project rows recursively to obtain $\widehat{B}_{1:H}$ and marginals $\pi_{0:H}$.
for $s = 1, \dots, S$ **do**
  Sample a discrete path from $T_h(i,j) = \widehat{B}_h(i,j)/\pi_{h-1}(i)$ on positive-mass rows.
  Recursively apply the selected pairwise local functions and output innovations to produce $\widehat{y}^{(s)}_{t+1:t+H}$.
end for
return Continuous ensemble, discrete paths, joint tables, checked frequencies, and upstream change ensemble.

A.6 Proofs and inverse-error identities

Proposition 2 (Forecast-time causality)**.** *Fix trained parameters, the fitted horizon, and innovation streams. Forecasts through step h depend only on observed history and innovations through that step. Changing decoder feature inputs after h does not change its predictions through h.*

Proof. The encoder and frequency lift use trailing observations. The sparse map uses prior coordinates and current innovations; memory uses its preceding state and the current coordinate. Integration uses observed anchors or preceding forecasts. Left-padded convolutions use only current and earlier forecast features, and their normalizers are fixed training statistics. Probability projection uses the preceding marginal. State sampling and the continuous writer recurse on preceding values and current random draws. Induction over the forecast steps proves the statement. Forecast-prefix comparisons hold for a fixed fitted horizon and coupled innovation streams; validation may use target values to select parameters before deployment. □

Proof of Proposition 1. Projection yields $\widehat{B}_h \mathbf{1} = \pi_{h-1}$, $\widehat{B}_h \geq 0$, and total mass one by induction from π_0. If the previous state has marginal π_{h-1}, its next pair probability is $\pi_{h-1}(i)T_h(i,j) = \widehat{B}_h(i,j)$ on all positive-mass rows. Zero-mass rows contribute zero. Summing over source states yields π_h. Induction gives the joint marginals at every step. Linearity of expectation then shows that convolving these pair marginals and the observed boundary gives the expected window table under the decoded conditional law. □

Proposition 3 (Hard inverse and propagation of approximation error)**.** *Given exact frequency anchors and the last $W-1$ observed pair tables, exact future D uniquely determines future F and B through*

$$F_u = D_u + F_{u-L}, \qquad B_u = WF_u - \sum_{k=1}^{W-1} B_{u-k}. \tag{16}$$

Table 1: Evaluation scope. Planned tasks are the chosen full experiments, not the much larger configuration-only inventory. Evaluated tasks include all completed result records, even if a rollout was incomplete or not fully scorable. Units describe evaluated systems/source corpora, rather than independent replications. All comparisons preserve their original data splits, seeds, horizons and forecast origins.

Family	Planned	Evaluated	Fully scorable	Units
BOOM	24	23	23	23
FinStressTS	30	30	30	6
GIFT-Eval	10	10	7	10
MDBench	108	102	101	66
MDGen pilot	1	1	1	1
ProbTS	24	24	24	3
StocBench	1	1	1	1
dysts	22	22	21	8
fev	20	17	12	15
Total	240	230	220	133

For the unconstrained recursion applied to approximate changes, errors satisfy

$$e_u^F = e_u^D + e_{u-L}^F, \qquad e_u^B = We_u^F - \sum_{k=1}^{W-1} e_{u-k}^B. \tag{17}$$

If the anchors are exact and $\|e_{t+h}^D\| \le \epsilon$*, then* $\|e_{t+h}^F\| \le \lceil h/L \rceil \epsilon$.

Proof. Rearranging the definitions in Equation (1) gives the two recursions and uniqueness by induction. Subtracting the corresponding exact recursions gives the error identities. Expansion along a residue class modulo L accumulates at most $\lceil h/L \rceil$ errors, so the triangle inequality gives the bound. □

The bound quantifies worst-case accumulation in the unconstrained algebraic inverse. The experimental hard-inverse control additionally projects each output to feasible probabilities, introducing projection corrections into its error recurrence. Appendix F compares this control with the learned readout while holding the upstream forecasts fixed.

B Data, configuration, and baseline protocols

Dynamical and simulation families. The dysts subset uses chaotic systems and sampling granularities from Gilpin (2021), with chronological 60/20/20 splits. MDBench (Bideh et al., 2026) is adapted to forecasting observed time trajectories. The adapter reads observations `u` and timestamps, but not the simulator-provided derivative labels `du`. The derivative fingerprints used to define q are instead estimated from trailing observation windows by the local polynomial fit in Appendix A. Thus the encoder uses observation-derived derivatives available at prediction time. Its noise variants are grouped by underlying system. The FinStressTS software benchmark (Sun et al., 2026) contributes six synthetic mechanism families with five difficulty settings each. The supplied chronological splits and fixed forecast endpoints are retained. The benchmark configuration registry supplies each task's context and horizon.

Time-series families. ProbTS (Zhang et al., 2024) supplies ETT, exchange-rate, and weather series with its chronological splits and frozen horizons. The conservative source grouping combines the four ETT variants into one corpus, with exchange rate and weather as two additional groups. GIFT-Eval (Aksu et al., 2024), fev (Shchur et al., 2025), and BOOM (Cohen et al., 2025) use declared subsets and preserved official test origins, with a validation segment preceding test and supervised fitting on the remaining prefix. These are supervised forecasting adaptations, not full-benchmark zero-shot leaderboards. We do not use external covariates. BOOM's large configuration inventory does not mean thousands of completed full model comparisons.

Table 2: Fixed configurations for four diagnostic tasks. C/H: observation context/forecast horizon; K: discrete states; W: transition-pair window; L: difference lag; r: POD rank; p: finite-memory nodes. The causal encoder uses nine observed points in each task. Writer degree is chosen from validation before testing.

Task	C/H	K	W	L	r	p	Degree
Aizawa	96/64	2	7	8	2	17	3
GARCH	96/1	4	7	4	14	17	1
Exchange	256/96	2	7	8	1	1	1
StocBench	96/10	4	7	4	15	17	1

Molecular and field observations. The MDGen pilot (Jing et al., 2024) uses 20 molecules split into 12 training, four validation, and four test molecules. The observations are 12-dimensional sine/cosine backbone-angle features, subsampled with stride 100 to approximately 10 ps between observations. This is molecular-observable forecasting, not an all-atom generation evaluation. StocBench (Pfister et al., 2026) uses 500 trajectories per split, generation seeds 42/43/44, with 200 time steps and 64×64 scalar fields. A training spatial POD retains 1,013 modes and 95.003% of training field variance. Its history-conditioned forecast task differs from the source benchmark's single-state conditional reference task. Native-field scores additionally account for discarded spatial components where available; the tabled principal Energy uses the declared modeled coordinates.

Eligibility and selection. Ten of the 240 planned tasks cannot support the frozen representation protocol. The fev tasks Rossmann weekly, US consumption yearly, and world CO_2 have training lengths 16, 9, and 10 with horizons 13, 5, and 5: their remaining histories are 3, 4, and 5 points, below the seven-point minimum. Six MDBench PDE tasks fail the requirement of 32 training D nodes even at the smallest $W = L = 1$: heat laser and uniform 3D soil heat each provide at most 8 nodes; Navier–Stokes channel, Navier–Stokes cylinder, and reaction–diffusion cylinder each provide 24; 2D advection–diffusion provides 30. Their observations vary with time, but their training records are too short for this protocol. BOOM task `ds-1815-D` has a constant 77-point training prefix and cannot support the minimum two distinct fingerprint states. These exclusions precede forecast evaluation and are separate from forecast failures.

State counts $K \in \{2, 4, 8, 12\}$ require sufficient distinct training fingerprints; StocBench uses $K \in \{2, 4\}$. The initial W grid is $\{3, 7, 15, 31\}$ for long contexts and $\{1, 3, 7\}$ for short ones, and $L \in \{1, 2, 4, 8\}$. For each representation, at most 4,096 sampled training changes give sorted covariance eigenvalues λ_j. Candidate ranks are the smallest integers r_γ satisfying $\sum_{j \le r_\gamma} \lambda_j / \sum_j \lambda_j \ge \gamma$, for $\gamma \in \{0.90, 0.95, 0.99\}$, within the numerical rank. Feasibility also requires enough history for the chosen convolution, difference lag, and coordinate model.

The selection procedure preserves structural diversity and an eligible original $K = 4, W = 7, L = 4$ derivative encoding. Up to 12 representations are shortlisted by validation projection retention and transition support. Each basis is then refitted to all eligible training changes. Its rank candidates and feasible histories $p \in \{1, \min(8, N_D), \min(17, N_D)\}$ are compared using deterministic validation D MSE over $\min(H, 16)$ steps, divided by last-D persistence MSE on the same windows. Up to three finalists, retaining the eligible original configuration when necessary, enter full-horizon validation of the complete chain and local-function settings. The selected basis contains the leading r full-training eigenvectors. Since candidate ranks originated from a sampled spectrum, their final retained variance need not equal the screening threshold; Table 3 records the realized values. All selection uses training and validation data. Full locks, origins, normalization hashes, and seeds remain in the artifacts.

Deterministic shortlist construction. For every feasible representation, let v be validation projection retention and a the validation transition mass assigned to pairs with fewer than eight training occurrences. Representations are sorted lexicographically by $(\max(0, 0.90 - v), a, r)$. The eligible `jet4` configuration with $K = 4, W = 7, L = 4$ is inserted first. Scanning the sorted list adds at most two further configurations per encoder identifier, skips duplicate $(\text{encoder}, W, L)$ triples, and stops at 12 entries. Each encoder identifier specifies its fingerprint family, observation stencil, and state count. After the rank/history continuation tests, successful entries are sorted by normalized

Table 3: Derivative-fingerprint states and retained transition-POD variance in four focused diagnostic tasks. All use a causal nine-observation stencil. `jet4` uses derivative orders 1–4; `level_jet2` uses the current level and orders 1–2. The percentage is measured on all eligible training D nodes using the final selected basis; it is not a nominal screening threshold or a spatial-POD retention rate.

Task	Fingerprint	K	r	Training D variance (%)
Aizawa	`level_jet2`	2	2	99.035
GARCH	`jet4`	4	14	99.116
Exchange	`level_jet2`	2	1	96.094
StocBench	`jet4`	4	15	100.000

validation D MSE. The first three enter complete-chain validation; an eligible original configuration replaces the last finalist when necessary. The frozen implementation and its hash manifest accompany the source package.

B.1 Comparator implementations

Shared task interface and comparison scope. The fourteen baselines receive the task's observation coordinates, training-only normalization, splits, context C, horizon H, masks, and frozen test origins. Field tasks share the frozen spatial observation-POD adapter; this is separate from STRIDE's transition-POD model. Baselines do not receive STRIDE's q or D; the same-q ablation is a separate module control. Independent recordings retain their boundaries, and no future external covariates are supplied. Training and selection are repeated separately for each task; neural weights may be shared across channels or series within that task.

Markov–Gaussian, first- and third-order Gaussian HMMs, ARHMM, continuous Gaussian Markov, and VAR cover all 230 evaluated tasks. Direct SINDy, NVAR, and context parroting cover 125 low-dimensional dynamical tasks (observation dimension 1–12). Seasonal naive and Holt cover 104 conventional time-series tasks. DLinear and TimePrism cover 87 tasks selected using metadata and a fixed hash ranking; DeepAR covers a smaller, declared 24-task subset. Selection of these applicability sets does not use test performance. Table 4 records the complete method-specific coverage.

Predictive distributions and deterministic controls. Markov–Gaussian, continuous Gaussian Markov, Gaussian HMM, third-order HMM, ARHMM, VAR, DeepAR, and TimePrism return 32 paths, as does STRIDE. Direct SINDy, NVAR, context parroting, DLinear, seasonal naive, and Holt each return one deterministic path. That path is passed directly to the scorer as a point mass, without replication or invented residual noise. Its Energy is $\mathrm{ES}(\delta_x, y) = \|x - y\|_2/\sqrt{M}$, its marginal CRPS is absolute error, and its empirical prediction intervals have zero width. The three protocol seed records for an identical deterministic fit are not independent training replicates.

Markov–Gaussian. K-means partitions normalized observations directly, with $K \in \{2, 4, 8\}$ when supported by training size, three initializations, and at most 16,000 training transitions. Empirical transition counts use additive smoothing 0.01. Each destination state has a diagonal Gaussian emission with variance floor 10^{-4}. The last observation initializes its nearest state; recursive state and emission sampling produces 32 paths. This comparator has its own observation clusters and emissions. The same-q Markov ablation instead preserves STRIDE's derivative-fingerprint encoder and local functions.

Gaussian HMM. We fit diagonal Gaussian emissions using `hmmlearn` 0.3.3, with 2, 4, or 8 states, at most 100 EM iterations, and likelihood tolerance 0.01. Training uses contiguous blocks of at most 256 observations, totaling at most 16,000, with explicit block boundaries. Forward filtering of the observed context gives the final-state posterior. Sampling future states and emissions produces 32 paths.

ARHMM. Each switching state has an affine vector AR(1) emission and diagonal innovation covariance. Forward–backward probabilities use `hmmlearn`; emission updates use independently implemented weighted ridge regression with penalty 0.001 and variance floor 10^{-4}. State counts

and training blocks match the HMM protocol. Observed adjacent pairs initialize filtering, followed by recursive generation of 32 paths. This is an implementation of the model family, with inference checked against small enumerated examples.

Continuous Gaussian Markov control. This baseline fits the continuous first-order model $x_{t+1} = Ax_t + b + \epsilon_t$ directly in training-standardized observation coordinates. Ordinary least squares uses all within-recording training transitions, without ridge tuning; the innovation covariance is the full residual covariance $\Sigma = R^\top R/N$. Prediction initializes from the last observed vector and samples 32 recursive paths. It uses no clustering, transition-POD, inverse readout, or STRIDE local function. Unlike the clustered Markov–Gaussian baseline above, its state is the continuous observation itself. Its three seeds affect sampling, not the deterministic fitted coefficients.

Third-order Gaussian HMM. The hidden transition is $A_{abcd} = \Pr(s_{t+1} = d \mid s_{t-2} = a, s_{t-1} = b, s_t = c)$, with K shared diagonal Gaussian emissions. Exact forward–backward inference tracks K^3 hidden triples and permits only overlapping transitions $(a, b, c) \to (b, c, d)$. Baum–Welch estimates the transition tensor, initial triple distribution, and shared emission parameters; independent training sequences retain their boundaries. State counts $K \in \{2, 4, 8\}$ are selected by seed-1729 validation Energy and then fixed across the three seeds. The training-block and emission-update protocols match the Gaussian HMM above: at most 16,000 rows in blocks of length 256, up to 100 EM iterations, and likelihood-increment tolerance 0.01. Initial and transition counts use additive smoothing 0.01. Filtering the observed context and sampling the resulting hidden triples and emissions gives 32 forecast paths. This fixed-order model is evaluated on all 230 tasks.

All Gaussian state baselines follow each task's frozen split, normalization, context, horizon, and test origins, and have their own output distributions. The same-q ablation instead holds STRIDE's encoder and local writer fixed to isolate the transition forecaster.

Direct SINDy. This control fits a discrete observation-space map $y_{t+1} = f(y_t)$ with a polynomial library of degree 2 or 3. Validation selects threshold 0.01, 0.05, or 0.1. At most 16,000 training transitions are used, with ridge 10^{-6}, up to 25 threshold/refit passes, and a final unpenalized fit on the selected support. Iteration gives one deterministic path. The adapter does not use simulator-provided derivatives or clip divergent predictions.

Seasonal naive and Holt. Seasonal naive repeats each channel's last cycle, selecting a feasible period from $\{1, 5, 7, 12, 24, 52, 168\}$ by validation Energy. Holt initializes its level and trend from the current context. Candidates use level smoothing $\alpha \in \{0.2, 0.5, 0.8\}$, with either simple exponential smoothing or damped additive trend ($\beta = 0.1$, damping 0.98). Both generate one channelwise deterministic continuation; the latter is the stated Holt/SES grid, not AutoETS or AutoARIMA.

Gaussian VAR. An affine vector autoregression uses at most 16,000 training transitions within recording boundaries. Validation selects lag 1, 2, or 4 and ridge 10^{-6}, 10^{-3}, or 0.1; dimensions above 128 use lag one. Full residual covariance is eigendecomposed with negative numerical eigenvalues clipped to zero. Independent-in-time multivariate Gaussian innovations generate 32 paths, retaining cross-channel innovation dependence.

NVAR. Following Gauthier et al. (2021), we concatenate p consecutive observed vectors, $p \in \{1, 2, 4\}$, and form an intercept, linear terms, and all upper-triangular quadratic monomials. Ridge regression uses the same three penalties as VAR and at most 16,000 training transitions. The predicted observation change is added to the last value at each step. The adapter returns one deterministic path and is restricted to the declared low-dimensional tasks.

Context parroting. The pinned author retrieval-and-continuation function (Zhang & Gilpin, 2026) is applied separately to each channel. Validation selects embedding dimension 4, 8, 16, or 30 subject to twice that dimension being smaller than C; delay is one. The routine produces the H-step continuation directly from the current context. There are no globally fitted forecast weights, added noise, or cross-channel retrieval terms; output is one deterministic multivariate path.

DLinear. A width-seven moving average separates each channel's context into trend and seasonal parts. Two shared linear C-to-H heads predict these parts and sum them (Zeng et al., 2023). Training

pools channel/series windows within the current task under MSE. The shared-channel variant returns one deterministic path and does not fit a residual distribution. Validation MSE selects its checkpoint.

DeepAR. The pinned GluonTS Student-t implementation (Salinas et al., 2020) uses two LSTM layers of width 40, dropout 0.1, scaling, and lag set $\{1\}$. It shares the network across univariate channels/series within a task. Calendar and static placeholders are zero; the nominal frequency is daily, with lag explicitly fixed to one and no future covariates. Native likelihood selects checkpoints. It samples 32 autoregressive paths per channel and assembles them into multivariate paths. This adapter's conditional joint law factorizes across channels.

TimePrism. The pinned author model (Dai et al., 2026) uses 625 hypotheses, width-seven DLinear decomposition, interaction rank eight, relaxed winner-take-all loss with $\epsilon = 0.01$, and classification-loss weight one. Mean scaling and the recorded standard-deviation settings are retained. Training pools channel windows within the task; native validation loss selects checkpoints. The author network returns 32 scenario draws per channel; the wrapper reshapes them into multivariate paths without an additional resampling step. Channels are conditionally independent in this adapter. Native validation excludes partially labeled channels because its path loss requires complete paths; common masked testing retains the prescribed origins.

Validation selection and reproducibility. Classical grids use seed-1729 validation Energy, then keep settings fixed across fitting seeds. For neural baselines, the native validation loss selects the retained checkpoint under the frozen protocol. The supplementary file `evidence/optimization_protocol.json` records optimizer settings, update limits, validation frequency, stopping rules, and source hashes. Table 4 reports comparison availability and forecast completion for each method.

Table 4: Fourteen-comparator availability across 2,161 planned task–method pairs (6,483 seed records). Available requires complete scorable forecasts from both methods at all prescribed origins and all three seeds. S_f and B_f count STRIDE/comparator forecast failures; Shared is their intersection. Missing denotes unobservable targets and may overlap failures. Third-order HMM covers all 230 tasks in nine families; its forecasts complete on all tasks, with two tasks lacking observable targets at some origins.

Comparator	Planned	Available	S_f	B_f	Shared	Missing
Markov–Gaussian	230	220	10	0	0	2
Continuous Markov	230	220	10	0	0	2
Gaussian HMM	230	220	10	0	0	2
Third-order HMM	230	220	10	0	0	2
ARHMM	230	220	10	0	0	2
Direct SINDy	125	114	2	10	1	0
VAR	230	219	10	1	0	2
NVAR	125	110	2	13	0	0
Context parroting	125	123	2	0	0	0
Seasonal naive	104	96	8	0	0	2
Holt grid	104	96	8	0	0	2
DLinear	87	78	9	0	0	2
DeepAR	24	21	3	0	0	0
TimePrism	87	78	9	0	0	2

C Statistical estimands, failures, and diagnostics

C.1 Family-level comparisons

For comparator b, let $\mathcal{C}_{u,b}$ contain conditions in unit u for which both methods succeed at all prescribed origins for all three fitting seeds. If $\ell_{a,c,s}$ is a method's origin-averaged score, the paired unit

difference is

$$\delta_{u,b} = \frac{1}{|\mathcal{C}_{u,b}|} \sum_{c \in \mathcal{C}_{u,b}} \frac{1}{3} \sum_{s=1}^{3} (\ell_{b,c,s} - \ell_{\text{STRIDE},c,s}), \qquad \widehat{\Delta}_b = \frac{1}{U_b} \sum_{u=1}^{U_b} \delta_{u,b}. \tag{18}$$

Bootstrap replicates resample U_b units with replacement using 2,000 draws and random seed 20260922, then use percentile endpoints. Point estimates and intervals have exactly the same unit weights. The intervals condition on fitted seeds and selected configurations; they do not estimate variability from rerunning the entire search. No interval is reported for a single unit. Few-unit intervals, particularly the three ProbTS source groups, have limited resolution. Across all reported family/baseline comparisons these intervals are descriptive and not adjusted for multiplicity.

Grouping merges dysts sampling rates, MDBench noise versions, related Lorenz/Rossler regimes within MDBench, and FinStress difficulty levels. The ETT variants form one source group; matching fev hospital frequency versions are merged. A sensitivity map additionally records recurring systems across benchmark families. A family-specific group count is not a count of globally independent physical systems, and we do not average raw Energy across families. Point estimates and intervals use the same system-weighted estimand. Grouping operates on fixed per-task predictions.

C.2 Completion and missing targets

Every scheduled task-method-seed record is retained in the failure inventory. Training or selection failure, nonfinite forecasts, finite but extremely large scores, and missing target labels are distinct cases. A task can have both forecast failures and unobservable origins, so their counts are not disjoint. A forecast origin with any failed ensemble member is not scored on its surviving members. A fully generated forecast with no observed targets is complete but unscorable. The finite Energy comparisons condition on complete three-seed paired tasks; their results are not claims about missing or failed tasks. No arbitrary finite penalty is substituted for failures.

Numerical failures in the dynamical baselines remain in their scheduled comparison cohorts.

C.3 Within-task paired tests

The context-only decoder comparison resamples both three fitting seeds and independent test trajectories or chronological blocks spanning $C + H$. A sensitivity analysis uses blocks spanning H. There are 11 principal blocks for Aizawa, five for GARCH, five for Exchange, and 213 sampled test-trajectory clusters for StocBench. A cluster is a within-task sampling unit, not a new underlying dynamical system. Approximate independence of time blocks is an assumption; dynamics can preserve longer dependence.

The saved whole-horizon Energy comparisons use paired block sign-flip tests with Holm adjustment over the original family of eight comparisons. The activity-stratified trajectory diagnostics retain their recorded correction families, as specified in Appendix E.5. Secondary $q/B/F$ diagnostics use their saved paired intervals and are not a jointly multiplicity-adjusted family. Earlier hard-inverse/module-ablation intervals average seeds first and then resample recording clusters or blocks covering H overlap. We explicitly distinguish these conditional-seed intervals from the later seed-and-block intervals.

C.4 Distribution and representation scores

The Energy score in Equation (8) uses the forecast path as one vector and applies as written for $S > 1$. For deterministic baselines, $S = 1$ and the exact point-mass score is $\|x - y\|_2/\sqrt{M}$; no ensemble-pair correction is needed. All observed time-coordinate pairs form its mask; $M = Hd$ for a complete target. The denominator $S(S-1)$ excludes self-pairs. For independent draws from the issued law conditional on the decoded table sequence, this gives an unbiased estimator of its Energy score. The analogous fair CRPS averages the scalar absolute-distance estimator across observed coordinates. Coverage and interval widths use empirical ensemble quantiles, and variogram scores use exponent $1/2$, temporal lags 1, 2, 4, and 8, and at most 256 fixed-seed coordinate pairs. A one-step task has no temporal variogram. These scores evaluate different aspects of a distribution and are not interchangeable.

Joint Brier and state Brier compare the decoded K^2 and K probabilities against observed one-hot targets. Change MSE evaluates the upstream predicted mean. Frequency MSE evaluates the checked convolution of decoded joint tables; the direct-F comparison in Appendix E.6 instead measures $\widehat{F}^{\mathrm{raw}}$ before decoding. Because (W, L) changes the change scale, the window grid separately saves normalized change errors; raw D MSE across different grids is not treated as the sole selection criterion.

D COMPLETE END-TO-END BENCHMARK RESULTS

In the tables below, n counts fully scorable paired *task conditions*: a fixed dataset variant, sampling/noise setting, and forecast horizon, evaluated at its prescribed origins and three protocol seeds. It does not count individual seeds, origins, sampled paths, or channels. The number u counts the distinct underlying systems or source corpora represented by those conditions. We average seeds within conditions, conditions within each group, and groups equally. For example, Aizawa at three sampling rates gives three conditions but one system; the sixteen ETT variant–horizon combinations give sixteen conditions but one source group. ProbTS therefore has $n = 24, u = 3$ against Markov–Gaussian, while DeepAR covers $n = 6, u = 3$. Both counts depend on which tasks are available for the comparator.

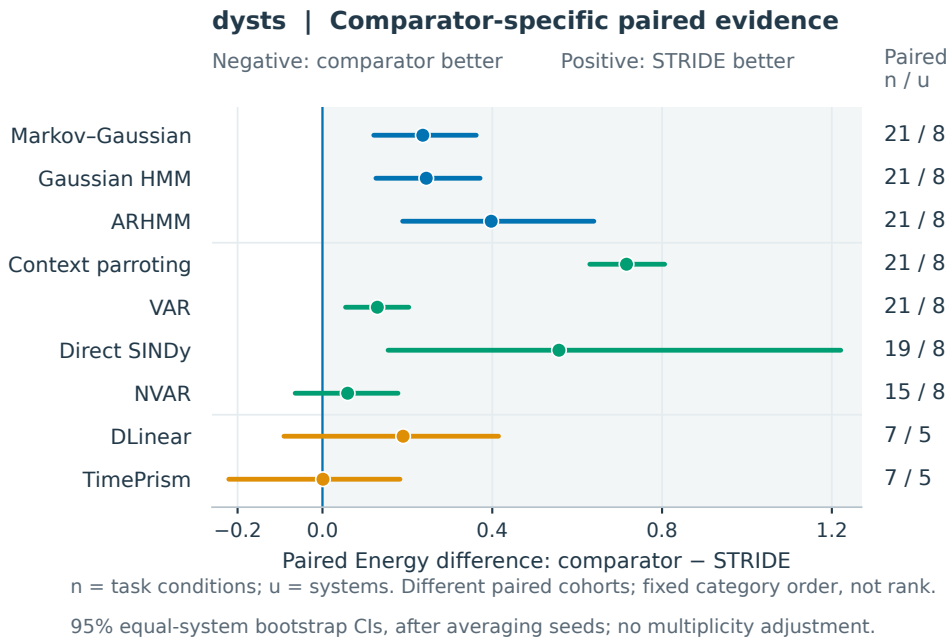


Figure 4: Whole-horizon chaotic-system forecasts. Points are system-weighted comparator-minus-STRIDE Energy differences; positive favors STRIDE. Lines are descriptive 95% group-bootstrap intervals. Cohort sizes are retained beside each row.

Table 5: Chaotic-system forecasts (dysts), including the independently fitted continuous Markov and third-order HMM controls. Lower path Energy is better; Δ = comparator minus STRIDE. Each row uses its own fully scorable three-seed cohort: n task conditions and u systems. Means weight systems equally; brackets are descriptive 95% system-bootstrap intervals. The third-order HMM row is the 22-task dysts subset of its full 230-task evaluation; all other rows retain their original coverage.

Comparator	n/u	STRIDE	Comparator	Δ	95% CI
Markov–Gaussian	21/8	0.3838	0.6202	0.2364	[0.1208, 0.3622]
Continuous Markov	21/8	0.3838	0.6125	0.2286	[0.0894, 0.3798]
Gaussian HMM	21/8	0.3838	0.6283	0.2445	[0.1259, 0.3713]
Third-order HMM	21/8	0.3838	0.5979	0.2141	[0.1015, 0.3423]
ARHMM	21/8	0.3838	0.7813	0.3975	[0.1888, 0.6394]
Direct SINDy	19/8	0.3257	0.8828	0.5571	[0.1545, 1.221]
VAR	21/8	0.3838	0.5131	0.1293	[0.05436, 0.2033]
NVAR	15/8	0.2721	0.3314	0.0593	[-0.06434, 0.1779]
Context parroting	21/8	0.3838	1.1	0.7163	[0.6298, 0.8061]
DLinear	7/5	0.4919	0.6818	0.1899	[-0.09082, 0.4147]
TimePrism	7/5	0.4919	0.493	0.001173	[-0.2207, 0.1826]

The MDBench family-score and source-diagnostic tables in this appendix exclude the Allee population-growth sequence. Figure 2 and the overall win fractions retain the full benchmark inventory.

Table 6: Separate within-family differences in path Energy, Δ = baseline − STRIDE (positive favors STRIDE). Each task is trained and evaluated on its own splits. Parentheses give available task conditions/system or source groups (n/u); every cell uses its own matched cohort. No cross-family transfer or pooling is performed. Full means and intervals appear in Table 7. Large finite errors are retained.

Family	Markov–Gaussian	VAR	DLinear	TimePrism
BOOM	0.05149 (23/23)	-0.0777 (23/23)	0.2639 (8/8)	0.3911 (8/8)
FinStressTS	-0.03626 (30/6)	-0.02438 (30/6)	0.2381 (6/6)	-0.01481 (6/6)
GIFT-Eval	0.07755 (7/7)	0.06331 (7/7)	0.06284 (7/7)	-0.04526 (7/7)
MDBench	0.589 (100/65)	-0.1964 (100/65)	-0.6178 (12/11)	-0.4329 (12/11)
MDGen pilot	0.01643 (1/1)	-0.02292 (1/1)	0.173 (1/1)	-0.006134 (1/1)
ProbTS	0.1454 (24/3)	1.097×10^{141} (23/3)	-0.123 (24/3)	0.0767 (24/3)
StocBench	-0.006921 (1/1)	-0.01032 (1/1)	0.2744 (1/1)	-0.0115 (1/1)
dysts	0.2364 (21/8)	0.1293 (21/8)	0.1899 (7/5)	0.001173 (7/5)
fev	1.534 (12/11)	-0.5023 (12/11)	-0.5842 (12/11)	0.6406 (12/11)

Table 7 reports the separate within-family comparisons for all fourteen forecasters, with comparator-specific means and availability records. Methods with different paired cohorts do not share a common STRIDE mean. Table 4 covers the full scheduled scope; Table 14 summarizes the five independent state-space baselines evaluated on all 230 tasks. The study includes 220 tasks fully scorable for STRIDE across all three fitting seeds. Each of the five full-inventory state-space baselines pairs with this same 220-task set.

Table 7: Complete within-family paired path-Energy results. n/u counts paired task conditions/system or source groups. Δ = comparator minus STRIDE; positive favors STRIDE. Each row uses its own three-seed cohort and equal group weights; intervals are descriptive 95% group-bootstrap intervals, omitted for one group. Every finite score in the stated reporting cohort is retained.

Comparator	n/u	STRIDE	Comparator	Δ	95% CI
BOOM					
Markov–Gaussian	23/23	1.538	1.589	0.05149	[-0.0264, 0.1494]
Continuous Markov	23/23	1.538	1.54	0.002675	[-0.1348, 0.102]
Gaussian HMM	23/23	1.538	1.578	0.04072	[-0.03205, 0.1308]
Third-order HMM	23/23	1.538	1.524	-0.01369	[-0.1046, 0.09559]
ARHMM	23/23	1.538	147.8	146.3	[-0.3368, 439.2]
VAR	23/23	1.538	1.46	-0.0777	[-0.2153, 0.03122]
Seasonal naive	23/23	1.538	2.143	0.6049	[-0.1537, 1.822]
Holt grid	23/23	1.538	1.603	0.06485	[-0.2882, 0.4793]
DLinear	8/8	2.571	2.835	0.2639	[0.06122, 0.5142]
DeepAR	4/4	4.696	4.912	0.2161	[0.02194, 0.4979]
TimePrism	8/8	2.571	2.962	0.3911	[-0.04425, 1.061]
FinStressTS					
Markov–Gaussian	30/6	0.7277	0.6914	-0.03626	[-0.04429, -0.02486]
Continuous Markov	30/6	0.7277	0.7026	-0.02504	[-0.03628, -0.01358]
Gaussian HMM	30/6	0.7277	0.6881	-0.03955	[-0.05043, -0.02581]
Third-order HMM	30/6	0.7277	0.6945	-0.03319	[-0.04675, -0.01769]
ARHMM	30/6	0.7277	0.6998	-0.02788	[-0.03535, -0.02279]
VAR	30/6	0.7277	0.7033	-0.02438	[-0.03573, -0.01232]
Seasonal naive	30/6	0.7277	1.291	0.5638	[0.4823, 0.6267]
Holt grid	30/6	0.7277	0.9995	0.2719	[0.2544, 0.2872]
DLinear	6/6	0.6931	0.9311	0.2381	[0.2205, 0.2511]
DeepAR	6/6	0.6931	0.6583	-0.03475	[-0.04124, -0.02873]
TimePrism	6/6	0.6931	0.6782	-0.01481	[-0.03039, 0.01313]
GIFT-Eval					
Markov–Gaussian	7/7	0.5729	0.6505	0.07755	[0.03392, 0.1236]
Continuous Markov	7/7	0.5729	0.6455	0.0726	[0.02154, 0.1278]
Gaussian HMM	7/7	0.5729	0.6363	0.06342	[0.02355, 0.1072]
Third-order HMM	7/7	0.5729	0.6204	0.04752	[0.00741, 0.08764]
ARHMM	7/7	0.5729	1.355	0.7824	[0.008952, 2.284]
VAR	7/7	0.5729	0.6362	0.06331	[0.02095, 0.112]
Seasonal naive	7/7	0.5729	0.7229	0.15	[-0.009495, 0.3077]
Holt grid	7/7	0.5729	0.8004	0.2275	[0.1367, 0.3106]
DLinear	7/7	0.5729	0.6358	0.06284	[-0.07014, 0.1901]
DeepAR	2/2	0.5657	0.6791	0.1135	[-0.02512, 0.2521]
TimePrism	7/7	0.5729	0.5277	-0.04526	[-0.16, 0.07013]
MDBench					
Markov–Gaussian	100/65	0.3851	0.9741	0.589	[0.03012, 1.049]
Continuous Markov	100/65	0.3851	0.3375	-0.04759	[-0.505, 0.2336]
Gaussian HMM	100/65	0.3851	1.12	0.7352	[0.1483, 1.214]
Third-order HMM	100/65	0.3851	1.172	0.7868	[0.1956, 1.274]
ARHMM	100/65	0.3851	0.3293	-0.05583	[-0.4801, 0.2033]
Direct SINDy	94/64	0.4029	0.2862	-0.1167	[-0.5576, 0.1529]
VAR	100/65	0.3851	0.1887	-0.1964	[-0.6539, 0.07886]
NVAR	94/65	0.3986	7.623×10^{72}	7.623×10^{72}	$[1.616\times10^{38}, 2.287\times10^{73}]$
Context parroting	100/65	0.3851	0.4065	0.02138	[-0.4253, 0.3061]
DLinear	12/11	1.593	0.9754	-0.6178	[-3.526, 1.127]
TimePrism	12/11	1.593	1.16	-0.4329	[-3.45, 1.403]
MDGen pilot					
Markov–Gaussian	1/1	0.5779	0.5944	0.01643	–
Continuous Markov	1/1	0.5779	0.5939	0.01596	–
Gaussian HMM	1/1	0.5779	0.6608	0.08285	–

Table 7, continued

Comparator	n/u	STRIDE	Comparator	Δ	95% CI
Third-order HMM	1/1	0.5779	0.5755	-0.002377	–
ARHMM	1/1	0.5779	0.6261	0.0482	–
Direct SINDy	0/0	–	–	–	–
VAR	1/1	0.5779	0.555	-0.02292	–
NVAR	0/0	–	–	–	–
Context parroting	1/1	0.5779	0.9754	0.3974	–
DLinear	1/1	0.5779	0.7509	0.173	–
TimePrism	1/1	0.5779	0.5718	-0.006134	–
ProbTS					
Markov–Gaussian	24/3	0.6617	0.8072	0.1454	[0.02274, 0.3596]
Continuous Markov	24/3	0.6617	0.6355	-0.0262	[-0.0459, 4.118×10^{-5}]
Gaussian HMM	24/3	0.6617	0.7903	0.1285	[0.003787, 0.3123]
Third-order HMM	24/3	0.6617	0.782	0.1203	[-0.008498, 0.3118]
ARHMM	24/3	0.6617	4.576×10^{43}	4.576×10^{43}	[0.00102, 1.373×10^{44}]
VAR	23/3	0.6517	1.097×10^{141}	1.097×10^{141}	[-0.0536, 3.291×10^{141}]
Seasonal naive	24/3	0.6617	0.6188	-0.04296	[-0.215, 0.109]
Holt grid	24/3	0.6617	0.6196	-0.04209	[-0.1937, 0.08677]
DLinear	24/3	0.6617	0.5387	-0.123	[-0.3823, 0.04165]
DeepAR	6/3	0.5181	0.5277	0.00956	[-0.3015, 0.1841]
TimePrism	24/3	0.6617	0.7384	0.0767	[-0.2065, 0.2431]
StocBench					
Markov–Gaussian	1/1	0.7138	0.7069	-0.006921	–
Continuous Markov	1/1	0.7138	0.6958	-0.01796	–
Gaussian HMM	1/1	0.7138	0.7067	-0.007118	–
Third-order HMM	1/1	0.7138	0.7067	-0.007077	–
ARHMM	1/1	0.7138	0.702	-0.0118	–
VAR	1/1	0.7138	0.7035	-0.01032	–
DLinear	1/1	0.7138	0.9882	0.2744	–
TimePrism	1/1	0.7138	0.7023	-0.0115	–
dysts					
Markov–Gaussian	21/8	0.3838	0.6202	0.2364	[0.1208, 0.3622]
Continuous Markov	21/8	0.3838	0.6125	0.2286	[0.0894, 0.3798]
Gaussian HMM	21/8	0.3838	0.6283	0.2445	[0.1259, 0.3713]
Third-order HMM	21/8	0.3838	0.5979	0.2141	[0.1015, 0.3423]
ARHMM	21/8	0.3838	0.7813	0.3975	[0.1888, 0.6394]
Direct SINDy	19/8	0.3257	0.8828	0.5571	[0.1545, 1.221]
VAR	21/8	0.3838	0.5131	0.1293	[0.05436, 0.2033]
NVAR	15/8	0.2721	0.3314	0.0593	[-0.06434, 0.1779]
Context parroting	21/8	0.3838	1.1	0.7163	[0.6298, 0.8061]
DLinear	7/5	0.4919	0.6818	0.1899	[-0.09082, 0.4147]
TimePrism	7/5	0.4919	0.493	0.001173	[-0.2207, 0.1826]
fev					
Markov–Gaussian	12/11	2.159	3.693	1.534	[0.01771, 4.515]
Continuous Markov	12/11	2.159	3.68	1.521	[-0.009264, 4.543]
Gaussian HMM	12/11	2.159	3.707	1.548	[0.007042, 4.584]
Third-order HMM	12/11	2.159	3.588	1.429	[-0.02642, 4.284]
ARHMM	12/11	2.159	11.38	9.218	[0.07234, 26.07]
VAR	12/11	2.159	1.657	-0.5023	[-1.5, 0.01432]
Seasonal naive	12/11	2.159	1.137	-1.022	[-3.628, 0.3819]
Holt grid	12/11	2.159	0.9448	-1.214	[-4.153, 0.2981]
DLinear	12/11	2.159	1.575	-0.5842	[-1.841, 0.1034]
DeepAR	3/3	0.7697	0.7861	0.01632	[-0.07089, 0.07869]
TimePrism	12/11	2.159	2.8	0.6406	[-0.1581, 2.127]

System/source-level relative accuracy. Table 8 reports system/source-level Energy ratios against VAR on all nine families. Each ratio compares the two methods after averaging the same task

conditions within that source; values below one favor STRIDE. MDBench has a median ratio of 0.582. The source-level diagnostic below identifies the individual systems contributing most to its mean score. Forecast completion is recorded separately in Table 4.

Table 8: Finite-loss distribution against VAR on 218 fully scorable paired task conditions in the stated Appendix D cohort. For each system/source group, R_u is its mean STRIDE Energy divided by its mean VAR Energy, averaging the same task conditions within each group. $R_u < 1$ favors STRIDE. Quantiles weight groups equally within each family and use linear interpolation; Max retains the largest finite ratio. No score is clipped, trimmed, or replaced by a failure penalty. With one group, all three summaries coincide. These descriptive ratios complement the raw family means and task win fractions.

Family	n/u	Median R_u	90th percentile	Max R_u
dysts	21/8	0.693	1.014	1.022
MDBench	100/65	0.582	2.804	8.216
FinStressTS	30/6	1.040	1.050	1.057
ProbTS	23/3	1.030	1.093	1.108
MDGen-pilot	1/1	1.041	1.041	1.041
StocBench	1/1	1.015	1.015	1.015
GIFT-Eval	7/7	0.918	0.998	1.003
fev	12/11	1.001	1.241	1.471
BOOM	23/23	0.997	1.199	1.251

Which systems determine the MDBench mean? The 100 paired conditions represent 65 equally weighted systems. STRIDE wins 75 conditions and 48 systems, with mean Energy 0.3851 versus 0.1887 for VAR. Table 9 identifies the four systems with the largest STRIDE system means in this cohort. If Allee is included, the 101 paired conditions represent 66 equally weighted systems: STRIDE wins 75 conditions and 48 systems, with mean Energy 25.4585 versus 0.2239 for VAR. Allee contributes 98.51% of the sum of STRIDE system scores; together with Kuramoto–Sivashinsky it contributes 99.50%. These full-inventory results locate the effect of the population-growth case examined below.

Table 9: Source-resolved MDBench diagnosis on the 100-condition, 65-system cohort. The four largest STRIDE system means are ranked by absolute Energy. Both methods average the same conditions within each system. Share is the percentage of the sum of the 65 STRIDE system means and measures contribution to the family mean.

System	STRIDE	VAR	Ratio	Share (%)
Kuramoto–Sivashinsky	16.597120	3.222938	5.150	66.301
Driven pendulum (quadratic damping)	1.121416	0.136491	8.216	4.480
Cell cycle	1.071909	0.181876	5.894	4.282
Landau equation	0.902971	3.033532	0.298	3.607

Allee: continuous extrapolation and recursive amplification. Allee is a population-growth dataset with a carrying capacity that limits its late-time population. The MDBench generator uses $\dot{x} = 0.14x(1 - x/130)(x/4.4 - 1)$ with $x(0) = 6.123$, so the growing population approaches 130 (Bideh et al., 2026). This 150-point sequence uses a chronological 90/30/30 split, context 75, horizon 15, and two test origins. Saved local-function targets span $[-1.078, 2.384]$ in training and $[9.303, 57.427]$ in testing, in training-standardized coordinates. The selected quadratic writer has 80 fitted rows; no state pair reaches the 40-row threshold for a local correction, so its pair-specific predictions use the shared polynomial. The two origin-averaged Energy scores are 0.7108 and 3309.7408, locating the large error in the second rollout.

Previously saved interventions distinguish the roles of state inputs and continuous recursion. Supplying true future states and removing innovations leaves autonomous Energy at 1657.6359, whereas refreshing the true continuous history at every step gives Energy 0.3097. The corresponding late MSEs are 1.0300×10^7 and 0.3240. Together with the training/test support ranges, these measurements identify extrapolation by the shared local function and amplification through recursive

continuous feedback. The true-input conditions are diagnostic measurements; the full-inventory record retains the original autonomous forecast scores.

Other systems have different support conditions. Kuramoto–Sivashinsky has 150 training points and three test origins; its field adapter reduces 1,024 native coordinates to nine spatial coordinates, retaining 99.36% of training variance. Among evaluated projected coordinates, 28.47% lie outside their training ranges, and the training state sequence contains only two switches. The quadratically damped driven pendulum likewise has 53.33% of test coordinates outside the corresponding training ranges. These records document limited transition support and continuous extrapolation exposure. The cell-cycle task has no such coordinate-range exceedance, so that explanation does not apply to its score difference. The supplementary record `evidence/mdb_source_diagnostics.json` supplies all 66 system scores, the saved controls, support statistics, and provenance hashes.

Task-resolved advantages. Table 10 shows lower whole-horizon Energy for STRIDE than both DLinear and TimePrism at every Weather horizon: $H = 96, 192, 336, 720$. The same two comparisons favor STRIDE on GIFT hospital $H = 12$, GIFT Jena $H = 48$, fev hospital admissions $H = 13$, and fev Jena $H = 24$. These task-level results complement the equal-source family averages. They are descriptive three-seed means, and horizons or recurring hospital/weather sources do not receive additional independent-system weight. The evidence archive retains all planned task/comparator entries and their complete-pair status, including comparisons outside this illustrative table.

Table 10: Task-resolved whole-horizon path Energy (lower is better). All four Weather horizons are shown, followed by illustrative hospital and Jena tasks. Each row uses the same frozen test origins and three-seed means for STRIDE, DLinear, and TimePrism. Reduction is $100(\text{comparator} - \text{STRIDE})/\text{comparator}$. These are descriptive task-level differences; horizons and recurring sources are not independent datasets. Complete task/comparator results and availability are retained in the evidence archive.

Task	STRIDE	DLinear	Red. (%)	TimePrism	Red. (%)
Weather, $H = 96$	0.37319	0.41782	10.68	0.67215	44.48
Weather, $H = 192$	0.43734	0.45890	4.70	0.64095	31.77
Weather, $H = 336$	0.46847	0.52921	11.48	0.64920	27.84
Weather, $H = 720$	0.54628	0.58595	6.77	0.63694	14.23
GIFT hospital, $H = 12$	0.75632	1.13237	33.21	0.95228	20.58
GIFT Jena, $H = 48$	0.46820	0.47879	2.21	0.59835	21.75
fev admissions, $H = 13$	0.72740	1.08626	33.04	0.80540	9.68
fev Jena, $H = 24$	0.38344	0.45405	15.55	0.41542	7.70

Source-resolved performance differences. The ProbTS comparison with DLinear has family-level Energy 0.661739 for STRIDE and 0.538728 for DLinear, a 22.8% difference relative to DLinear. The ETT corpus accounts for the larger gap: across its 16 conditions, the means are 0.981101 and 0.598764. Exchange gives 0.547797 versus 0.519449, whereas Weather favors STRIDE at 0.456319 versus 0.497969. Thus the aggregate contrast is localized rather than shared uniformly across the three sources. The descriptive transition-activity inventory is reported separately in Appendix E.3.

D.1 Whole-horizon wins, late accuracy, and growth

The lead-resolved comparison covers 200 tasks with $H > 1$ and 1,903 planned task–baseline comparisons across fourteen forecasters. Of these, 1,777 have complete paired whole- and late-horizon Energy scores. The 30 single-step tasks remain in the principal whole-horizon tables; a late slope is not assigned to them. The analysis uses saved predictions, checkpoints, and per-lead scores under the frozen task configurations. Coverage records distinguish forecast completion from availability of the target observations needed for scoring.

The last half uses zero-based array indices $\lfloor H/2 \rfloor, \ldots, H-1$, and the last quarter uses the final $\lceil H/4 \rceil$ entries. Path Energy is recomputed after restricting both samples and targets to that subpath. Its observed time–coordinate entries have equal weight in the standardized Euclidean norm, with no increasing lead-time weight. System/source weighting enters the cross-task comparisons below. Single-step Energy, ensemble-mean MSE, fair CRPS, and 90% coverage are also saved at every lead. Let e_h be the Energy of the single-step predictive distribution at lead h. For each origin, its late growth rate is

$$\beta = \frac{\sum_{h \in \mathcal{H}} (h - \bar{h})(e_h - \bar{e})}{\sum_{h \in \mathcal{H}} (h - \bar{h})^2}, \qquad \mathcal{H} = \{\lfloor H/2 \rfloor + 1, \ldots, H\}. \tag{19}$$

This OLS slope is computed before aggregation; negative values indicate decreasing error within the specified interval. The normalized slope $H\beta$, last-minus-first difference, and final error are additional descriptive fields. These quantities do not differentiate the joint path-Energy score.

For each method, scores are first averaged over the prescribed origins within a task condition and fitting seed, then over the three seeds. For comparator b and task condition c, define $w_{c,b} = 1$ if this task mean for STRIDE is smaller, 0 if larger, and $1/2$ for equality. Let $\mathcal{C}_{u,b}$ be the complete paired conditions in system/source group u, with U_b represented groups. The reported win fraction averages task indicators within each group, then weights groups equally:

$$\widehat{w}_b = \frac{1}{U_b} \sum_{u=1}^{U_b} \frac{1}{|\mathcal{C}_{u,b}|} \sum_{c \in \mathcal{C}_{u,b}} w_{c,b}. \tag{20}$$

The displayed task counts are unweighted: three wins from three conditions of one system and no win from one condition of another give $3/4$ task wins but a group-weighted fraction of $(1+0)/2 = 50\%$. The same aggregation is applied separately to whole-horizon Energy, late Energy, and the late slope. A score difference within 10^{-12} is a tie; there are no ties in the multi-step comparison plotted in Figure 2. Groups are resampled for the reported descriptive 95% intervals. This dimensionless comparison permits aggregation across differently scaled tasks without averaging their raw errors. It remains conditional on each comparator's complete paired cohort. A conservative global grouping merges recurring systems across sources; Table 13 reports that sensitivity. The merged inventory retains all 1,903 planned comparisons, including incomplete rows. Slope summaries use the paired tasks with both recorded slopes available. First- and third-order Gaussian HMMs, ARHMM, and continuous Gaussian Markov have 189 such tasks: the restaurant task has unavailable slopes under their complete-lead scoring rule. Whole- and late-path Energy retain all 190 paired multi-step tasks for these methods.

Figure 2 uses identical comparator-specific paired task sets in its whole-horizon and late-horizon panels. Its fourteen rows contain 1,777 complete comparisons out of 1,903 planned, drawn from the same union of 200 multi-step tasks. Markov–Gaussian, first- and third-order Gaussian HMMs, ARHMM, and continuous Gaussian Markov share 190 paired tasks from 120 system/source groups. Task counts and equal-group percentages are both reported in Table 11; they use different denominators. Group bootstrap uses 2,000 resamples and seed 20260922. All scores and task win classifications use frozen benchmark records. Single-step tasks remain in the whole-horizon benchmark tables.

Table 11: Task counts underlying Figure 2. Both endpoints use identical paired $H > 1$ tasks per comparator. Whole/late wins compare three-seed means of origin-averaged whole-horizon/final-half path Energy. Weighted percentages average win indicators within each system/source, then weight groups equally; ties count one half (10^{-12} tolerance). Missing comparisons remain planned. Counts are task conditions, not independent replications. Third-order HMM uses 190 paired of 200 planned multi-step tasks across all nine families.

Comparator	Paired/planned	Groups	Whole wins	Weighted (%)	Late wins	Weighted (%)
Markov–Gaussian	190/200	120	156	82.1	146	78.5
Continuous Markov	190/200	120	139	72.2	134	74.6
Gaussian HMM	190/200	120	153	80.5	143	76.4
Third-order HMM	190/200	120	145	76.3	140	76.0
ARHMM	190/200	120	139	72.7	137	73.6
Direct SINDy	114/125	73	90	77.3	89	77.0
VAR	189/200	120	122	64.3	123	65.6
NVAR	110/125	74	65	51.9	64	51.6
Context parroting	123/125	75	103	82.7	102	81.5
Seasonal naive	66/74	44	41	67.5	39	64.6
Holt grid	66/74	44	44	72.2	42	69.3
DLinear	72/81	47	44	77.7	44	77.3
DeepAR	15/18	12	8	66.7	8	66.7
TimePrism	72/81	47	35	55.5	36	52.5

STRIDE remains strongly competitive in late-horizon accuracy: its system-weighted win fractions against the five full-inventory state-space baselines range from 73.6% to 78.5%. Tables 12 and 13 report all fourteen forecasters and sensitivity to the global system/source grouping. Late error-growth slopes are a separate endpoint, with their own available-pair counts.

Table 12: Late-horizon comparisons on multi-step tasks. Energy and slope columns report equally weighted group win percentages with descriptive 95% group-bootstrap intervals (2,000 draws). n_E/u_E and n_β/u_β give paired tasks/groups for each endpoint. Gaussian HMM, third-order HMM, ARHMM and continuous Markov require every late step mean to be finite for the slope, excluding one restaurant task with missing observations; saved original slopes use available finite steps. Lower slopes can be negative. Rows use comparator-specific cohorts.

Comparator	n_E/u_E	n_β/u_β	Lower Energy (%)	Lower slope (%)
Markov–Gaussian	190/120	190/120	78.5 [71.6, 85.4]	52.0 [43.8, 60.6]
Continuous Markov	190/120	189/119	74.6 [66.9, 81.9]	55.5 [47.6, 64.1]
Gaussian HMM	190/120	189/119	76.4 [69.1, 83.6]	47.9 [39.8, 57.0]
Third-order HMM	190/120	189/119	76.0 [69.0, 83.1]	41.7 [34.1, 50.3]
ARHMM	190/120	189/119	73.6 [65.7, 81.4]	64.7 [56.6, 73.1]
Direct SINDy	114/73	114/73	77.0 [68.3, 85.7]	63.1 [53.4, 73.3]
VAR	189/120	189/120	65.6 [57.3, 73.3]	50.5 [42.3, 58.8]
NVAR	110/74	110/74	51.6 [41.1, 62.2]	34.4 [24.5, 44.9]
Context parroting	123/75	123/75	81.5 [72.5, 89.5]	65.6 [55.5, 75.8]
Seasonal naive	66/44	66/44	64.6 [50.0, 78.3]	51.0 [36.5, 65.2]
Holt grid	66/44	66/44	69.3 [55.7, 81.8]	34.8 [21.7, 49.0]
DLinear	72/47	72/47	77.3 [64.5, 87.9]	57.2 [43.1, 71.0]
DeepAR	15/12	15/12	66.7 [41.7, 91.7]	64.6 [39.6, 89.6]
TimePrism	72/47	72/47	52.5 [38.3, 65.7]	58.5 [43.6, 72.3]

Table 13: Sensitivity to merging recurring systems across benchmark sources; forecast scores and paired tasks are unchanged. Energy and slope columns report equally weighted group win percentages with descriptive 95% group-bootstrap intervals (2,000 draws). n_E/u_E and n_β/u_β give paired tasks/groups for each endpoint. Gaussian HMM, third-order HMM, ARHMM and continuous Markov require every late step mean to be finite for the slope, excluding one restaurant task with missing observations; saved original slopes use available finite steps. Lower slopes can be negative. Rows use comparator-specific cohorts.

Comparator	n_E/u_E	n_β/u_β	Lower Energy (%)	Lower slope (%)
Markov–Gaussian	190/116	190/116	78.1 [70.6, 85.0]	52.3 [43.9, 60.7]
Continuous Markov	190/116	189/115	74.1 [66.3, 81.5]	56.0 [47.7, 64.5]
Gaussian HMM	190/116	189/115	75.9 [68.0, 83.2]	48.2 [39.7, 56.8]
Third-order HMM	190/116	189/115	75.6 [67.7, 82.8]	41.7 [32.8, 50.4]
ARHMM	190/116	189/115	73.5 [65.5, 80.8]	64.9 [56.7, 73.2]
Direct SINDy	114/70	114/70	77.5 [68.4, 86.7]	63.9 [53.6, 74.3]
VAR	189/116	189/116	64.8 [55.8, 72.7]	50.8 [42.6, 59.6]
NVAR	110/71	110/71	53.2 [41.7, 64.2]	35.3 [25.7, 45.7]
Context parroting	123/72	123/72	81.2 [72.1, 89.5]	66.0 [55.5, 76.4]
Seasonal naive	66/43	66/43	64.0 [50.0, 76.7]	51.0 [37.0, 66.1]
Holt grid	66/43	66/43	68.7 [54.8, 81.4]	34.4 [21.1, 49.4]
DLinear	72/46	72/46	78.9 [66.9, 89.3]	56.3 [42.7, 70.5]
DeepAR	15/12	15/12	66.7 [41.7, 91.7]	64.6 [39.6, 89.6]
TimePrism	72/46	72/46	53.6 [39.5, 67.4]	57.6 [43.5, 70.7]

Full-inventory independent state-space comparisons. Markov–Gaussian, first- and third-order Gaussian HMMs, ARHMM, and continuous Gaussian Markov each use the prescribed splits, contexts, horizons, origins, and three protocol seeds on all 230 tasks from nine families. Each method fits its own dynamics and observation model. None receives STRIDE states, transition-change inputs, or local output functions. Across all horizons, each comparator has 220 complete paired tasks from 126 system/source groups; their common multi-step subset has 190 tasks from 120 groups. Table 14 gives both endpoints and their denominators. Against third-order HMM, STRIDE has lower Energy on 148 of 220 whole-horizon pairs and 140 of 190 late-horizon pairs.

Table 14: Five comparators evaluated on the same 230 task conditions in nine families. Whole-horizon results include $H = 1$: each comparator has 220 paired tasks across 126 system/source groups. Late Energy uses the same 190 paired $H > 1$ tasks across 120 groups (200 planned). Wins are unweighted task counts; percentages weight within-system/source win indicators equally across groups. Third-order HMM has fixed order three and shared diagonal Gaussian emissions.

Comparator	Whole n/N	Wins	Weighted (%)	Late n/N	Wins	Weighted (%)
Markov–Gaussian	220/230	158	78.5	190/200	146	78.5
Continuous Markov	220/230	142	69.2	190/200	134	74.6
Gaussian HMM	220/230	155	77.0	190/200	143	76.4
ARHMM	220/230	139	69.3	190/200	137	73.6
Third-order HMM	220/230	148	73.2	190/200	140	76.0

E Representation and trajectory-evolution evidence

E.1 Common protocol and matched module controls

Module controls measure trajectory evolution, transition decoding, and continuous prediction at their respective outputs, using three fitting seeds per condition. Within each task, the encoder and continuous-output functions are shared. Table 15 examines trajectory evolution on Aizawa and StocBench: both have multiple retained POD directions and frequently changing observed transition frequencies. The known three-dimensional generators in Appendix E.4 isolate coupling and memory under specified generating equations. The window-selection study additionally uses GARCH and Exchange under their original configurations.

Matched replacements. Same-q Markov uses empirical transitions of the existing encoder with smoothing 0.01, and the same local writer. No coupling restricts the deterministic map to each direction's own features while retaining the residual covariance estimator. No memory removes both explicit lagged nodes and all four filtered channels, using $p = 1$. These two variants keep the POD basis and common training endpoints and retrain the decoder under the same budget. Hard inverse preserves the original D sample array, replaces the learned readout with the recursive window inverse, and applies source-mass simplex projection at each step. Context-only decoding (No-D) leaves the upstream module unchanged but masks the first $2K^2$ dynamic feature channels consistently during training and prediction. The frozen historical boundary and step features remain available.

Table 15: Trajectory-dynamics case studies on Aizawa and StocBench. Both representations retain multiple POD directions and frequently changing transition targets. Scores are full-horizon D MSE (lower is better), averaged over the original three fitting seeds. The encoder, POD basis, eligible training endpoints, and forecast protocol are matched. No coupling removes cross-direction drift inputs; no memory removes finite lag features and filtered memory channels. Active (%) counts distinct observed targets with $\frac{1}{2}\|D\|_1 > 10^{-12}$. These are descriptive case-study means; known-operator controls are given in Table 20.

Task	r	Active (%)	STRIDE	No coupling	No memory
Aizawa	2	100.00	0.00501638	0.00528282	0.0797243
StocBench	15	99.30	0.00481715	0.00500528	0.00741082

Table 15 scores the predicted increment trajectory directly. Removing memory raises D MSE from 0.005016 to 0.079724 on Aizawa and from 0.004817 to 0.007411 on StocBench. Their active-target fractions are 100% and 99.30%, respectively, and their training active fractions are 100% and 99.28%. Thus these case studies contain changing transition mass for the dynamics module to predict. The reported means describe these fixed cases; the controlled-generator action and autonomous-response tests provide the dedicated coupling and memory comparisons. Appendix F evaluates learned inversion and local continuous output, and Appendix G evaluates their long-horizon use.

E.2 Window length and difference lag

The window study uses feasible local neighborhoods of $W \in \{1, 3, 7, 15, 31\}$ and $L \in \{1, 2, 4, 8, 16\}$. Each task holds its encoder, state count, rank, finite-history setting, and writer fixed. The POD basis is refitted for each window definition at the fixed rank. Training endpoints are shared using the largest eligible window/lag/history requirement, capped at 16,000 rows. Consequently, the reference is a matched-endpoint refit at the original (W, L), rather than the original main-study checkpoint. Only seed-1729 validation Energy selects a setting; all three seeds' test results are then reported. Table 16 reports the validation-selected setting and its test score for each of the four tasks. Figure 5 shows the full StocBench grid, with the same validation-selected cell marked in both panels. The grid documents parameter selection; it does not assume that validation selection minimizes the held-out test score.

Table 16: Window/difference-lag sensitivity. A validation-only choice (seed 1729) is evaluated with all three seeds. Both the selected configuration and its reference are refitted on the same eligible training endpoints with fixed encoder, writer and POD rank. The reference uses the original (W, L) and the common grid endpoints. The full 60-configuration grid is retained in the accompanying evidence.

Task	Reference (W, L)	Selected (W, L)	Reference ES	Selected ES
Aizawa	(7,8)	(3,8)	0.00679429	0.00682371
GARCH	(7,4)	(7,4)	0.742682	0.742682
Exchange	(7,8)	(7,16)	0.262738	0.262738
StocBench	(7,4)	(15,2)	0.713787	0.714841

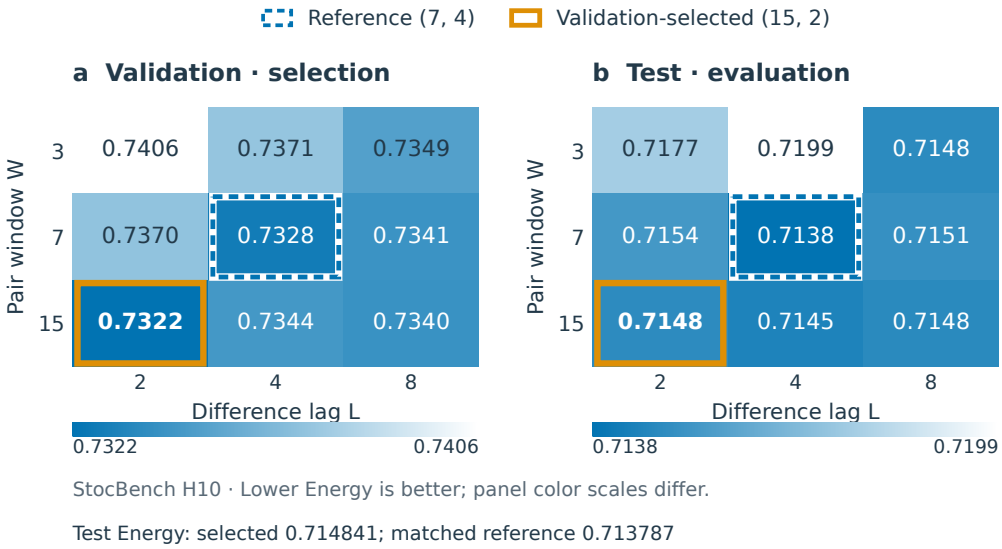


Figure 5: StocBench window/lag sensitivity. Each cell is path Energy, lower is better. Validation selects $(W, L) = (15, 2)$; the matched reference is $(7, 4)$. Both are refitted on identical eligible endpoints. The test panel evaluates this fixed choice and is not a selection surface.

E.3 True transition-frequency activity

Twelve task/horizon settings describe the transition statistics under their fixed encoders, windows, difference lags, seeds, and test origins. These settings specify the descriptive activity inventory under the declared representation.

Table 17: Frozen representations in the expanded diagnostic study. J is the causal encoder length; $N_D = C - (J-1) - W - L$ counts available historical change nodes. Hospital admissions uses $J = 3$ and therefore $N_D = p = 5$; the other encoders use $J = 9$. All rows satisfy $p \le N_D$. The last column is the percentage of distinct evaluated target timestamps with true $D = 0$, deduplicated across overlapping forecast windows. Horizons of one source are not independent datasets.

Task	C	J	K	W	L	r	p	N_D	$D = 0$ (%)
Weather H96	256	9	8	15	8	6	8	225	80.7
Weather H192	256	9	4	15	8	4	17	225	85.4
Weather H336	256	9	8	15	8	15	1	225	80.9
Jena weather H48	192	9	8	31	8	10	8	145	57.2
Hospital H12	48	9	2	7	8	2	17	25	18.8
Hospital admissions H13	12	3	8	1	4	37	5	5	3.3
Lorenz H15	75	9	2	3	2	2	8	62	0.0
Mackey-Glass H64	96	9	4	15	8	5	1	65	85.4
Exchange H96	256	9	2	7	8	1	1	233	98.8
StocBench H10	96	9	4	7	4	15	17	77	0.7
Aizawa H64	96	9	2	7	8	2	17	73	0.0
Aizawa H256	96	9	2	7	8	2	17	73	0.0

True transition tables are reconstructed with the frozen causal encoder. A node is active when $v_h = \frac{1}{2}\|D_{t+h}\|_1 > 10^{-12}$, and quiet otherwise. Overlapping forecast targets are deduplicated by recording and absolute timestamp within each task setting. Table 18 reports quiet fractions and change magnitudes over these distinct targets; it also gives the corresponding training statistics. No model predictions enter these activity measurements.

Table 18: Transition activity under each frozen representation. Quiet means $D = 0$ to tolerance 10^{-12}. Training percentages use all eligible training nodes; evaluation percentages and TV quantiles use distinct forecast target timestamps, deduplicated within recording. TV is $\frac{1}{2}\|F_t - F_{t-L}\|_1$. The final column counts active/all origins at $h = \lceil 0.8H \rceil$. These populations have different weights by definition.

Task	Train quiet (%)	Eval quiet (%)	TV median	TV 95th	Active / all
Weather H96	81.45	80.67	0.000	0.400	53/256
Weather H192	86.49	85.41	0.000	0.333	33/256
Weather H336	81.28	80.91	0.000	0.400	46/256
Jena weather H48	62.71	57.24	0.000	0.194	7/19
Hospital H12	6.98	18.78	0.286	0.714	209/256
Hospital admissions H13	2.08	3.31	1.000	1.000	122/128
Lorenz H15	2.60	0.00	0.667	0.667	2/2
Mackey-Glass H64	86.56	85.43	0.000	0.467	3/31
Exchange H96	99.64	98.81	0.000	0.000	2/256
StocBench H10	0.72	0.70	0.429	0.571	252/256
Aizawa H64	0.00	0.00	0.714	1.000	31/31
Aizawa H256	0.00	0.00	0.714	1.000	28/28

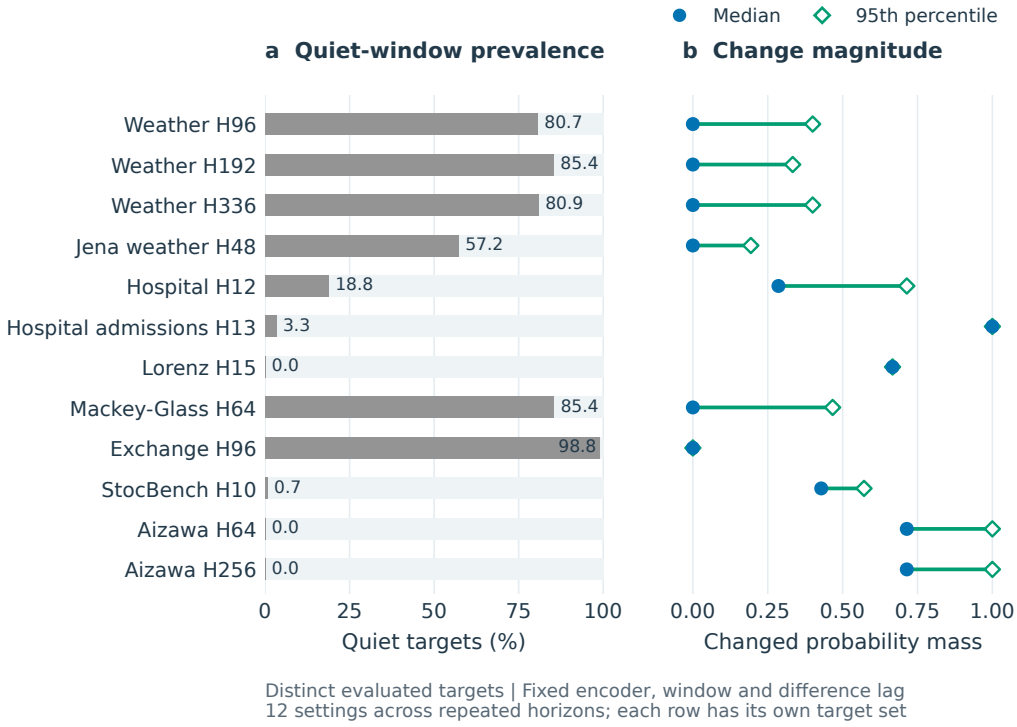


Figure 6: **How often transition frequencies change, and by how much.** Left: quiet fraction among distinct evaluated targets under each frozen representation. Right: median and 95th-percentile total-variation distance between the two lagged frequency windows. Weather combines long quiet periods with occasional appreciable changes; Aizawa supplies a continuously changing target. Each row is descriptive for its own (q, W, L) and forecast target set.

Exchange has 98.81% quiet evaluated targets and Mackey–Glass 85.43%. Weather has 80.67–85.41% quiet targets, with 95th-percentile change magnitude 0.33–0.40. Their training quiet fractions are 99.64%, 86.56%, and 81.28–86.49%, respectively. Thus infrequent change is already visible before evaluation. At $h = \lceil 0.8H \rceil$, Exchange has only 2 active origins out of 256 and Mackey–

Glass has 3 out of 31. These measurements characterize how much changing transition mass is present in the declared representation. Aizawa and StocBench instead contain many active targets.

Adaptation to the represented dynamics. The framework adapts its representation and finite-history length through training/validation selection (Appendix B). Exchange selects one POD direction and $p = 1$, while Aizawa selects two directions and $p = 17$, and StocBench selects 15 directions and $p = 17$. These recorded configurations allocate different directional and historical capacity to different tasks. A rank-one representation has no cross-direction interaction to remove, so coupling is tested using the multidirectional cases and the three-dimensional generators. The quiet-target score comparisons below assess behavior where the observed lag-window change is zero.

Quietness refers specifically to $F_{t+h} = F_{t+h-L}$. It need not mean equality of adjacent windows or constancy of the underlying continuous observations. The magnitude $v_h = \frac{1}{2}\|D_{t+h}\|_1$ is the total-variation distance between two empirical joint-frequency tables, so frequency of activity and size of change are reported separately.

Observed transition counts behind the trajectory. The example rule takes the first active origin in fixed order at the common 80% forecast position, independently of predictive performance. On Weather $H = 336$, origin 0:42833 has a continuously nonzero lag difference at leads 250–278, with 21 actual changes of adjacent frequency tables in 28 comparisons. At lead 269, $q_0 \to q_0$ changes from $5/15$ to $9/15$ relative to eight leads earlier. The five changing entries in Table 19 sum to zero, as required of a frequency difference.

Table 19: Audited Weather $H = 336$ example, origin 0:42833 and $h = 269$. Entries compare the observed pair counts in the current window and its $L = 8$ predecessor, with $W = 15$. Only changed entries are shown. The example is the first active origin in fixed order at the common 80% position, selected independently of prediction accuracy.

Pair	F_{t+h-8}	F_{t+h}	Change (pp)
$q_0 \to q_0$	5/15	9/15	+26.667
$q_0 \to q_4$	2/15	1/15	-6.667
$q_0 \to q_5$	2/15	1/15	-6.667
$q_4 \to q_5$	1/15	0/15	-6.667
$q_5 \to q_0$	3/15	2/15	-6.667

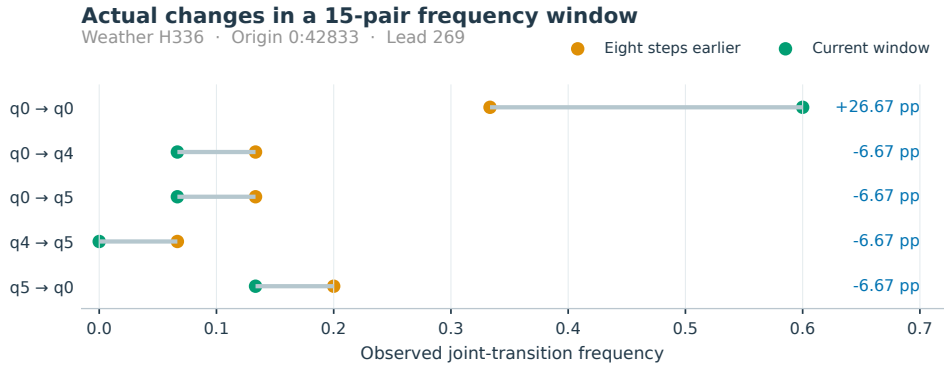


Figure 7: **A trajectory node is an observed change in joint-transition frequencies.** The Weather example shows all changing entries between two real windows. These are directly recounted from the true encoded sequence. The example illustrates the mass redistribution represented by one trajectory node.

Other audited active segments include Weather $H = 96$, origin 0:42222 at leads 73–96 (19/23 adjacent changes), Weather $H = 192$, origin 0:42580 at leads 152–163 (4/11), Aizawa $H = 64$,

origin 0:8000 at leads 1–64 (49/63), and StocBench, origin 0:106 at leads 1–10 (9/9). A zero-difference counterexample is StocBench origin 33:166 at lead 8: $F_8 = F_4$ while $F_8 \neq F_7$, and neighboring D values are nonzero. By contrast, Weather $H = 336$, origin 0:42187, has no adjacent frequency changes over leads 210–336. These examples distinguish lag-window equality from a continuously unchanged frequency path.

E.4 Coupling and memory in known trajectory dynamics

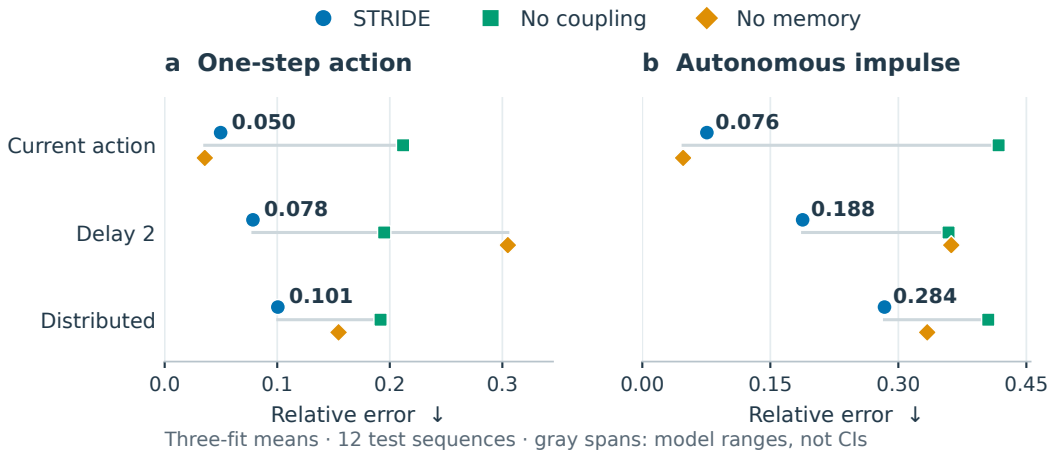


Figure 8: **Known-operator controls isolate coupling and memory.** The two delayed systems benefit from both components; the current-only system provides a control with no generating memory term. Left: held-out one-step action error. Right: autonomous impulse-response error. Points average three fitting seeds. All three generating mechanisms and matched controls are shown. These directly observed zero-sum trajectories evaluate the dynamics module.

Controlled generators. These controls isolate the POD/sparse-dynamics module from discrete encoding and window inversion. Let $z_t \in \mathbb{R}^3$ and observe $d_t = Vz_t \in \mathbb{R}^{16}$, where the columns of V are orthonormal and orthogonal to the all-ones vector. Generate

$$z_{t+1} = Az_t + Bz_{t-1} + \sum_{j=1}^{4} M_j m_t^{(j)} + \xi_t, \quad \xi_t \sim \mathrm{Uniform}([-0.035, 0.035]^3). \tag{21}$$

The base current operator is

$$A_0 = \begin{pmatrix} 0.68 & 0.16 & 0 \\ -0.13 & 0.64 & 0.08 \\ 0 & -0.09 & 0.60 \end{pmatrix}. \tag{22}$$

The current-only system uses $A = A_0$, $B = M_j = 0$. The two-step predictor uses $A = 0.65A_0$ and

$$B = \begin{pmatrix} 0.20 & 0.05 & 0 \\ 0 & 0.20 & 0.03 \\ 0.04 & 0 & 0.18 \end{pmatrix}, \qquad M_j = 0. \tag{23}$$

Here the two predictor states are z_t and z_{t-1}; the latter is two sample steps before target z_{t+1}. The distributed-memory system uses $A = 0.70A_0$, $B = 0$,

$$M_1 = \begin{pmatrix} 0.13 & 0.03 & 0 \\ 0 & 0.12 & 0.03 \\ 0.02 & 0 & 0.10 \end{pmatrix}, \quad M_3 = 0.12I, \quad M_2 = M_4 = 0, \tag{24}$$

with memory channels ordered as the two stages at 0.6 followed by the two stages at 0.9. Each mechanism has 24 training, eight validation, and 12 test sequences of length 256. Training/validation initial states are uniform in $[-0.2, 0.2]^3$, while test initial states use $[-0.5, 0.5]^3$. The initial previous

state is zero and all memory stages start at z_0. A deterministic name-derived seed based on 20260922 fixes generation.

Because $V^\top V = I$, an error in generating coordinates has the same Euclidean norm after lifting: $\|V(\widehat{z} - z)\|_2 = \|\widehat{z} - z\|_2$. The trajectory occupies a known three-dimensional subspace of the zero-sum hyperplane $\{d \in \mathbb{R}^{16} : \mathbf{1}^\top d = 0\}$, whose dimension is 15. We fit its POD directions from training observations, and score predictions in the original trajectory coordinates.

POD rank is three, full-model history length is eight, and forecast horizon is 32. Controls use the same data and row endpoints, and the fit sees only observed trajectories. Truth operators are transformed into fitted POD coordinates solely for evaluation. We measure held-out one-step conditional action, augmented partial operators, impulse responses, and zero-innovation rollouts against the generator's conditional means. Correlated feature channels can make separate coefficients less identifiable than their combined action, so coefficient error alone is not used as proof of recovery. The one-step test uses indices 95–254 of each of the 12 held-out sequences and compares the predicted action with the known conditional next-trajectory mean. Zero-innovation rollouts start at origins 96, 128, 160, and 192 in each sequence, giving 48 continuations per fit. The learned predictor and generating equations then evolve for 32 steps with future innovations set to zero. Because the generating system is linear with zero-mean innovations, its zero-innovation continuation is its conditional mean. The learned nonlinear predictor is evaluated against that reference; its zero-innovation rollout is not asserted to equal its own stochastic conditional mean. For the impulse test, each of the three generating-coordinate axes receives an initial perturbation of 10^{-4} with zero earlier state and memory. The learned zero-input response is subtracted before normalizing by the perturbation. Relative impulse error compares the complete 32-step response tensor with the known response. These metrics therefore evaluate future trajectory continuation, in addition to matching an individual local action.

Table 20: Known-operator trajectory prediction. Action is relative RMSE of the held-out one-step conditional action; impulse is autonomous impulse-response relative error; future mean is autonomous conditional-mean MSE over the 32-step horizon. Values are lower-is-better means across three fits. Action and future mean use 12 held-out sequences; the impulse test uses three basis perturbations. The same training endpoints, POD rank, and data apply to the matched variants. These controls evaluate the POD/dynamics module in the known zero-sum trajectory space.

Mechanism	Model	Action	Impulse	Future mean MSE
Current action	STRIDE	0.049706	0.07566	2.8154×10^{-7}
Current action	No coupling	0.2118	0.41752	5.1907×10^{-7}
Current action	No memory	0.035628	0.047833	2.4641×10^{-7}
Two-step	STRIDE	0.078414	0.18785	1.3729×10^{-7}
Two-step	No coupling	0.19501	0.35884	3.5809×10^{-7}
Two-step	No memory	0.30487	0.36201	4.7923×10^{-7}
Distributed memory	STRIDE	0.10053	0.28377	1.9062×10^{-7}
Distributed memory	No coupling	0.19179	0.40534	3.9025×10^{-7}
Distributed memory	No memory	0.15448	0.33384	9.8532×10^{-7}

Figure 8 reports the operator-action and impulse-response errors. The table additionally reports zero-innovation rollout MSE against the generator's conditional future mean for all three mechanisms and all three fitted variants. These predictions evolve autonomously after their observed prefix; no future trajectory nodes are supplied.

The saved paired action-MSE intervals resample the 12 held-out sequences after averaging the three fits within each sequence. For the two-step generator, ablation-minus-full intervals are $[1.0309, 1.1810] \times 10^{-6}$ without coupling and $[2.9464, 3.0562] \times 10^{-6}$ without memory; for distributed memory they are $[0.9786, 1.0799] \times 10^{-6}$ and $[0.4474, 0.6131] \times 10^{-6}$. These descriptive 95% intervals quantify the matched-generator effects. The current-action control, whose generating memory is zero, favors the simpler no-memory fit, with interval $[-0.07487, -0.06257] \times 10^{-6}$; removing coupling instead gives $[2.3064, 2.5761] \times 10^{-6}$. Thus the controls distinguish a generating dependence from additional model capacity.

E.5 FORECASTING OBSERVED CHANGES AND TRANSITION PROBABILITIES

On Aizawa, the context-only minus full-model differences have paired 95% intervals $[0.20950, 0.22328]$ for state Brier, $[0.23174, 0.25305]$ for joint Brier, and $[0.02821, 0.03093]$ for frequency MSE. These support a transition-layer contribution on the measured task. Continuous-path Energy evaluates the subsequent decoded distribution and local dynamics. The activity-stratified comparisons below complement this full-horizon test; the independent descriptive activity measurements are given in Appendix E.3.

Forecast scores are stratified using the true-data activity definition in Appendix E.3. The zero group compares equal windows separated by L; it need not have $F_{t+h} = F_{t+h-1}$. Errors are averaged over the appropriate origins separately at each lead, with seed pairing preserved. An origin may change groups over the forecast horizon. The percentages in Table 21 divide by the comparator error. Near-zero denominators below 10^{-12} are not converted into a percentage. The absolute errors remain in the source evidence.

Table 21: Twelve settings at the common position $h = \lceil 0.8H \rceil$, conditioned on true $D \neq 0$. n_a/n counts active/all origins at that step. Entries are percentage D-prediction error reductions relative to zero change and last-change persistence. * requires both the simultaneous band and conditional sign-flip test to support STRIDE after full-family correction; † similarly supports the comparator. Unmarked numbers are point estimates.

Task	h	n_a/n	D/zero	D/last
Weather H96	77	53/256	-0.05	+20.04
Weather H192	154	33/256	-0.59	+4.58
Weather H336	269	46/256	+0.67	+18.76
Jena weather H48	39	7/19	-0.16	+64.37
Hospital H12	10	209/256	+5.22	+59.29*
Hospital admissions H13	11	122/128	-89.11	+5.06
Lorenz H15	12	2/2	+71.94	+81.29
Mackey-Glass H64	52	3/31	-0.49	-0.49
Exchange H96	77	2/256	+4.08	+4.08
StocBench H10	8	252/256	+20.20*	+71.52*
Aizawa H64	52	31/31	+97.75	+99.43
Aizawa H256	205	28/28	+97.36	+97.39

Simultaneous inference. The primary stratified analysis uses 9,999 paired seed-and-recording/time-block bootstrap draws, with chronological span $C + H$ and a $2(C + H)$ sensitivity. Each draw recomputes stratum means and relative effects. A studentized max-T band covers all leads in the final 40%, both strata, seven comparisons, and five active-minus-zero effect contrasts within each task. Error probability is then allocated across the 17 settings. The seven comparisons are continuous Energy, continuous MSE, checked-frequency MSE, pair Brier, state Brier, D MSE against zero, and D MSE against persistence.

An additional max-T block sign-flip test conditions on the three fitted seeds, enumerating signs for at most 16 blocks and using 9,999 draws otherwise. Support requires both the globally corrected absolute-effect band to exclude zero in the favorable direction and the corrected sign-flip $p < 0.05$. These time-block procedures assume approximately independent blocks and, for the sign test, joint sign symmetry. Only three fitted seeds and a small number of blocks limit inferential precision. They do not imply independent observations at every origin or lead.

For active nodes, StocBench supports improvement over zero D at leads 6–9 and over persistence at 6–10. Aizawa $H = 64$ supports a pair-Brier improvement at lead 39, leads 42–53 and 55–56, and leads 60 and 64. Hospital $H = 12$ and StocBench $H = 10$ have supported D-forecast comparisons. The complete interval registry retains all 17 settings, original test outcomes, both block choices, and the original correction family. Earlier unstratified tests use a different family; their markers are not transferred into these stratified tables.

Quiet-target equivalence and conditional contribution. The same stratified analysis evaluates whether the forecast-frequency inputs materially change single-step Energy on quiet targets. Equiv-

alence requires the entire simultaneous relative-effect band to lie inside a declared margin, using the recorded $\pm1\%, \pm3\%, \pm5\%$ bounds and the unchanged full-family correction. Table 22 uses the same common 80% lead as the active-node summaries. The reported bound is met under both block lengths. Across leads, Exchange satisfies the $\pm1\%$ bound at $h = 58$–96, and Weather $H = 336$ satisfies the $\pm3\%$ bound at $h = 202$–336. Weather also meets $\pm1\%$ at $h = 299$, 302–314, 316–319, and 321–324. With its frequency inputs retained, the full model therefore preserves comparable quiet-target scores within these measured margins. This behavior complements the active-target transition improvements and the separately fitted long-horizon forecasts in Appendix G.

Table 22: Quiet-stratum score equivalence at the common $h = \lceil 0.8H \rceil$ used for the active-node summaries. Scores are single-step Energy, averaged over quiet origins and three fits; n_q counts quiet origins per fit. Context-only masks the forecast-frequency mean and standard deviation during both fitting and prediction. Relative difference is $100(\mathrm{ES}_{\mathrm{context}} - \mathrm{ES}_{\mathrm{STRIDE}})/\,\mathrm{ES}_{\mathrm{context}}$. The displayed band uses the primary $C + H$ blocks and the existing full-family correction. The indicated equivalence bound contains the entire relative band under both $C + H$ and $2(C + H)$ blocks. The bounds come from the recorded $\pm1\%, \pm3\%, \pm5\%$ analysis margins; equivalence is determined by the band, not by a nonsignificant superiority test.

Task	h	n_q	STRIDE	Context-only	Relative band (%)	Bound
Exchange $H = 96$	77	254	0.3426208	0.3426209	[-0.000729, 0.000828]	$\pm1\%$
Weather $H = 336$	269	210	0.4845507	0.4845833	[-1.1702, 1.1836]	$\pm3\%$

For a fixed lead h, let $\delta_{i,h}$ be the context-only minus full-model single-step Energy of paired evaluation record i. With normalized protocol weights w_i over origins and fits and true active indicator $A_{i,h}$, set $p_h = \sum_i w_i A_{i,h}$. Renormalizing these weights within each stratum gives

$$\Delta_{\mathrm{all}}(h) = p_h \Delta_{\mathrm{active}}(h) + (1 - p_h)\Delta_{\mathrm{quiet}}(h). \tag{25}$$

Partitioning $\sum_i w_i \delta_{i,h}$ by $A_{i,h}$ proves the identity; an empty stratum contributes zero. Thus an overall ablation effect depends on both the conditional effects and the prevalence of each regime. The observed active-stratum effect must be measured alongside its prevalence. This identity concerns single-step scores; whole- and late-horizon path Energy score their joint subpaths directly.

E.6 Predicting increments versus frequency levels

Why model $D_t = F_t - F_{t-L}$ rather than F_t directly? We compare these targets in three selected case studies, evaluating the frequency forecasts supplied to the inverse readout. STRIDE recovers $\widehat{F}^{\mathrm{raw}}$ by Equation (5); the direct-F control forecasts frequency levels using the same POD–SINDy family, cross-direction inputs, and memory kernels. Both branches share the frozen encoder, observed prefix, training prediction endpoints, forecast origins, rank, history length, and sparse-fitting penalties. These settings retain the original validation-selected configuration. The direct-F branch fits its own training POD basis, coordinate scaling, coefficients, and residual covariance at these fixed settings.

Weather and Hospital each use 256 test origins, with $(C, H, K, W, L, r, p) = (256, 96, 8, 15, 8, 6, 8)$ and $(48, 12, 2, 7, 8, 2, 17)$, respectively. A separate Aizawa cohort uses 64 independent initial-condition draws (seed 2026092602), 512 burn-in steps, 96 context observations, and 1024 future steps, with $(K, W, L, r, p) = (2, 7, 8, 2, 17)$. It retains the original system, sampling interval, normalization, encoder, and already fitted $H = 64$ upstream operators; all 64 trajectories are retained. This cohort differs from the $H = 256$ continuous-output study in Appendix G.2. All three tasks use fitting seeds $1729, 2718, 3141$ and $S = 32$ upstream paths.

The reference F is recomputed from true future observations using the frozen causal encoder and the trailing transition windows in Equation (1). We score the mean of the 32 predicted frequency paths against this reference, averaging squared errors over the K^2 cells, indicated leads, test cases, and fits. This is pre-readout F_{raw} MSE, without added clipping or projection; checked-frequency MSE and continuous-output Energy are separate endpoints.

Task	Leads	STRIDE	Direct F	Lower mean (gain)	95% interval for Δ
Weather	49–96	**0.001317**	0.001616	STRIDE (18.51%)	$[-0.000314, 0.000846]$
Hospital	7–12	**0.050208**	0.141230	STRIDE (64.45%)	$[0.073981, 0.108463]$
Aizawa	1–64	0.004444	**0.002964**	Direct F (33.29%)	$[-0.001676, -0.001268]$
Aizawa	129–256	0.004408	**0.003761**	Direct F (14.68%)	$[-0.001061, -0.000207]$
Aizawa	257–512	**0.004911**	0.007869	STRIDE (37.59%)	$[0.002034, 0.003917]$
Aizawa	513–1024	**0.006146**	0.025657	STRIDE (76.05%)	$[0.017466, 0.021657]$
Aizawa	769–1024	**0.006579**	0.032885	STRIDE (79.99%)	$[0.023720, 0.028831]$

Table 23: Increment versus level prediction. Both columns score F_{raw} MSE (lower is better); STRIDE predicts D and recovers F. $\Delta = \text{MSE}(\text{direct } F) - \text{MSE}(\text{STRIDE})$. Reduction divides the mean-error difference by the larger mean, with the lower-mean branch named explicitly. Bold identifies the smaller mean, not an inferential decision. All Aizawa rows use the same new 64-trajectory cohort.

Weather's late-window mean is 18.51% lower with STRIDE, although its paired interval includes zero; Hospital's is 64.45% lower with a positive interval. On the same new Aizawa trajectories, direct F is better at leads 1–64 and 129–256, whereas STRIDE is better at 257–512 and later. In the prespecified primary interval 513–1024, STRIDE lowers frequency MSE by 76.05%. Each fit's mean favors STRIDE, as do 64/64 trajectories after averaging the three fits. The late MSE slope is 3.6153×10^{-6} per lead for STRIDE versus 5.6394×10^{-5} for direct F, a 93.59% reduction; the paired slope-difference interval is $[4.8140, 5.7481] \times 10^{-5}$. The increment target yields substantially lower long-range frequency error and slower late error growth on this cohort.

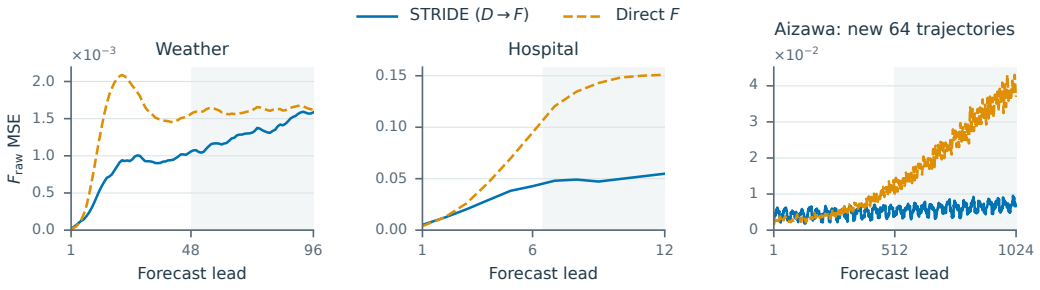


Figure 9: Frequency-prediction error at every saved lead, averaged over test cases and three fits without smoothing. Shading marks the reported late window for Weather and Hospital and the primary 513–1024 interval for the new Aizawa cohort. Each panel has its own vertical scale; lines are descriptive means.

Pointwise 95% intervals retain the original paired bootstrap: 2,000 draws using overlapping-origin blocks for Weather and source-series clusters for Hospital; 4,000 draws over fits and paired trajectories for Aizawa. Slopes fit per-lead MSE by ordinary least squares. Aizawa windows reuse one cohort, with 769–1024 nested within 513–1024; fits and windows are not independent replications. No multiplicity correction is applied.

F Inverse decoding and local continuous output

F.1 Matched readout with fixed upstream predictions

The learned-versus-hard inverse comparison preserves the complete predicted D ensemble and inverse-difference recurrence. On StocBench, origin-averaged Energy is 0.713792 for the learned readout and 0.730442 for recursive inversion, a 2.28% reduction. The recording-equally weighted paired hard-minus-learned effect has interval $[0.01507, 0.01771]$. Figure 10 places this readout comparison alongside the Aizawa state-probability control from Appendix E.5.

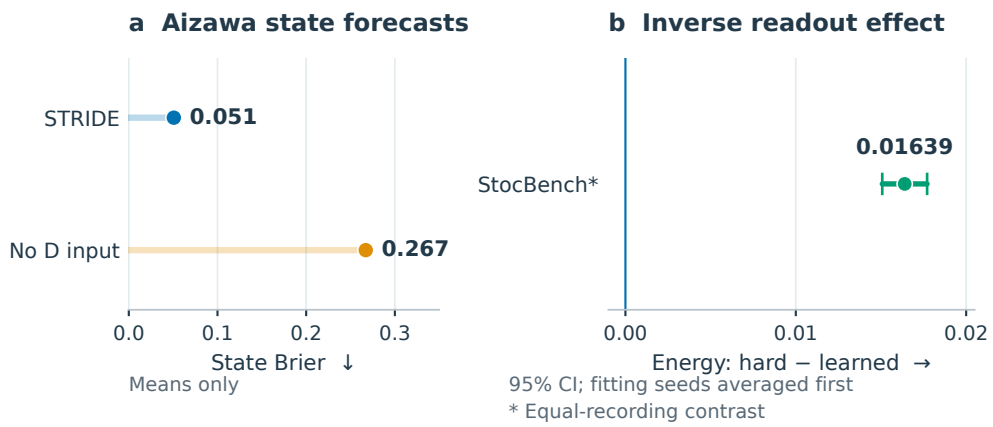


Figure 10: Matched transition-readout controls. Left: Aizawa state Brier for full and context-only decoding with the same encoder and writer. Right: the StocBench hard-minus-learned inverse Energy effect with fixed upstream D samples and a seed-conditional 95% interval, weighting recordings equally.

F.2 Accuracy of the frozen inverse mapping

Each independent one-step diagnostic starts with true history. The original encoder maps true observations to q, then to the true pair and frequency tables. True D and lagged anchors reconstruct the current F; the decoder receives its original first-step input format and frozen weights, including the true source state and boundary tables. No upstream predictor, path sampler, or local writer is called. The network has no separate continuous-observation input. A label error is the disagreement between its deterministic argmax state and the true next state.

Table 24: Frozen decoder accuracy on Aizawa at two horizons with independently refreshed true history and true D/F inputs at every step. Labels are deduplicated by absolute timestamp within each seed. Accuracy is the argmax hit rate; brackets are seed-and-time-block 95% intervals. Brier and mean probability of the true state additionally evaluate the probabilities. Exact algebraic inversion under the same true inputs and boundaries has 100% accuracy in both settings. These are privileged diagnostics, not forecasting leaderboard entries.

Task / H	K	Times	Accuracy (%) [CI]	Brier	$p(q_{\text{true}})$
Aizawa64	2	1984	95.21 [95.1, 95.3]	0.056644	0.90475
Aizawa256	2	1984	92.02 [90.4, 95.2]	0.093842	0.83906

Within each seed, overlapping evaluations of the same absolute timestamp are deduplicated. Repeated calls have maximum probability difference zero. Counts in the table identify unique times per seed; intervals resample the three fitting seeds and time blocks, rather than treating repeated times as independent. The separate fixed-origin diagnostic retains the original along-horizon convolution history, time features, and marginal propagation. Its original-marginal error rates on Aizawa are 1.915% at $H = 64$ and 2.265% at $H = 256$; the complete per-task records are retained in the evidence. It answers a different question from refreshing true history at every node.

The exact inverse checks recover F to maximum error 2.64×10^{-16} and B to 1.56×10^{-14}, with zero state-label errors. Manual softmax and marginal projection match the original interface, and perturbing future features leaves earlier outputs unchanged. These privileged-input measurements evaluate the frozen inverse readout under the specified histories; autonomous forecast performance is measured by the matched full-chain controls.

An additional equal-capacity decoder is trained on true future frequency windows as a diagnostic. Relative to the full chain, pair Brier changes from 0.09527 to 0.04609 on Aizawa in the matched

diagnostic fit. This fit uses privileged windows and is not entered into the forecasting comparison. Frozen-decoder true-F inputs with zero or original predicted standard deviations are retained as separate conditions in the evidence.

F.3 Local-function regression

Table 25: Local-function fit with true preceding continuous values and true state pairs. RMSE uses training-standardized observation coordinates. Test rows are distinct forecast target transitions; these teacher-forced diagnostics do not replace autonomous forecast scores.

Task	Train RMSE	Val RMSE	Test RMSE	Shared RMSE
Aizawa	0.000105478	0.00010031	0.000110649	0.000158124

The local-function test supplies the two true preceding observations and the true state pair. Validation targets exclude the training-history prefix; test transitions are deduplicated from the frozen forecast targets. Pair-specific support, acceleration errors, and R^2 are retained in the evidence. Table 25 reports Aizawa's local and shared-function errors on the training, validation, and held-out transitions.

F.4 Continuous-path distribution diagnostics

Tables 26 and 28 report the full model's continuous-output scores under the frozen test protocol, using three fitting seeds and 32 sampled paths per origin. Energy and fair CRPS assess forecast accuracy, while variograms measure temporal and cross-coordinate dependence.

Table 26: Full-model distribution checks on four focused tasks, averaged over three fitting seeds and prescribed origins. Energy and fair CRPS are computed in training-standardized observation coordinates; lower is better. Coverage is the empirical fraction inside the ensemble 90% marginal interval, and width is measured in the same coordinates. These are descriptive means; no interval from a different metric is reused.

Task	Energy	Fair CRPS	Coverage (%)	Width
Aizawa	0.00689284	0.00454724	62.73	0.0143431
GARCH	0.743643	0.609243	75.45	2.5428
Exchange	0.262738	0.191132	81.54	0.858993
StocBench	0.713792	0.568127	79.96	2.67196

Aizawa error and interval scale. A companion sampling study retains the Aizawa task, encoder, coordinate dynamics, and local functions, with the readout fitted under the original configuration. It evaluates 31 origins, three fitted seeds, and four Monte Carlo repetitions per sample count. Table 27 places coverage alongside the size of the prediction errors and intervals in training-standardized observation coordinates. With 8,192 forecast paths, RMSE is 0.01118, the mean 90% interval width is 0.01563, and its mean half-width is 0.00781; empirical coverage is 67.42%. Thus both the error and interval width are on the order of 10^{-2} in these coordinates.

Table 27: Sampling scale in the companion full-chain Aizawa study ($H = 64$). RMSE is the square root of pooled ensemble-mean MSE; width and half-width summarize the empirical 90% marginal intervals. All three use training-standardized observation coordinates.

Samples S	RMSE	Width	Half-width	Coverage (%)
32	0.01122	0.01422	0.00711	62.60
8192	0.01118	0.01563	0.00781	67.42

Table 28: Full-model dependence diagnostics. VS denotes the variogram score with exponent $1/2$; temporal comparisons use lags 1, 2, 4 and 8 when available, and channel comparisons use at most 256 fixed channel pairs. Fair versions remove the finite-ensemble variance term. A one-step task has no temporal pair, and a scalar task has no channel pair; undefined scores are shown as dashes. Scores are descriptive three-seed means.

Task	Temporal VS	Fair temporal VS	Channel VS	Fair channel VS
Aizawa	5.7413×10^{-5}	5.6797×10^{-5}	0.00016582	0.0001633
GARCH	–	–	0.14791	0.14382
Exchange	0.013933	0.013601	0.047644	0.046702
StocBench	0.15568	0.15132	0.18139	0.17636

Marginal intervals use the 0.05 and 0.95 empirical quantiles with NumPy's linear interpolation: for S sorted draws, quantile u interpolates at zero-based index $(S-1)u$. Coverage is the observed fraction of targets within these ensemble intervals, reported together with interval width.

G Matched long-horizon Aizawa case study

Two matched $H = 256$ experiments assess whether the transition trajectory improves long-horizon continuous prediction on the same fixed system. The first uses the original chronological test sequence; the second uses independent draws of perturbed initial conditions with a separately locked fitting protocol. Their sampling units and fitting protocols are distinguished below.

G.1 Chronological test-sequence experiment

The Aizawa $H = 256$ case freezes the original encoder, sparse operator, and local functions, and fits full and context-only decoders using matched long-horizon windows. No D masks the two forecast-frequency channels (mean and standard deviation) consistently during fitting and evaluation. The upstream D arrays remain fixed. The original nine-contrast correction family is retained; the extended horizon is not an additional independent system.

Table 29: Aizawa long-horizon decoder case study at $H = 256$. No D retains context and masks the forecast-frequency mean and standard-deviation channels during both decoder training and prediction. Energy is recomputed on the final half-path. Reduction is $(\text{No } D - \text{STRIDE})/(\text{No } D)$. Slopes are OLS fits of step Energy within the final half; the displayed slopes are multiplied by 10^3. This extended horizon uses its separately recorded protocol.

Task / H	STRIDE	No D	Reduction (%)	$10^3\beta$: STRIDE	No D
Aizawa 256	0.0907158	0.221461	59.04	0.72485	2.2392

Table 30: Paired No-D minus STRIDE inference. Intervals resample seeds and recording/time blocks; p_E, p_β are block sign-flip tests conditional on the three fits, with Holm correction over nine comparisons per endpoint. A positive bootstrap interval and a non-significant conditional test may coexist. b is the number of candidate blocks. A single block gives no interval.

Task / H	Energy difference CI	p_E	Slope difference CI	p_β	b
Aizawa 256	[0.07666, 0.1781]	0.4375	[0.0008756, 0.00209]	0.4375	5

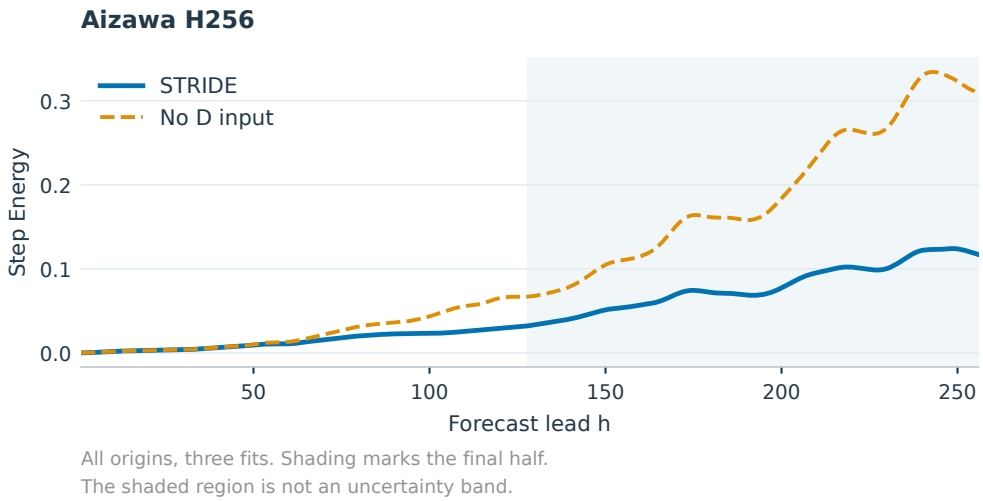


Figure 11: **Per-lead error under matched decoder ablations.** Means use all test origins and three fitting seeds. The highlighted region is the final half used for the late score and slope. The Aizawa $H = 256$ curves compare full and context-only decoders under matched long-horizon fitting. Lines are descriptive means, not confidence bands. The numerical joint-subpath scores and their inference are reported separately in Tables 29–30.

The paired intervals resample fitting seeds and independent recordings or non-overlapping chronological origin blocks of span $C + H$. A shorter H-block sensitivity is retained for this study. The block sign-flip tests first average the three fitted seeds, so their randomness differs from the seed-and-block bootstrap. Holm adjustment is applied separately to the nine late-Energy and nine slope tests. For Aizawa $H = 256$, five blocks limit the minimum two-sided exact sign-flip value to $2/2^5 = 0.0625$ before correction. The positive bootstrap interval therefore accompanies a non-significant corrected conditional test. No interval is inferred when only one candidate block is available.

G.2 Independent-initial-condition replication

The preceding experiment uses five chronological time blocks. This complementary experiment uses one newly generated trajectory as its sampling unit, with one forecast origin per trajectory. The system, sampling interval, protocol, trajectory count and four primary contrasts are fixed before generating any new test data. Three fitting seeds are $1729, 2718, 3141$. The original upstream models, encoders, normalization and writers are frozen. STRIDE, No-D, and history-direct readouts are fitted afresh under the same long-horizon protocol.

Controls with matched available history. No D masks the predicted frequency mean and standard deviation in the causal readout during both fitting and evaluation. History direct receives the entire available state prefix: each historical input contains a one-hot state, its causal one-hot joint transition, trailing F, and availability masks. A width-16 GRU summarizes this prefix. Its final hidden vector augments the context-only readout features ($W - 1$ boundary transitions, L frequency anchors, last observed state q_t, and time). The same width-16 causal decoder blocks and probability-flow projection produce future B directly. All methods share the frozen local continuous writer.

The duration control conditions on current state and observed dwell age, censored by the prefix. Age-bin upper boundaries are $1, 2, 3, 4, 6, 9, 13, 19, 28, 42, 63, 95$, followed by an overflow bin. Training counts shrink toward the same-q transition estimate; validation joint Brier selects prior strength from $\{1, 10, 100\}$. Conditional means propagate the augmented state/age distribution, and sampled paths retain the corresponding age dependence.

Selection before new test generation. For each seed, all neural readouts use the same 256 training windows and all 28 original validation origins that admit $H = 256$ (the prespecified cap was 32). Training uses 1,600 updates and validation full-horizon joint-Brier checkpoint selection every 50 updates. STRIDE and No-D each have 5,876 readout parameters; history direct has 7,444. The duration model is selected on the same validation targets. Hashes of all three completed fitting records are locked before test generation; no new test outcome changes training or settings.

The 64 initial conditions are independent draws from the fixed training-based distribution: uniformly sampled training indices 100–5899 plus Gaussian perturbations of 0.01 training standard deviations, using seed 2026092404. Each trajectory has 512 burn-in steps, 96 context observations and 256 forecast steps. The original Aizawa right-hand side is integrated with DOP853, relative tolerance 10^{-11} and absolute tolerance 10^{-12}, at the calibrated interval 0.0861319465. Training-only one-step checks agree with the source trajectory to below 6×10^{-13}. Independence is conditional on this fixed system and initial-condition distribution; the trajectories are not independent physical systems.

Method	Whole	Late half	STRIDE−control	95% interval	Holm p
STRIDE	0.065654	0.090483	–	–	–
No-D	0.134121	0.186657	-0.096174	$[-0.126918, -0.067056]$	0.0004
Same-q	0.146580	0.203570	-0.113087	$[-0.154254, -0.076306]$	0.0004
History direct	0.135540	0.188623	-0.098140	$[-0.129477, -0.068654]$	0.0004
Duration	0.085995	0.119340	-0.028857	$[-0.040447, -0.017730]$	0.0004

Table 31: Independent-initial-condition Aizawa replication at $H = 256$. Lower Energy is better. The primary endpoint is late-half path Energy (leads 129–256); the difference and interval columns concern this endpoint only. Each of 64 newly generated trajectories contributes one origin, evaluated with three fits and 32 paths. Four paired seed-conditional sign-flip tests use Holm correction; 9,999 random flips give a minimum unadjusted reported $p = 0.0001$. Intervals are descriptive paired fit/trajectory bootstrap intervals, not simultaneous intervals. This is one fixed system and initial-condition distribution, not 64 independent systems.

Long-horizon improvement under matched controls. Late-half path Energy is 0.0904831 for STRIDE, compared with 0.1866571 for No-D, 0.2035699 for same-q Markov, 0.1886235 for history direct and 0.1193397 for duration. The corresponding relative reductions are 51.5%, 55.6%, 52.0% and 24.2%. All four contrasts pass the fixed Holm family, with adjusted $p = 0.0004$ at the resolution of the random sign-flip test. Tests average the three fits before flipping trajectory-level paired differences and assume sign symmetry under the null. Bootstrap intervals resample fits and trajectories conditional on the selected configurations. These results demonstrate a long-horizon continuous-prediction benefit over all four controls on independently initialized Aizawa trajectories.

Method	q Brier	B Brier	F^{check} MSE
STRIDE	0.043368	0.083800	0.001964
No-D	0.441240	0.523273	0.061122
Same-q Markov	0.499006	0.584523	0.070402
History direct	0.441182	0.523223	0.061110
Duration	0.153393	0.230608	0.012634

Table 32: Intermediate forecast scores on the same 64 independently initialized Aizawa trajectories as Table 31, averaged over the three prescribed seeds. All methods share the frozen encoder/writer; lower is better. Scores average the full $H = 256$ horizon, with F^{check} obtained from decoded transition tables and observed boundary pairs. These are saved descriptive means; the four primary late-Energy tests do not provide inference for these metrics.

The same trajectories also have lower full-horizon state Brier, joint-transition Brier and checked-frequency MSE for STRIDE than for each control (Table 32). For example, joint Brier is 0.083800 for STRIDE and 0.230608 for the duration control. These descriptive measurements connect intermediate prediction accuracy and continuous-path benefit within one fixed cohort. The prespecified inferential comparisons concern late-half Energy.

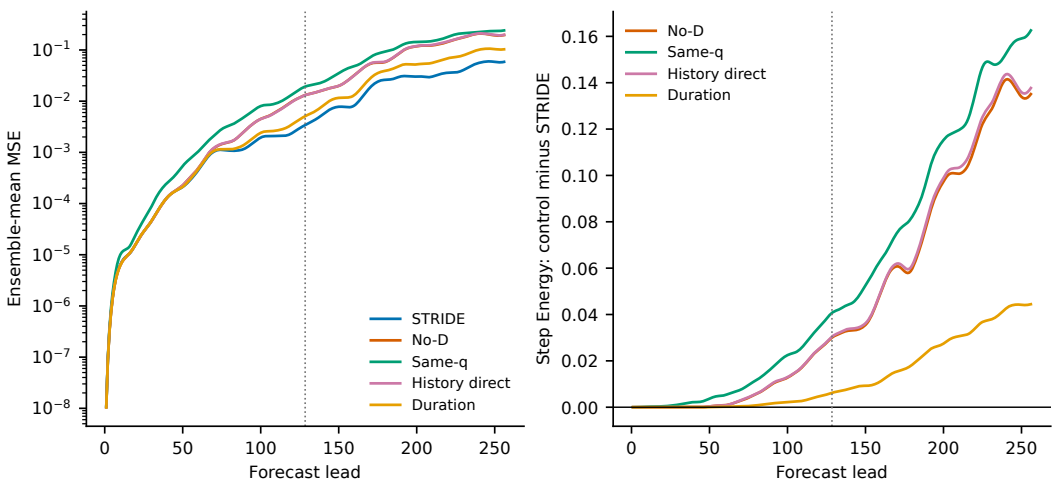


Figure 12: Secondary lead-wise diagnostics on the same 64 new Aizawa trajectories. Left: ensemble-mean MSE on a logarithmic axis. Right: mean per-step Energy of each control minus STRIDE. The dotted line starts the primary late-half subpath. Curves average trajectories and fits, with no per-lead significance claims. Per-step Energy is not the joint late-subpath endpoint in Table 31.

Reduced long-horizon error growth. STRIDE's mean MSE rises from 0.0001338 over the first 64 leads to 0.041996 over the last 64. The corresponding No-D, same-q, history-direct and duration final-quarter MSEs are 0.159339, 0.189953, 0.161620 and 0.076726. All four first-to-last-quarter MSE-growth contrasts favor STRIDE, with descriptive paired intervals retained in the evidence. These secondary measurements describe reduced error accumulation over the evaluated horizon and complement the primary late-Energy comparisons.

H REPRODUCIBILITY ARTIFACTS

The source package separates the following reproducibility materials.

1. `evidence/` contains frozen aggregate scores, comparison cohorts, paired-test records, configuration audits, and table generators. `SOURCE_MANIFEST.json` records packaged file hashes. The manuscript and figures compile directly from these files.
2. `reference_implementation/` contains the frozen benchmark encoder, sparse dynamics, readout, local functions, selection, scoring, data adapters, and command-line entry point. Its own manifest records original file hashes. `TASK_REGISTRY.json` identifies the 230 evaluated tasks, with environment-specific path prefixes replaced by `EXPERIMENT_ROOT`; `environment_frozen.txt` records the benchmark software environment.
3. The independent-initial-condition study's recorded protocol, selected fit settings, aggregate scores, and original source hashes are in `evidence/closure_followup.json`. Its separate simulation/training scripts and fitted archives are external artifacts.

The README maps each entry point to its role. Table and figure generation uses existing scores. Full training replay additionally requires the original benchmark inputs, baseline repositories, fitted artifacts, and the corresponding study-specific environment and scripts.