FORECASTING POST-CHECKPOINT TRAINING WITH MOVING-FRAME DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how far the future trajectory of a training run can be forecast from its current checkpoint. Because learned features continue to evolve, coordinates fixed at the checkpoint become progressively inaccurate. We introduce a moving-frame predictor that transports the checkpoint-conditioned mean and covariance with the evolving representation, using optimizer-induced feature motion to update the frame. This models the forecastable component of the continuation. We then establish a complementary, method-independent limit: the irreducible state-prediction MSE induced by post-checkpoint stochasticity is locally linear in the forecast horizon, whereas the additional error from freezing feature evolution admits a local quartic bound. The resulting error budget explains how inaccuracies accumulate over a completed forecast. Experiments on end-to-end multilayer perceptrons and Transformers for next-token prediction on real text verify the predicted error structure and show that the moving-frame predictor substantially outperforms checkpoint-fixed and learning-curve extrapolation baselines. Together, these results separate forecast error that can be reduced by modeling representation dynamics from uncertainty intrinsic to the future training path.

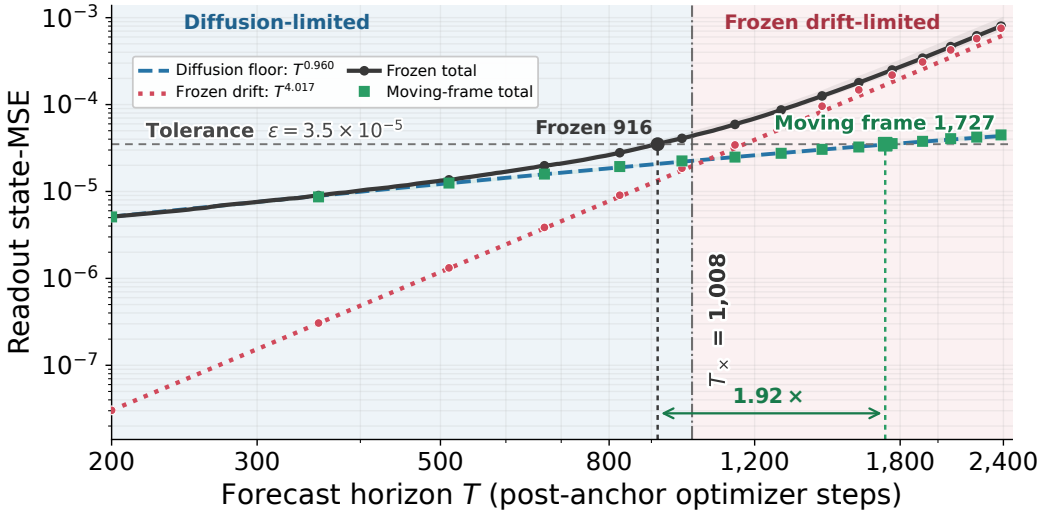


Figure 1: **What limits post-checkpoint predictability?** At short horizons, prediction error is dominated by post-checkpoint diffusion, whereas at longer horizons feature drift limits the Frozen predictor, a checkpoint-fixed NTK approximation. We show that the diffusion floor is intrinsic to causal state prediction: future minibatch innovations are unavailable at the checkpoint and therefore cannot be predicted from its information, leaving an irreducible state-MSE that is locally $O(T)$ even for the exact conditional mean. We also show that freezing feature evolution adds an avoidable excess state-MSE bounded locally as $O(T^4)$. Across 16 end-to-end tanh MLP problems, the empirical exponents (0.960 and 4.017) match the predicted $O(T)$ and $O(T^4)$ scalings and exhibit the theoretically predicted crossover between the two error regimes of the Frozen predictor. Moving-frame dynamics reduce this avoidable error and extend the useful prediction horizon by a median $1.92\times$ over Frozen.

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.

1 INTRODUCTION

A central challenge in deep-learning optimization is to characterize the dynamics of learning. We study this problem through the predictability of loss curves, a standard summary of learning dynamics. Specifically, for a particular training run, we ask how far its future trajectory can be forecast from the state and statistical information available at an intermediate checkpoint, and what aspects of the future are intrinsically unpredictable due to post-checkpoint stochasticity.

We focus on forecasting the continuation of a training run from an intermediate checkpoint. Let $\mathcal{F}_t$ denote the information observed through time t. At checkpoint t_0, this information includes the target dataset D, the model and optimizer configuration, the current state, and the training history. The checkpoint forms the anchor from which we predict the subsequent loss trajectory. To quantify uncertainty, we consider replicated runs with different initializations, data orders, and minibatch draws, and propagate the conditional mean and covariance of their possible continuations.

An end-to-end network changes its feature representation during training, so coordinates fixed at the checkpoint quickly become outdated. Our key idea is to move the coordinates with the representation. Starting from the checkpoint, a full-batch update predicts the next feature Gram matrix and how its feature frame rotates. We then transport the readout mean and covariance into that updated frame before taking the next forecast step. This separates two effects that fixed coordinates mix together: motion of the readout within the current representation and motion of the representation itself. The *Frozen* baseline, i.e., a checkpoint-fixed NTK approximation, keeps the checkpoint frame fixed, while our operator-form update remains well defined even when eigenvalues cross.

The geometry yields a two-regime predictability picture. Post-checkpoint stochasticity creates an irreducible $O(T)$ state-prediction error, even for the exact conditional mean. Freezing the feature frame adds an avoidable excess state-MSE from feature drift, which grows locally as $O(T^4)$ and eventually limits the forecast. Figure 1 previews this transition from a diffusion-limited regime to a feature-drift-limited regime and shows how moving-frame transport extends the useful forecast horizon.

This leads to three contributions:

- We derive a checkpoint-conditioned stochastic process in an optimizer-induced moving feature frame. It propagates the conditional mean and replica covariance while accounting explicitly for feature-frame motion, and its ambient-operator implementation remains regular through eigenvalue crossings.
- We characterize two complementary aspects of forecast reliability: a retrospective decomposition of an upper bound on local prediction error and an intrinsic causal predictability limit induced by post-checkpoint stochasticity. Locally, irreducible diffusion scales as $O(T)$, whereas freezing feature evolution incurs an additional $O(T^4)$ state-MSE penalty.
- We validate the mechanism on end-to-end multilayer perceptrons and character-level Transformer training on real text. The experiments verify that the learned-feature error budget contains the endpoint error, identify its dominant components, isolate the benefit of frame updating, and report run-level calibrated coverage. On both model classes, the moving-frame predictor improves over a frozen-frame reference and learning-curve extrapolation baselines.

1.1 SETUP

To estimate the checkpoint-conditioned mean and covariance used by the forecast, we run M training replicas that share the target dataset and objective but differ in initialization, data order, or minibatch sampling. This ensemble includes anchor-state variation; the same-state causal limit in Theorem 1 instead conditions on one observed anchor. The feature frame is evaluated on a common clean central path initialized at the replica mean, not generally on the evolving population-mean path (Lemma 1); every replica readout uses that frame. Write $\Phi_t \in \mathbb{R}^{P\times d}$ for this feature matrix with P rows and d coordinates, and let $\Sigma_t := \Phi_t^\top \Phi_t / P = V_t \Lambda_t V_t^\top$, where V_t contains its eigenvectors and Λ_t its eigenvalues. If $a_{m,t}$ is replica m's readout, then $c_{m,t} := V_t^\top a_{m,t}$ is its instantaneous spectral coordinate. At anchor t_0, let $\mathcal{G}_{t_0}^{\mathrm{shared}} \subseteq \mathcal{F}_{t_0}$ contain the dataset and common anchor information but not a generic replica's state or innovations. Define $\mathbb{E}_m[\cdot] := \mathbb{E}[\cdot \mid \mathcal{G}_{t_0}^{\mathrm{shared}}]$, with Var_m and Cov_m

defined analogously, and write

$$\bar{c}_t = \mathbb{E}_m[c_{m,t}], \qquad \Sigma_{\mathrm{mem},t} = \mathrm{Cov}_m(c_{m,t}). \tag{1}$$

Observed replicas estimate these population moments by the sample mean and the $1/(M-1)$ sample covariance. Their finite-M fluctuations determine how precisely the checkpoint distribution is known; Appendix C.1 records the corresponding sampling conventions.

For a replica-common quadratic evaluation measurable at the anchor, let $E^{\mathrm{loc}}_{\mathrm{test}}(t)$ be the conditional local test loss and $L_{\mathrm{mean},t}$ its loss-at-the-mean term. Then

$$E^{\mathrm{loc}}_{\mathrm{test}}(t) = L_{\mathrm{mean},t}(\bar{c}_t) + \mathrm{Tr}(C_{\mathrm{te},t}\Sigma_{\mathrm{mem},t})\,, \tag{2}$$

where $C_{\mathrm{te},t}$ is one half of the local loss Hessian in c. Expected loss is thus loss at the mean plus a curvature-weighted variance penalty. For random common coefficient paths, apply this decomposition conditionally along each path, then average (Appendix B.4).

At the anchor we measure $(\bar{c}_{t_0}, \Sigma_{\mathrm{mem},t_0}, \Lambda_{t_0}, \Omega_{t_0}, Q_{t_0})$ together with the model and optimizer state, where $\Omega_{t_0} := V_{t_0}^\top \dot{V}_{t_0}$ is the frame connection and Q_{t_0} is the refreshed-source covariance. Let R_{clean} denote the full-batch objective shared by the replicas after removing replica-specific target perturbations and refreshed minibatch noise. Its gradient supplies the central trajectory, Ω_{t_0} rotates the coordinates along that trajectory, and Q_{t_0} spreads the replica cloud around it. We call a quantity *causal* when it is available at the anchor (together with completed calibration runs); quantities measured after the endpoint are used to explain the forecast error. Appendix C.1 gives the full conditioning and finite-ensemble conventions. We employ the following assumption and Lemma 1; Appendix B.1 provides a proof.

Assumption 1 (A0: local regularity and replica sampling)**.** On each forecast window, Φ_t is twice differentiable; R_{clean} is three times continuously differentiable; its third derivative and the required coefficient paths are locally bounded and Lipschitz. The parameter neighborhood is convex and contains the replica states and their mean; state deviations have finite third moments. Generic-replica innovations are centered and satisfy the stated conditional martingale/bracket hypotheses. Write $B_3 := \sup \|\nabla^3 R_{\mathrm{clean}}\|$ on that neighborhood and let L_3 be a Lipschitz constant of $\nabla^3 R_{\mathrm{clean}}$. Replica-fixed heterogeneity is represented separately. The finite-ensemble rates use conditionally i.i.d. replicas with independent driving noises and the usual conditional variance and maximal bounds; Appendix C.1 states the variants based on zero cross-covariances.

Lemma 1 (Replica averaging: the mean follows the clean drift)**.** *For replica parameters θ_m, set $\bar{\theta} := \mathbb{E}_m[\theta_m]$ and $\delta\theta_m := \theta_m - \bar{\theta}$. Under A0, suppose each replica follows the common clean-gradient drift with learning-rate scale η plus a perturbation with zero conditional mean given its current state and $\mathcal{G}^{\mathrm{shared}}_{t_0}$. Then*

$$\dot{\bar{\theta}} = -\eta \nabla R_{\mathrm{clean}}(\bar{\theta}) - \frac{\eta}{2}\nabla^3 R_{\mathrm{clean}}(\bar{\theta}) : \Sigma_\theta + O\big(\eta\, \mathbb{E}_m \|\delta\theta_m\|^3\big)\,, \qquad \Sigma_\theta = \mathrm{Cov}(\theta_m \mid \mathcal{G}^{\mathrm{shared}}_{t_0}). \tag{3}$$

Here ":" denotes covariance contraction. Writing $\sigma^2 = \mathbb{E}_m\|\delta\theta_m\|^2$, the clean-path mean-drift bias is at most $\eta B_3 \sigma^2/2$; the displayed curvature correction leaves at most $\eta L_3 \mathbb{E}_m\|\delta\theta_m\|^3/6$, which is $O(\eta\sigma^3)$ under a uniformly bounded standardized third absolute moment. At each fixed checkpoint, the empirical mean has an additional $O_p(M^{-1/2})$ sampling error. Here Σ_θ is the full parameter covariance; replica-fixed heterogeneity contributes a separate conditional mean-gradient term.

2 The Local Replica Process After an Anchor

2.1 Fixed Frame: Mean and Replica Variation

With fixed features, every mode can be solved independently. Let Φ have Gram matrix $\Sigma = V\Lambda V^\top$. A replica target is $y_m = y + \xi_m$, where y is fixed by the dataset and ξ_m is a centered replica-fixed perturbation. Minibatch sampling adds a second, temporally refreshed source with covariance Q. With learning rate η and readout weight decay λ, mode k has one-step contraction $\rho_k = 1 - \eta(\lambda_k + \lambda)$ toward $g_k + n_k^m$, the clean and replica-specific spectral components of its fixed point.

Under the centered, uncorrelated-source conditions detailed in Appendix B.1, with shared $c_{0,k}$, $|\rho_k| < 1$, and $\nu_k = \mathrm{Var}_m(n_k^m)$, the scalar recursion after $r \geq 0$ optimizer steps has moments

$$\bar{c}_k(r) = \rho_k^r c_{0,k} + g_k(1-\rho_k^r), \qquad \mathrm{Var}_m[c_{m,k}(r)] = \nu_k(1-\rho_k^r)^2 + \underbrace{\textstyle\sum_{j=0}^{r-1} \rho_k^{2j} Q_{kk}}_{\text{refreshed-noise accumulation}}. \tag{4}$$

Replica-fixed target variation therefore has quadratic onset, whereas refreshed noise has linear onset. In gradient flow, ρ_k^r becomes $e^{-\eta(\lambda_k+\lambda)t}$ and the geometric sum becomes the corresponding Ornstein–Uhlenbeck (OU) integral, preserving the same two growth mechanisms in continuous time. Appendix B.1 gives the projection formulas.

Replica-fixed target noise changes each replica's limiting point, whereas refreshed minibatch noise continually broadens the distribution. The former is a persistent latent displacement and is propagated together with its cross-covariance; the latter appears as the source term in the Lyapunov equation below. This distinction explains the quadratic and linear onsets in (4).

2.2 Moving Frame: the Connection from the Optimizer Step

In an end-to-end network, the features themselves change. A coefficient can therefore change for two reasons: the predictor moves within the current basis, or the basis itself rotates. Ignoring the second effect introduces artificial discontinuities between spectral modes.

Let X be the evaluation inputs, $\bar{\theta}(t)$ the clean mean-parameter path, and $a(t)$ its readout, so $\Phi_t = \phi_{\bar{\theta}(t)}(X)$, $\Sigma_t = V_t \Lambda_t V_t^\top$, and $c(t) = V_t^\top a(t)$. Differentiation introduces the connection $\Omega_t = V_t^\top \dot{V}_t$ and moving-frame derivative $\nabla_t^{\mathrm{mov}} c := \dot{c} + \Omega_t c$. Writing g_t for the target projection and I for the identity, basis motion gives

$$\nabla_t^{\mathrm{mov}} c = -\eta(\Lambda_t + \lambda I)(c - g_t), \qquad (\Omega_t)_{pq} = \frac{v_p(t)^\top \dot{\Sigma}_t v_q(t)}{\lambda_q(t) - \lambda_p(t)} \quad (p \neq q). \tag{5}$$

The denominator in the component expression is singular when two eigenvalues coincide, although the underlying operator evolution and predicted loss remain regular. The implementation, therefore, evaluates the transport through the ambient matrix functions of Σ_t in Procedure 1; optional grouped-projector diagnostics use a separate classical Davis–Kahan estimate.

The frame motion follows directly from the optimizer step. For activations $\Phi_t = \phi_\theta(X)$ at a twice differentiable cut and parameter step $\Delta\theta_t$, first-order expansion with parameter Jacobian $J_\phi(\theta)$ gives

$$\Delta\Phi = J_\phi(\theta)\Delta\theta_t + O(\|\Delta\theta_t\|^2), \qquad \Delta\Sigma = \frac{\Delta\Phi^\top \Phi + \Phi^\top \Delta\Phi}{P} + O(\|\Delta\Phi\|^2), \tag{6}$$

and, for distinct eigenvalues, $(\Omega_t)_{pq} = v_p^\top \dot{\Sigma}_t v_q / (\lambda_q - \lambda_p)$ in continuous time. Thus the architecture and optimizer determine the accessible feature motion, while the residual determines its drive; Appendix B.2 gives the feature-kernel factorization and the contribution of regularization below the cut. Appendix B.2 also derives this rotation law from the chain rule and the optimizer update.

Combining spectral relaxation, frame transport, and stochastic forcing gives the local replica process

$$dc_{m,t} = \big(A_t c_{m,t} + b_t\big)dt + G_t\, dW_{m,t}, \tag{7}$$

where $W_{m,t}$ are independent standard Brownian motions, $K_t = \eta(\Lambda_t + \lambda I)$, g_t is the instantaneous target projection, $A_t = -K_t - \Omega_t$, $b_t = K_t g_t$, and $Q_t = G_t G_t^\top$ is the refreshed-source covariance. Exposing b_t is essential because feature motion may change g_t even when A_t is constant. On forecast interval $\mathcal{I}$, condition on $\mathcal{H}_\mathcal{I} := \mathcal{G}_{t_0}^{\mathrm{shared}} \vee \sigma\{A_s, b_s, G_s : s \in \mathcal{I}\}$ and write $\bar{c}_t^{\mathcal{H}}$ and $\Sigma_{\mathrm{aff},t}$ for the affine-process mean and covariance. Then

$$\dot{\bar{c}}_t^{\mathcal{H}} = A_t \bar{c}_t^{\mathcal{H}} + b_t, \qquad \dot{\Sigma}_{\mathrm{aff},t} = A_t \Sigma_{\mathrm{aff},t} + \Sigma_{\mathrm{aff},t} A_t^\top + Q_t. \tag{8}$$

Let $U(r,s)$ be the propagator of A, defined by $\partial_r U(r,s) = A_r U(r,s)$ and $U(s,s) = I$. If, after conditioning on $\mathcal{H}_\mathcal{I}$, $d\mathcal{M}_t = G_t\, dW_t$ is a centered square-integrable martingale with bracket $d\langle\mathcal{M}\rangle_t = Q_t dt$, then (8) integrates to

$$\begin{aligned} \Sigma_{\mathrm{aff},s} &= U(s,t_0)\Sigma_{\mathrm{aff},t_0} U(s,t_0)^\top + \int_{t_0}^{s} U(s,r) Q_r U(s,r)^\top dr, \\ \mathrm{Cov}(c_{m,s} \mid \mathcal{G}) &= \mathbb{E}[\Sigma_{\mathrm{aff},s} \mid \mathcal{G}] + \mathrm{Cov}(\bar{c}_s^{\mathcal{H}} \mid \mathcal{G}), \qquad \mathcal{G} = \mathcal{G}_{t_0}^{\mathrm{shared}}. \end{aligned} \tag{9}$$

When the coefficient path is already determined by $\mathcal{G}$, $\Sigma_{\mathrm{aff},s} = \Sigma_{\mathrm{mem},s}$. Random coefficient paths add the second-line term: replicas then vary both within a path and through variation of the path-conditioned mean. Appendix B.3 derives (9) and explains the two conditioning levels.

Finite-M estimators fluctuate around these population moments as described in Section 1.1, while curvature shifts the population mean through Lemma 1. Because Ω_t is antisymmetric, $\mathrm{Tr}(\Omega_t \Sigma_{\mathrm{aff},t} + \Sigma_{\mathrm{aff},t}\Omega_t^\top) = 0$: rotation preserves total variance instantaneously, but reorients it toward modes with different damping and test curvature. Substituting the transported moments into (2) produces the forecast studied next.

Procedure 1 Moving-frame checkpoint forecast

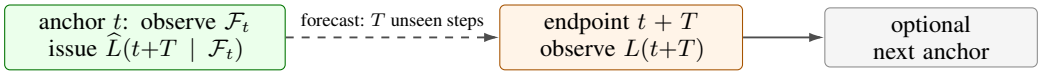
Require: Anchor t; M registered replicas; inputs X; clean objective R_{clean}; horizon $T = N$ optimizer steps

1: Estimate finite-M sample moments $\bar{\theta}_t^{[M]}$, $\bar{a}_t^{[M]}$, and $\Sigma_{\mathrm{mem},t}^{[M]}$
2: $\Phi_t \leftarrow \phi_{\bar{\theta}_t^{[M]}}(X)$; $\quad (V_t, \Lambda_t) \leftarrow \textsc{EigenDecompose}(\Phi_t^\top \Phi_t / P)$
3: $\bar{c}_t \leftarrow V_t^\top \bar{a}_t^{[M]}$; $\quad$ estimate g_t and one-step innovation covariance Q_t^{step}; $\quad \rho_t \leftarrow I - \eta(\Lambda_t + \lambda I)$
4: $\widehat{\theta}_N \leftarrow \textsc{FullBatchPropagate}(\bar{\theta}_t^{[M]}, R_{\mathrm{clean}}, N)$ using the training optimizer
5: $\widehat{\Phi}_{t+T} \leftarrow \phi_{\widehat{\theta}_N}(X)$; $\quad (\widehat{V}_{t+T}, \cdot) \leftarrow \textsc{EigenDecompose}(\widehat{\Phi}_{t+T}^\top \widehat{\Phi}_{t+T} / P)$
6: $R_{0,N} \leftarrow \widehat{V}_{t+T}^\top V_t$
7: $m_N \leftarrow \rho_t^N \bar{c}_t + (I - \rho_t^N) g_t$
8: $S_N \leftarrow \rho_t^N \Sigma_{\mathrm{mem},t}^{[M]} (\rho_t^N)^\top + \sum_{r=0}^{N-1} \rho_t^r Q_t^{\mathrm{step}} (\rho_t^r)^\top$
9: $\widetilde{m}_N \leftarrow R_{0,N} m_N$; $\quad \widetilde{S}_N \leftarrow R_{0,N} S_N R_{0,N}^\top$
10: $\widehat{L}(t+T \mid \mathcal{F}_t) \leftarrow \textsc{EvaluateEndpointLoss}(\widehat{\Phi}_{t+T}, \widetilde{m}_N, \widetilde{S}_N)$
11: **return** $\widehat{L}(t+T \mid \mathcal{F}_t), \widetilde{S}_N$

Repeated eigenvalues use ambient matrix polynomials; Appendix B.4 gives implementation details.

3 Prediction-Time Guarantees and Retrospective Error Analysis

Proposition 1 retrospectively bounds local forecast error, Theorem 1 identifies the unavoidable diffusion floor, and Proposition 2 bounds Frozen's avoidable excess. They share the affine process (7) but answer diagnosis and predictability questions separately. For a generic window, t is the observed anchor, the forecast is issued once using $\mathcal{F}_t$, and its target is the loss at $t+T$ after T optimizer steps.



The dashed segment denotes a forecast from the anchor to its endpoint. Rolling prediction re-anchors at each observed endpoint.

3.1 Retrospective Error-Bound Decomposition

At anchor t, the causal predictor, using only information available at the anchor, advances the replica mean under the clean gradient, transports the covariance through the predicted feature frame, and evaluates the resulting moments in test space. A measured-endpoint reference repeats the final evaluation in the observed future frame; its difference from the causal predictor isolates error attributable to frame prediction. Proposition 1 traces the remaining error through four objects in one aligned chart: the true continuous trajectory, a predicted continuous reference, its interleaved discrete recursion, and the endpoint-split implementation.

Writing hats for predictions, use Euclidean/operator $\|\cdot\|_2$ and Frobenius $\|\cdot\|_F$ norms. Appendix B.4 defines a C^1 Procrustes alignment; a superscript $\sharp$ denotes a predicted quantity expressed in that common chart. Define the integrated coefficient defects

$$\epsilon_A = \int_t^{t+T} \|A_s - \widehat{A}_s^\sharp\|_2 ds, \quad \epsilon_b = \int_t^{t+T} \|b_s - \widehat{b}_s^\sharp\|_2 ds, \quad \epsilon_Q^{\mathrm{int}} = \int_t^{t+T} \|Q_s - \widehat{Q}_s^\sharp\|_F ds, \tag{10}$$

and the evaluation defects $\epsilon_\Phi = \sup_s \|\Phi_s - \widehat{\Phi}_s\|_F$ and $\epsilon_{\mathrm{sub}} = \sup_s \|V_s V_s^\top - \widehat{V}_s \widehat{V}_s^\top\|_2$. Here, ϵ_Φ measures error in the predicted feature values, while ϵ_{sub} measures error in the predicted feature subspace. The accumulated defects ϵ_A and ϵ_b respectively quantify errors in the dynamics operator and affine forcing, and $\epsilon_Q^{\mathrm{int}}$ quantifies error in the stochastic covariance injected along the path.

Finally, ϵ_{disc} accounts for discretization error introduced by numerical integration and endpoint-split transport; Appendix B.4 defines its four mean/covariance components.

Proposition 1 (A posteriori local error bound)**.** *Consider a continuous horizon $T = Nh$, with N steps of size h, under one field that makes the common coefficient and evaluation paths measurable. Assume the martingale, bracket, and conditioning conditions stated with* (9)*, initialize all four objects described above from the same anchor moments, and suppose that the true and predicted frames have the same rank and a C^1 Procrustes alignment. If the reference moments and propagators are bounded and the local evaluation map is Lipschitz in its feature, subspace, mean, and covariance arguments, then there exist finite nonnegative constants $C_\Phi, C_{\text{sub}}, C_A, C_b, C_Q$ such that, pathwise under that conditioning,*

$$|\widehat{E}^{\text{loc}}_{\text{test}}(t+T) - E^{\text{loc}}_{\text{test}}(t+T)| \le C_\Phi \epsilon_\Phi + C_{\text{sub}}\epsilon_{\text{sub}} + C_A\epsilon_A + C_b\epsilon_b + C_Q\epsilon^{\text{int}}_Q + \epsilon_{\text{disc}} =: \mathcal{B}_{\text{loc}}. \tag{11}$$

The constants depend only on the stated local bounds and feature/readout/test norms. If the conditioning field is finer than $\mathcal{G}^{\text{shared}}_{t_0}$, outer conditioning and conditional Jensen give the same bound averaged over that field. Measured after the window, the five model defects and the discretization term reveal which mechanism consumed the error budget; anchor-time envelopes for them turn the same inequality into a forecast-time certificate.

Figure 2 isolates the geometric source of the frame terms in Proposition 1. Architecture Jacobians determine the directions in which features can move, the optimizer preconditioner sets their relative speed, and residuals provide the driving signal. The resulting frame rotation changes the coordinates of both the mean and covariance. Separately, the full parameter covariance supplies the mean-curvature correction in Lemma 1, while Σ_{mem} tracks dispersion in the readout frame. Appendix B.4 supplies the conditioning, alignment, recursion, and Duhamel details for Procedure 1 and proves Proposition 1.

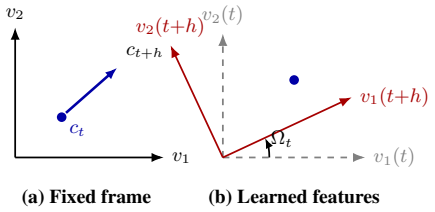


Figure 2: Moving coordinates: Ω_t rotates the basis over one step h, while $\nabla^{\text{mov}}_t c$ removes the induced coordinate change.

For a completely frozen reference, each predicted coefficient stays at its anchor value. Differentiability then gives $\epsilon_A \le \frac{1}{2}T^2 \sup_s \|\dot{A}_s\|_2$, $\epsilon_b \le \frac{1}{2}T^2 \sup_s \|\dot{b}_s\|_2$, and $\epsilon^{\text{int}}_Q \le T \sup_s \|Q_s - Q_t\|_F$. A first-order feature extrapolator similarly gives $\epsilon_\Phi \le \frac{1}{2}T^2 \sup_s \|\ddot{\Phi}_s\|_F$. Procedure 1 reduces these terms by following a time-varying frame. Its operator defect measures drift of both spectral relaxation K_s and frame rotation Ω_s through their combined generator $A_s = -K_s - \Omega_s$.

For an exact nonlinear ensemble evaluation, define $\epsilon_{\text{eval}} = |E^{\text{exact}}_{\text{test}} - E^{\text{loc}}_{\text{test}}|$. The triangle inequality then separates total endpoint error into the local-dynamics budget above and a measured local-to-network gap. The latter collects feature–readout interactions, persistent-shift cross-covariance, anchor sampling error, and finite-step products between frame and readout increments, such as $\Delta V^\top \Delta a$. This separation makes a completed forecast mechanistically interpretable: the first part diagnoses propagation, while ϵ_{eval} measures the fidelity of the local quadratic description.

3.2 Forecast Timing and Run-Level Calibration

For reported window j, write t_j for the anchor, T_j for its horizon, $\widehat{L}^{\text{mf}}_j := \widehat{L}(t_j + T_j \mid \mathcal{F}_{t_j})$ for the moving-frame forecast, $\widehat{L}^{\text{fr}}_j$ for its Frozen counterpart, and $L_j := L(t_j + T_j)$ for the endpoint. After the preceding window ends, $e_{j-1} := |L_{j-1} - \widehat{L}^{\text{mf}}_{j-1}|$ is known. The moving–Frozen disagreement measures feature motion, the squared horizon ratio rescales the previous error, and fixed padding $\delta_{\text{pad}} > 0$ keeps the anchor-time scale positive:

$$s_j = e_{j-1}(T_j/T_{j-1})^2 + |\widehat{L}^{\text{mf}}_j - \widehat{L}^{\text{fr}}_j| + \delta_{\text{pad}}. \tag{12}$$

The first window supplies the initial error scale. For completed calibration run i and target miscoverage $\alpha \in (0,1)$, form the run-indexed score $S_i = \max_j |L_{ij} - \widehat{L}^{\text{mf}}_{ij}|/s_{ij}$ over the common subsequent windows and take its split-conformal $(1-\alpha)$ quantile q_{conf}. The interval $\widehat{L}^{\text{mf}}_j \pm q_{\text{conf}} s_j$ is fixed at anchor t_j, before L_j is observed. Taking the maximum over windows preserves within-run dependence

and yields simultaneous run-level coverage under complete-run exchangeability. Appendix B.4.3 records the design.

3.3 Prediction Horizon and Irreducible Error

Future stochastic forcing imposes a method-independent limit on causal prediction. Theorem 1 identifies this irreducible component, and Proposition 2 adds the avoidable error caused by freezing feature learning.

Theorem 1 (Causal predictability limit)**.** *For the affine process* (7)*, let $\mathcal{F}_t$ be the information observed through time t and write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. Let the realized anchor state be observed, the post-anchor paths (A_s, b_s, Q_s) be $\mathcal{F}_t$-measurable with $Q_s = G_s G_s^\top$, and future Brownian innovations be independent of $\mathcal{F}_t$. Assume G_s is predictable and conditionally square-integrable, and $c_{t+T} \in L^2$ for each considered T. Define*

$$\mu_{t+T} = \mathbb{E}_t[c_{t+T}], \qquad q_t(T) = \frac{1}{T}\int_t^{t+T} \mathrm{Tr}\big(U(t+T,s) Q_s U(t+T,s)^\top\big)\, ds. \tag{13}$$

Then the exact conditional mean is the best causal state predictor, and

$$\operatorname*{ess\,inf}_{\widehat{c} \in L^2(\mathcal{F}_t;\mathbb{R}^d)} \mathbb{E}_t \|c_{t+T} - \widehat{c}\|_2^2 = q_t(T) T. \tag{14}$$

Thus $q_t(T)T$ is the variance created after the anchor and is attained by the exact conditional mean. Define the intrinsic horizon as the first tolerance violation,

$$T_{\mathrm{int}}(\varepsilon) = \inf\{T > 0 : q_t(T) T > \varepsilon\}, \tag{15}$$

with $\inf \varnothing = \infty$. Every $T < T_{\mathrm{int}}(\varepsilon)$ is feasible for the exact conditional mean. If Q_s is continuous at t and A_s is locally bounded, $q_t(T) = \mathrm{Tr}\, Q_t + o(1)$; for $\mathrm{Tr}\, Q_t > 0$, $T_{\mathrm{int}}(\varepsilon) \sim \varepsilon / \mathrm{Tr}\, Q_t$ as $\varepsilon \downarrow 0$.

The floor $q_t(T)T$ measures spread from the same observed state. Moving-frame prediction approximates its conditional mean; the next result bounds Frozen's deviation from that mean. Appendix B.5 gives the proof and weighted extension.

Proposition 2 (Frozen-frame drift penalty and certified horizon)**.** *In addition to Theorem 1, fix a prescribed local cap $T_{\max} > 0$ and assume the local boundedness conditions in Appendix B.5. Express both states in a common checkpoint-fixed chart. Let $\widehat{\mu}^{\mathrm{fr}}$ denote Frozen, initialized at the observed state with the drift and forcing in that chart held at their anchor values. Then there is a finite checkpoint-local coefficient $D_t \geq 0$ such that, for every $0 \leq T \leq T_{\max}$,*

$$0 \leq \underbrace{\mathbb{E}_t \|c_{t+T} - \widehat{\mu}^{\mathrm{fr}}_{t+T}\|_2^2}_{\mathrm{MSE}_{\mathrm{Frozen}}} - q_t(T) T = \|\mu_{t+T} - \widehat{\mu}^{\mathrm{fr}}_{t+T}\|_2^2 \leq D_t T^4. \tag{16}$$

Consequently $q_t(T)T \leq \mathrm{MSE}_{\mathrm{Frozen}} \leq q_t(T)T + D_t T^4$. Let $\bar{q}_t := \sup_{0<T\leq T_{\max}} q_t(T) < \infty$. For a fixed tolerance $\varepsilon \geq 0$ and $D_t + \bar{q}_t > 0$, let $T_{\mathrm{fr}}(\varepsilon)$ be the unique nonnegative solution of

$$D_t T_{\mathrm{fr}}^4 + \bar{q}_t T_{\mathrm{fr}} = \varepsilon. \tag{17}$$

Every $0 \leq T \leq \min\{T_{\max}, T_{\mathrm{fr}}(\varepsilon)\}$ satisfies the upper certificate; when $T_{\mathrm{fr}}(\varepsilon) \leq T_{\max}$, the root itself is certified by the local bound.

Under the local approximation $q_t(T) \approx q_t$, one may likewise use $\bar{q}_t \approx q_t$: drift dominance gives $T_{\mathrm{fr}} \sim (\varepsilon/D_t)^{1/4}$, whereas $T_{\mathrm{int}} \sim \varepsilon/q_t$. Figure 1 recovers the corresponding early-time powers on 16 end-to-end MLP anchors ($0.960/4.017$); the predicted and measured component switches occur at $1{,}055$ and $1{,}008$ updates, and the median moving-frame/Frozen horizon ratio is 1.92. These quantities describe state-space predictability; the calibration intervals above translate forecast uncertainty into the scalar loss observed by the experiment.

Proposition 2, proved in Appendix B.5, separates the irreducible term $q_t(T)T$ from the avoidable freezing penalty $D_t T^4$. Here D_t measures how rapidly the local relaxation and forcing drift, while $\bar{q}_t$ summarizes the strongest diffusion over the window. Solving where their sum reaches the tolerance gives a guaranteed local horizon. Figure 3(a,b) complements this state-space picture by showing how the separate propagation defects add to the upper bound on endpoint loss error.

4 EXPERIMENTS

We evaluate at fixed anchors in three settings: fixed-readout controls isolate the analytic mechanisms, end-to-end multilayer perceptrons (MLPs) expose learned-feature motion, and character-level Transformers test character next-token prediction (char-NTP) on real text. Each forecast uses anchor information, with its endpoint reserved for evaluation. Replicas estimate the checkpoint-conditioned moments; independent problems or complete runs are the comparison units. Appendix C.1 gives the architectures and metrics, and Appendix C.4 repeats Procedure 1 with LLM scale Pythia-1B.

Figure 3 follows the logic of the theory. With horizons measured in optimizer steps ($T = N$), panel (a) verifies $|\widehat{E}^{\mathrm{loc}}_{\mathrm{test}} - E^{\mathrm{exact}}_{\mathrm{test}}| \le \mathcal{B}_{\mathrm{loc}} + \epsilon_{\mathrm{eval}}$ on 16 problems and four horizons: every endpoint lies below the full-rank bound, with maximum error-to-budget ratio 0.410. Panel (b) explains that margin. Feature-map and affine-forcing drift supply most of the upper bound; operator drift is secondary, and covariance, numerical, and full-rank subspace terms remain small or zero. In panel (c), multi-step moving-frame propagation has lower error than a one-step Euler update in $47/50$ problems and closely matches the measured-endpoint reference. Panel (d) moves from retrospective diagnosis to an anchor-issued interval, with three exceedances among 960 negative-log-likelihood (NLL) forecasts and simultaneous coverage in $77/80$ complete text trajectories.

Table 1 compares point prediction under the same anchor information. Persistence keeps the current loss, Frozen propagates the readout in the checkpoint frame, and the learning-curve baselines extrapolate the observed scalar prefix; the moving-frame predictor also transports the representation. Its median error is smallest on both model classes, improving on Frozen by factors of 15.5 for MLPs and 13.7 for character prediction. The close match to the measured-endpoint reference indicates a small remaining frame-prediction error.

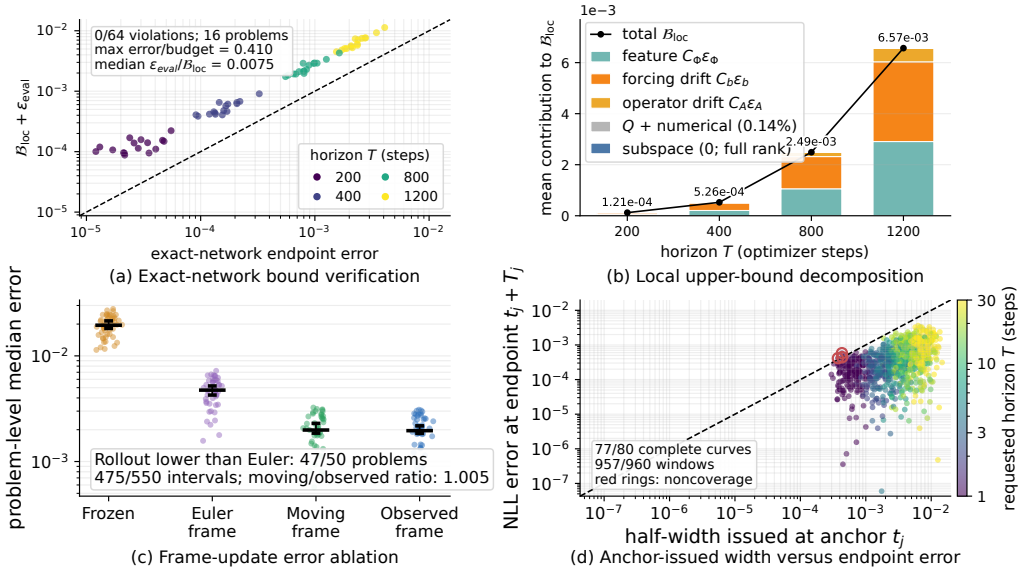


Figure 3: Reliability results: (a) exact-network error versus the full-rank bound $\mathcal{B}_{\mathrm{loc}} + \epsilon_{\mathrm{eval}}$; (b) decomposition of $\mathcal{B}_{\mathrm{loc}}$ (stack: total; subspace: zero); (c) frame-update ablation; and (d) run-level calibrated half-width issued at anchor t_j versus the error observed at endpoint $t_j + T_j$ (red rings: noncoverage).

Table 1: Median relative error (char-NTP NLL; MLP MSE); bold is best.

Predictor	MLP (50)	char-NTP (48)
Persistence	0.264	0.00717
Domhan `pow3` (Domhan et al., 2015)	0.367	0.00950
LC-PFN (Adriaensen et al., 2023)	0.0219	0.00259
FT-PFN (Rakotoarison et al., 2024)	0.0622	0.00253
Frozen frame	0.101	0.00189
Moving frame	**0.00652**	$\mathbf{1.38 \times 10^{-4}}$

5 Related Work

Forecasting training progress varies in both its conditioning information and its target. Learning-curve extrapolation maps an observed scalar prefix to the remaining curve or terminal validation performance for early stopping and hyperparameter optimization (Domhan et al., 2015; Klein et al., 2017; Adriaensen et al., 2023; Rakotoarison et al., 2024). Training-state methods predict time-to-target or near-future weights from task information and recent optimization history (Zancato et al., 2020; Jang et al., 2023). Our formulation conditions on a realized checkpoint—model and optimizer state, local feature geometry, and replica statistics—and predicts the continuation of that same run: its mean path, its spread under future minibatch sampling, and the horizon at which this spread reaches a tolerance.

Kernel and spectral theories explain learning curves through relaxation along data- or tangent-kernel modes (Jacot et al., 2018; Lee et al., 2019; Bordelon et al., 2020), while high-dimensional analyses connect generalization dynamics to task geometry (Advani et al., 2020). End-to-end networks also evolve their tangent kernels and align features with the task (Fort et al., 2020; Atanasov et al., 2022; Yang and Hu, 2021). The moving frame gives this evolution a checkpoint-local interpretation: Gram eigenspaces define the coordinates, the optimizer predicts their rotation, and the readout distribution moves with them. Frozen is the associated local fixed-kernel approximation.

Stochastic modified equations and diffusion approximations describe minibatch SGD (Li et al., 2017; Mandt et al., 2017); dynamical mean-field and finite-width theories describe trajectory laws and prediction fluctuations (Gerbelot et al., 2024; Bordelon and Pehlevan, 2024). Conditioning these dynamics on the observed checkpoint separates forecastable drift from uncertainty created by future innovations. Classical perturbation theory makes the spectral transport stable through eigenvalue crossings (Kato, 1966; Davis and Kahan, 1970; Stewart and Sun, 1990), and split conformal prediction converts completed-run errors into intervals for the checkpoint-to-endpoint sequence (Lei et al., 2018).

6 Limitations

The local approximation is strongest while curvature and optimizer state vary smoothly. Abrupt regimes call for shorter windows or a richer state; multimodal or heavy-tailed futures call for more than two moments. Predictive certificates use anchor-time envelopes, and run-level calibration transfers across exchangeable runs under the same design.

7 Conclusion

Checkpoint-conditioned forecasting characterizes both a future learning curve and what is predictable in principle from an observed training state. The moving frame tracks forecastable representation dynamics, while the diffusion floor isolates uncertainty created after the checkpoint. Experiments locate the transition from point prediction to calibrated uncertainty, with additional LLM-scale cases supporting the same forecasting framework. Thus learning-curve predictability delineates the frontier of what can be understood about learning dynamics from an observed checkpoint.

AI use statement

Generative AI assisted with experiment and figure code, Lean 4 formalization, and English proof-reading. The authors developed the research ideas, draft, and proofs, reviewed the assisted outputs, and take responsibility for the paper and artifacts.

Ethics statement

This theoretical study uses synthetic tasks and public text corpora to analyze aggregate training losses and states.

Reproducibility statement

The supplement provides experiment code and deterministic and stochastic Lean 4 formalizations.

REFERENCES

S. Adriaensen, H. Rakotoarison, S. Müller, and F. Hutter. Efficient Bayesian learning curve extrapolation using prior-data fitted networks. *NeurIPS*, 2023.

M. Advani, A. Saxe, and H. Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

A. Atanasov, B. Bordelon, and C. Pehlevan. Neural networks as kernel learners: the silent alignment effect. *ICLR*, 2022.

B. Bordelon, A. Canatar, and C. Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. *ICML*, 2020.

B. Bordelon and C. Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104021, 2024.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

T. Domhan, J. T. Springenberg, and F. Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. *IJCAI*, pages 3460–3468, 2015.

S. Fort, G. K. Dziugaite, M. Paul, S. Kharaghani, D. M. Roy, and S. Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *NeurIPS*, 2020.

C. Gerbelot, E. Troiani, F. Mignacco, F. Krzakala, and L. Zdeborová. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. *NeurIPS*, 2018.

J. Jang, W.-H. Yun, W. H. Kim, Y. Yoon, J. Kim, J. Lee, and B. Han. Learning to boost training by periodic nowcasting near future weights. *ICML*, pages 14730–14757, 2023.

T. Kato. *Perturbation Theory for Linear Operators*. Springer, 1966.

A. Klein, S. Falkner, J. T. Springenberg, and F. Hutter. Learning curve prediction with Bayesian neural networks. *ICLR*, 2017.

J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *NeurIPS*, 2019.

J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Q. Li, C. Tai, and W. E. Stochastic modified equations and adaptive stochastic gradient algorithms. *ICML*, 2017.

S. Mandt, M. Hoffman, and D. Blei. Stochastic gradient descent as approximate Bayesian inference. *JMLR*, 18(134):1–35, 2017.

H. Rakotoarison, S. Adriaensen, N. Mallik, S. Garibov, E. Bergman, and F. Hutter. In-context freeze–thaw Bayesian optimization for hyperparameter optimization. *ICML*, 2024.

G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

G. Yang and E. J. Hu. Tensor programs IV: feature learning in infinite-width neural networks. *ICML*, pages 11727–11737, 2021.

L. Zancato, A. Achille, A. Ravichandran, R. Bhotika, and S. Soatto. Predicting training time without training. *NeurIPS*, 2020.

APPENDIX CONTENTS

A Notation and Symbols **12**
B Theory: Proofs, Implementations, and Certificates **13**
B.1 The Post-Anchor Replica Process and Fixed-Frame Forms 13
B.2 Moving-Frame Results 15
B.3 Covariance Transport and Moving-Frame Geometry 15
B.4 Predictor Implementations and Local Error Certificates 18
B.5 The Predictability Horizon 22
B.6 Local Cross-Entropy Geometry 23
C Experimental Details **24**
C.1 Forecast Protocol and Conditioning 24
C.2 Theory-to-Experiment Correspondence 24
C.3 Prediction Experiments 27
C.4 LLM-Scale Same-Anchor Continuations 28

A NOTATION AND SYMBOLS

The first table contains the notation needed to read the main paper. The second isolates recurring symbols used only in the appendices; variables local to one proof are still defined where introduced. Hats denote predictions unless a row states otherwise.

A.1 CORE NOTATION USED IN THE MAIN TEXT

Symbol	Meaning	Symbol	Meaning
Indices, horizons, and conditioning			
$t, t_0, \mathcal{I}$	time, measured anchor, forecast interval	T, Δ, h, N	forecast horizons T, Δ, step duration h, and update count N
$T_{\mathrm{int}}, T_{\mathrm{fr}}, T_{\max}, \varepsilon$	intrinsic/Frozen horizons, local cap, and tolerance	$q_t(T), \bar{q}_t, D_t$	diffusion density, its local envelope, and drift coefficient
D, X, P, d	target dataset, evaluation inputs, number of feature rows, feature dimension	$M, m, [M]$	replica count, replica index, finite-replica estimator
$\mathbb{E}_m, \mathrm{Var}_m, \mathrm{Cov}_m$	conditional population moments over a generic replica	$\mathcal{F}_t$	all information observed through time t
$\mathcal{G}_{t_0}^{\mathrm{shared}}$	anchor information common to replicas	$\mathcal{H}_{\mathcal{I}}$	$\mathcal{G}_{t_0}^{\mathrm{shared}}$ plus the coefficient path on $\mathcal{I}$
t_j, T_j	anchor and horizon of forecast window j	$\|\cdot\|_2, \|\cdot\|_F$	Euclidean/operator and Frobenius norms
State, features, and geometry			
$\theta, \bar{\theta}, \delta\theta_m$	parameters, population mean, replica deviation	Φ_t, ϕ_θ	feature matrix and feature map
Σ_t	normalized feature Gram matrix $\Phi_t^\top \Phi_t / P$	V_t, Λ_t	spectral frame and eigenvalue matrix
v_k, λ_k	mode-k eigenvector and eigenvalue	$a_{m,t}, c_{m,t}$	readout and coordinates $V_t^\top a_{m,t}$
$\bar{c}_t, \Sigma_{\mathrm{mem},t}$	conditional mean and total replica covariance	$\Sigma_{\mathrm{aff},t}$	affine-process covariance under coefficient-path conditioning
Σ_θ	full parameter covariance		
Local dynamics and stochastic forcing			
η, λ, g_t	learning rate, readout decay, target projection	K_t	spectral decay $\eta(\Lambda_t + \lambda I)$
$\Omega_t, \nabla_t^{\mathrm{mov}} c$	frame connection and moving-frame derivative	A_t, b_t	affine drift matrix and forcing, $-K_t - \Omega_t$ and $K_t g_t$
$G_t, Q_t, W_{m,t}$	source factor, rate covariance $G_t G_t^\top$, Brownian innovation	Q_t^{step}	one-step innovation covariance in the discrete predictor
ρ_k, n_k^m, ν_k	mode contraction, fixed shift, and its variance	$U(r, s)$	continuous propagator from s to r
R_{clean}	shared full-batch objective defining the clean propagation	B_3, L_3	bound and Lipschitz constant for $\nabla^3 R_{\mathrm{clean}}$
Evaluation, error, and certificates			
$E_{\mathrm{test}}^{\mathrm{loc}}$	local test loss	$L_{\mathrm{mean},t}, C_{\mathrm{te},t}$	mean loss and one-half local loss Hessian
$\widehat{(\cdot)}, (\cdot)^\sharp$	predicted quantity and aligned-chart quantity	$\epsilon_A, \epsilon_b, \epsilon_Q^{\mathrm{int}}$	integrated coefficient defects
$\epsilon_\Phi, \epsilon_{\mathrm{sub}}$	feature and spectral-subspace defects	$\epsilon_{\mathrm{disc}}, \mathcal{B}_{\mathrm{loc}}, \epsilon_{\mathrm{eval}}$	discretization defect, total local budget, local-to-network gap
$L_j, \widehat{L}_j^{\mathrm{mf}}, \widehat{L}_j^{\mathrm{fr}}$	endpoint loss at $t_j + T_j$ and its anchor-issued forecasts	$e_{j-1}, s_j, \delta_{\mathrm{pad}}$	preceding error, anchor-time scale, positive padding
$S_i, q_{\mathrm{conf}}, \alpha, n_{\mathrm{cal}}$	maximum score, conformal quantile, miscoverage, calibration count		

A.2 NOTATION CONFINED TO THE APPENDICES

Symbol	Meaning	Symbol	Meaning
$\mathcal{H}_{\mathrm{eval}}$	field making a random evaluation path measurable	U_t, Γ_t, z_t	left singular frame, singular values, and function coordinate
$\mathsf{P}_t, K_{\mathsf{P}_t}^\phi, u_j$	optimizer preconditioner, feature kernel, and cut-level residual	O_s	Procrustes alignment into the true chart
R_j	one-step predicted-frame overlap	$\epsilon_{\mathrm{cd}}^{\mu,\Sigma}, \epsilon_{\mathrm{split}}^{\mu,\Sigma}$	continuous/discrete and endpoint-splitting defects
L_μ, L_Σ	endpoint Lipschitz constants for mean and covariance	M_U, ω_+	local propagator-bound constants
L_A, L_b	coefficient-derivative bounds	$H_{\max}, R_{\mu,\max}$	local propagator and solution bounds
$\Sigma^{(t)}$	covariance restarted from zero at the anchor	$F_i, F(\bar{z})$	per-example and mean-logit categorical Fisher matrices

B THEORY: PROOFS, IMPLEMENTATIONS, AND CERTIFICATES

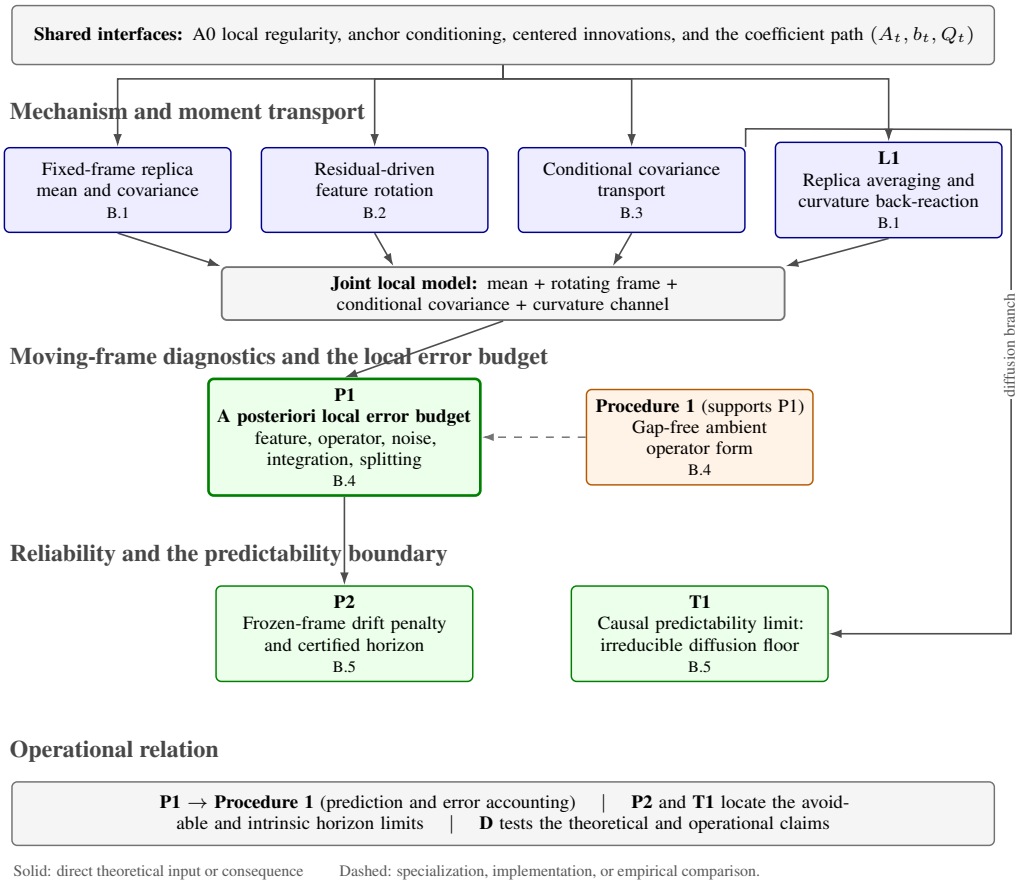


Figure 4: **Dependency map for the central results.** Read from top to bottom. The fixed-frame moments, optimizer-induced rotation law, covariance transport, and Lemma 1 supply the mean, frame, covariance, and curvature channels. Proposition 1 combines them into the local error bound; Proposition 2 and Theorem 1 separate avoidable feature-freezing error from the intrinsic diffusion limit. Procedure 1 implements the predictor, and Appendix C evaluates these claims.

Throughout Appendix B, population moments are conditional on the anchor field $\mathcal{G}_{t_0}^{\mathrm{shared}}$ unless a finer coefficient-path field is displayed. Under that finer field the coefficients are fixed and the innovations are centered; returning to the anchor field uses conditional Jensen or total covariance. True and predicted states are aligned before comparison, $U(r,s)$ propagates from s to r, and empty ordered products are identities. We write Σ_{mem} for total replica covariance and Σ_{aff} for the covariance of the coefficient-conditioned affine process. In (34), its first term propagates anchor covariance and its second term is the post-anchor refreshed contribution.

B.1 THE POST-ANCHOR REPLICA PROCESS AND FIXED-FRAME FORMS

We first separate population-mean dynamics from finite-replica sampling and then solve the fixed-frame readout recursion. These results provide the mean and covariance baselines used throughout the moving-frame analysis.

Proof of Lemma 1. Decompose a generic replica as $\theta_m = \bar{\theta} + \delta\theta_m$, where $\mathbb{E}_m[\delta\theta_m] = 0$, and split its gradient into the clean consensus part and centered replica noise. *Step 1 (population mean).* Conditional expectation removes the centered noise. Taylor expansion of the clean gradient around

the population mean gives

$$\mathbb{E}_m\big[\nabla R_{\text{clean}}(\bar{\theta}+\delta\theta_m)\big] = \nabla R_{\text{clean}}(\bar{\theta}) + \tfrac{1}{2}\,\nabla^3 R_{\text{clean}}(\bar{\theta}) : \Sigma_\theta + O\big(\mathbb{E}_m\|\delta\theta_m\|^3\big), \tag{18}$$

because the first-order term vanishes by centering. A0 bounds the uncorrected second-order term by $B_3\mathbb{E}_m\|\delta\theta_m\|^2/2$ and the remainder after retaining it by $L_3\mathbb{E}_m\|\delta\theta_m\|^3/6$; multiplication by $-\eta$ proves the population display and both bias bounds. *Step 2 (finite sample).* Replacing the conditional expectation by $M^{-1}\sum_m$ adds a centered empirical fluctuation. Under the conditional-iid variance condition in A0, its conditional variance is $O(M^{-1})$, hence its size is $O_p(M^{-1/2})$. The curvature contraction uses the full Σ_θ; in a closed block such as a readout-only model, this is exactly the covariance tracked by the forecast. □

Derivation of the fixed-frame moments. For squared loss with fixed features, let

$$L_m(a) = \frac{1}{2P}\|\Phi a - y_m\|_2^2 + \frac{\lambda}{2}\|a\|_2^2, \qquad y_m = y + \xi_m, \tag{19}$$

and diagonalize $\Phi^\top\Phi/P = V\Lambda V^\top$. We prove the result one spectral coordinate at a time.

Step 1 (one-step scalar recursion). Projecting one gradient step onto v_k gives

$$c_{m,k}(r+1) = \rho_k c_{m,k}(r) + (1-\rho_k)(g_k + n_k^m) + \varepsilon_{m,k}(r), \qquad \rho_k = 1-\eta(\lambda_k+\lambda), \tag{20}$$

where

$$g_k = \frac{v_k^\top \Phi^\top y/P}{\lambda_k+\lambda}. \tag{21}$$

The variable n_k^m is the projected replica-fixed target displacement, so it is constant in r, centered across replicas, and has variance ν_k. The increment $\varepsilon_{m,k}(r)$ is refreshed at step r, is centered conditionally on the past, and has variance Q_{kk}.

Step 2 (solve the recursion). Repeated substitution yields

$$c_{m,k}(r) = \rho_k^r c_{0,k} + (1-\rho_k^r)(g_k + n_k^m) + \sum_{j=0}^{r-1}\rho_k^j \varepsilon_{m,k}(r-1-j). \tag{22}$$

This identity can be checked by inserting $r+1$ and separating the newest innovation. Taking conditional expectation removes n_k^m and every refreshed increment, which gives

$$\bar{c}_k(r) = \rho_k^r c_{0,k} + g_k(1-\rho_k^r). \tag{23}$$

Step 3 (separate the two variance sources). Subtracting the mean gives

$$c_{m,k}(r) - \bar{c}_k(r) = (1-\rho_k^r)n_k^m + \sum_{j=0}^{r-1}\rho_k^j \varepsilon_{m,k}(r-1-j). \tag{24}$$

The persistent term and the refreshed terms are mutually uncorrelated by assumption, and distinct refreshed increments are uncorrelated. Therefore all cross terms vanish after conditional expectation and

$$\mathrm{Var}_m[c_{m,k}(r)] = \nu_k(1-\rho_k^r)^2 + \sum_{j=0}^{r-1}\rho_k^{2j} Q_{kk}. \tag{25}$$

This also explains the different short-time behavior: the persistent displacement is first multiplied by $1-\rho_k^r$, whereas one refreshed innovation enters immediately.

Step 4 (quadratic evaluation). Write $c_m = \bar{c} + \delta c_m$ in the quadratic test loss. Its linear term in δc_m has zero conditional mean, while the quadratic term satisfies

$$\mathbb{E}_m[\delta c_m^\top C_{\text{te}}\delta c_m] = \mathrm{Tr}(C_{\text{te}}\Sigma_{\text{mem}}). \tag{26}$$

This recovers the evaluation identity (2) in the fixed-frame quadratic specialization. For the continuous-time limit, introduce a step duration h, set $\eta = h\bar{\eta}$ and $Q_{kk} = h\bar{Q}_{kk}$, and hold $\tau = rh$ fixed as $h \to 0$. Then ρ_k^r converges to $e^{-\bar{\eta}(\lambda_k+\lambda)\tau}$ and the geometric variance sum becomes the corresponding Ornstein–Uhlenbeck integral with source rate $\bar{Q}_{kk}$. Thus contraction transports existing uncertainty while the source continuously injects new uncertainty.

B.2 Moving-Frame Results

We next derive the connection term Ω_t created by changing feature coordinates and express its motion through the current optimizer step. The same local expansion identifies the curvature back-reaction of replica dispersion on the mean.

Derivation of the feature-rotation law. *Coordinate identity and connection.* Differentiating $c = V^\top a$ and $V^\top V = I$ gives $\dot{c} = V^\top \dot{a} - \Omega c$ for a complete frame. For a reduced frame, closure of the modeled subspace is required; otherwise $\dot{V}^\top (I - VV^\top)a$ remains. Differentiating $\Sigma v_j = \lambda_j v_j$ and projecting onto v_i, $i \neq j$, gives $v_i^\top \dot{\Sigma} v_j = (\lambda_j - \lambda_i)\Omega_{ij}$.

Two-layer case. Let φ be the activation, w_j the hidden weight of unit j, a_j its scalar readout coefficient, and r_l the prediction residual on example l. For $\Phi_{ij} = \varphi(w_j^\top x_i)$, the chain rule gives $\dot{\Phi}_{ij} = \varphi'(w_j^\top x_i)\, x_i^\top \dot{w}_j$. Under gradient flow on the empirical squared risk, substituting $\dot{w}_j = -\eta a_j P^{-1} \sum_l r_l \varphi'(w_j^\top x_l) x_l$ gives the claimed residual–kernel factorization directly. *General case.* At an arbitrary differentiable cut, Taylor's theorem gives $\Delta\Phi = J_\phi \, \Delta\theta + O(\|\Delta\theta\|^2)$ and the product rule gives $\Delta\Sigma = (\Delta\Phi^\top \Phi + \Phi^\top \Delta\Phi)/P + O(\|\Delta\Phi\|^2)$. For the data-loss contribution, write the parameter block below the cut as $\nabla R = \frac{1}{P} \sum_j J_j^\top u_j$, with u_j the loss gradient backpropagated to the cut; since J_ϕ vanishes on parameters above the cut and P_t does not couple the blocks, only this block enters $\Delta\Phi$ to first order. Then row i of the feature increment factorizes as

$$\Delta\Phi_i = -\frac{\eta}{P} \sum_j \underbrace{J_i \, \mathsf{P}_t \, J_j^\top}_{K^\phi_{\mathsf{P}_t}(x_i, x_j)} \, u_j + O(\|\Delta\theta\|^2), \tag{27}$$

the matrix-valued feature-NTK acting on the upper error field. Dividing by the step duration and taking the continuous-time limit yields the displayed velocity and connection. The architecture Jacobians and optimizer preconditioner jointly determine $K^\phi_{\mathsf{P}_t}$, while the data loss enters through u. If $R = R_{\text{data}} + R_{\text{reg}}$, the same chain rule adds $-\eta J_i \mathsf{P}_t \nabla R_{\text{reg}}$; hence vanishing u removes the first-order drive only in the absence of another force below the cut.

B.3 Covariance Transport and Moving-Frame Geometry

Here we propagate conditional mean and covariance along a prescribed moving-frame coefficient path and separate coordinate rotation from changes in test evaluation. The resulting identities distinguish variance redistribution from genuine contraction, forcing, and feature-map error.

Figure 5 separates the two conditioning levels used in (9). The upper two branches construct the covariance conditional on a coefficient path; conditioning back on the anchor adds variation of the conditional mean across coefficient paths.

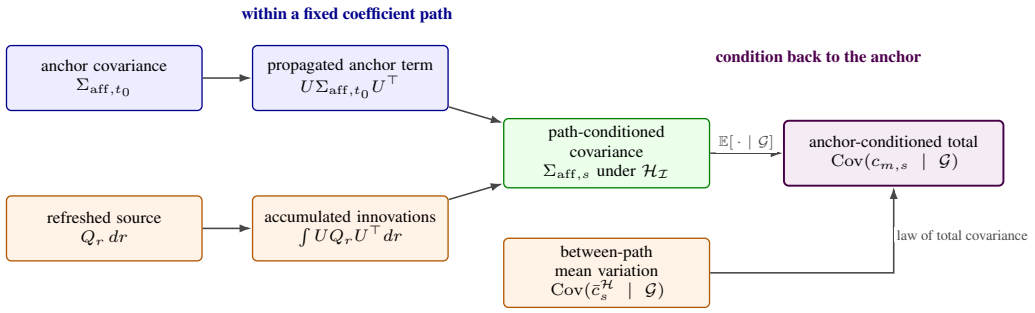


Figure 5: **Guide to the covariance-transport proof.** The Lyapunov integral combines transported anchor covariance and post-anchor refreshed noise under the finer field $\mathcal{H}_\mathcal{I}$. Returning to the anchor field $\mathcal{G}$ adds a distinct source of spread: variation of the path-conditioned mean across coefficient paths.

Derivation of conditional covariance transport. Condition on $\mathcal{H}_{\mathcal{I}}$. Thus A_t, b_t, G_t are fixed coefficient paths, while $\mathcal{M}_t = \int_{t_0}^{t} G_s\, dW_s$ remains a centered square-integrable martingale with bracket $d\langle \mathcal{M} \rangle_t = Q_t dt$.

Step 1 (conditional mean). Taking conditional expectation in (7) and using $\mathbb{E}[d\mathcal{M}_t \mid \mathcal{H}_{\mathcal{I}}] = 0$ gives

$$d\bar{c}_t^{\mathcal{H}} = (A_t \bar{c}_t^{\mathcal{H}} + b_t)dt. \tag{28}$$

Hence the centered state $x_t = c_{m,t} - \bar{c}_t^{\mathcal{H}}$ satisfies

$$dx_t = A_t x_t dt + d\mathcal{M}_t. \tag{29}$$

Centering cancels the common affine forcing b_t, so covariance responds to the linear deformation A_t and the stochastic source Q_t.

Step 2 (matrix product rule). Apply the Itô product rule to $x_t x_t^\top$:

$$d(x_t x_t^\top) = (dx_t)x_t^\top + x_t(dx_t)^\top + d[\mathcal{M}]_t. \tag{30}$$

Substitution of $dx_t = A_t x_t dt + d\mathcal{M}_t$ yields

$$\begin{aligned} d(x_t x_t^\top) = \big(A_t x_t x_t^\top + x_t x_t^\top A_t^\top + Q_t\big)dt \\ + d\mathcal{M}_t x_t^\top + x_t d\mathcal{M}_t^\top. \end{aligned} \tag{31}$$

The last two terms are martingale increments and have zero conditional expectation. Therefore

$$\dot{\Sigma}_{\mathrm{aff},t} = A_t \Sigma_{\mathrm{aff},t} + \Sigma_{\mathrm{aff},t} A_t^\top + Q_t. \tag{32}$$

The predictable bracket identifies Q_t as the instantaneous covariance injected by the martingale.

Step 3 (integral representation). Let $U(s,r)$ solve $\partial_s U(s,r) = A_s U(s,r)$ and $U(r,r) = I$. Variation of constants for the centered state gives

$$x_s = U(s,t_0)x_{t_0} + \int_{t_0}^{s} U(s,r)\, d\mathcal{M}_r. \tag{33}$$

Taking the conditional second moment, the initial–innovation cross terms vanish and conditional Itô isometry gives

$$\Sigma_{\mathrm{aff},s} = U(s,t_0)\Sigma_{\mathrm{aff},t_0}U(s,t_0)^\top + \int_{t_0}^{s} U(s,r)Q_r U(s,r)^\top dr. \tag{34}$$

Differentiating this display recovers the Lyapunov equation, so the differential and integral forms are equivalent.

Step 4 (return to the anchor field). Let $\mathcal{G} = \mathcal{G}_{t_0}^{\mathrm{shared}}$. The conditional law of total covariance gives

$$\mathrm{Cov}(c_{m,s} \mid \mathcal{G}) = \mathbb{E}[\mathrm{Cov}(c_{m,s} \mid \mathcal{H}_{\mathcal{I}}) \mid \mathcal{G}] + \mathrm{Cov}(\mathbb{E}[c_{m,s} \mid \mathcal{H}_{\mathcal{I}}] \mid \mathcal{G}), \tag{35}$$

which is exactly the second line of (9). When the coefficient path is random under $\mathcal{G}$, the second term carries its between-path variation. The coarse-field covariance therefore combines Lyapunov evolution within each path with dispersion between their conditional means.

Step 5 (population versus a finite ensemble). The preceding calculation describes a generic-replica population covariance. For M conditionally iid replicas with independent noises, the empirical covariance fluctuates around it through an additional martingale term. With $1/M$ normalization its drift source is $(1 - 1/M)Q_t$, while the $1/(M-1)$ convention in Section 1.1 is unbiased for $M \geq 2$. Persistent replica-specific forcing has a different geometry: it moves each replica toward its own center and is propagated with its readout–shift cross-covariance, as in the fixed-frame formula.

Coordinate transport. Feature-frame motion also changes the coordinates of a fixed ambient predictor. Let B_s, b_s, Ω_s be continuous and compare

$$\dot{c}_s^{\mathrm{mov}} = (B_s - \Omega_s)c_s^{\mathrm{mov}} + b_s, \qquad \dot{c}_s^{\mathrm{fr}} = B_s c_s^{\mathrm{fr}} + b_s, \qquad c_t^{\mathrm{mov}} = c_t^{\mathrm{fr}}. \tag{36}$$

Writing $\delta c_s = c_s^{\mathrm{fr}} - c_s^{\mathrm{mov}}$ gives $\dot{\delta c}_s = B_s \delta c_s + \Omega_s c_s^{\mathrm{mov}}$ and $\delta c_t = 0$. Variation of constants therefore gives

$$\delta c_{t+\Delta} = \int_{t}^{t+\Delta} U_B(t+\Delta, s)\Omega_s c_s^{\mathrm{mov}} ds. \tag{37}$$

If the propagator U_B is bounded by H_{rot} and E is L_E-Lipschitz in its mean coordinate, omitting basis rotation contributes at most

$$|\delta E_{\mathrm{rot}}| \le L_E H_{\mathrm{rot}} \int_t^{t+\Delta} \|\Omega_s c_s^{\mathrm{mov}}\|_2 ds. \tag{38}$$

This display isolates the mean-coordinate part of frame rotation. The same rotation acts on an anisotropic covariance through $-\Omega_s S_s^{\mathrm{mov}} - S_s^{\mathrm{mov}} \Omega_s^\top$, while eigenvalues, target projections, source covariance, and the evaluation map evolve through their own channels. Proposition 1 assigns these mechanisms to separate defects so that each term has a direct dynamical interpretation.

B.3.1 Aligned Comparison Chart

An exactly degenerate eigenspace admits many orthonormal bases, and nearby eigenvectors may rotate rapidly even when the subspace changes smoothly. We therefore treat a grouped invariant subspace as the geometric object and allow any orthonormal gauge, provided its within-group operator is transformed consistently. Under $V_t \mapsto V_t R_t$, $c_t \mapsto R_t^\top c_t$ and $\Omega_t \mapsto R_t^\top \Omega_t R_t + R_t^\top \dot{R}_t$. The derivative transforms as $\nabla_t^{\mathrm{mov}} c \mapsto R_t^\top \nabla_t^{\mathrm{mov}} c$, so function-space predictions and $E_{\mathrm{test}}^{\mathrm{loc}}(t)$ are gauge-invariant. The experiments accordingly compare activation-feature subspaces rather than individual neuron identities whenever possible.

Figure 6 shows why predicted coordinates are aligned before taking norms: alignment removes arbitrary basis orientation and leaves the dynamical mismatch between the two representations of the modeled subspace.

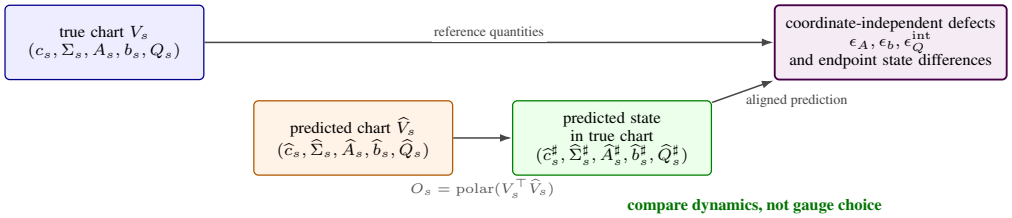


Figure 6: **Guide to the aligned comparison chart.** The Procrustes map transports every predicted quantity into the true chart. The derivative of this transport enters $\widehat{A}_s^\sharp$, placing connection mismatch inside the operator defect.

Projectors identify a subspace but leave its internal coordinates free. To compare a true frame V_s and predicted frame $\widehat{V}_s$, on a same-rank interval where $\sigma_{\min}(V_s^\top \widehat{V}_s) > 0$, choose the Procrustes identification

$$O_s = \mathrm{polar}(V_s^\top \widehat{V}_s), \qquad O_t = I \tag{39}$$

after fixing the predicted anchor gauge. For continuous coefficient comparison assume this choice is C^1 on the interval and define

$$\begin{aligned} \widehat{c}_s^\sharp &= O_s \widehat{c}_s, & \widehat{\Sigma}_s^\sharp &= O_s \widehat{\Sigma}_s O_s^\top, \\ \widehat{A}_s^\sharp &= O_s \widehat{A}_s O_s^\top + \dot{O}_s O_s^\top, & \widehat{b}_s^\sharp &= O_s \widehat{b}_s, \qquad \widehat{Q}_s^\sharp = O_s \widehat{Q}_s O_s^\top. \end{aligned} \tag{40}$$

Every coefficient and state difference in Proposition 1 is taken in this chart. With Euclidean, subordinate operator, and Frobenius norms, define

$$\epsilon_A = \int_t^{t+T} \|A_s - \widehat{A}_s^\sharp\|_2 ds, \quad \epsilon_b = \int_t^{t+T} \|b_s - \widehat{b}_s^\sharp\|_2 ds, \quad \epsilon_Q^{\mathrm{int}} = \int_t^{t+T} \|Q_s - \widehat{Q}_s^\sharp\|_F ds. \tag{41}$$

Assume the local evaluation map is Lipschitz in its feature, mean, and covariance arguments, the reference moments are bounded, and $\|U(r,s)\|_2 \le M_U e^{\omega_+(r-s)}$, with $\omega_+ = \max(\omega, 0)$. Set $\epsilon_\Phi = \sup_s \|\Phi_s - \widehat{\Phi}_s\|_F$ for the evaluation feature map. The ambient projector defect ϵ_{sub} directly measures subspace separation. A Davis–Kahan estimate connects it to Gram-matrix error when an external spectral gap is available; otherwise the comparison remains in ambient coordinates.

B.4 Predictor Implementations and Local Error Certificates

This subsection turns the continuous equations into the checkpoint-to-checkpoint predictor used in the experiments. It first constructs the discrete and gap-free implementations, then follows each approximation from the anchor to the endpoint to obtain the local error budget. The final part converts past endpoint errors into run-level calibrated intervals.

B.4.1 Discrete-Time Predictor

Let the continuous horizon be $T = Nh$, with $t_j = t + jh$ for $j = 0, \dots, N$. Condition on the predicted frame path $\widehat{V}_j$ and define $R_j = \widehat{V}_{j+1}^\top \widehat{V}_j$. In the optimizer-step normalization used in the experiments, $h = 1$ and $\rho_t = I - \eta(\Lambda_t + \lambda I)$. For a rescaled time coordinate, ρ_t is formed only after replacing the drift coefficient by its one-step integral. Let Q_t^{step} be the innovation covariance over one discrete step in the pre-transport coordinates; under a frozen continuous source rate Q_t, the first-order relation is $Q_t^{\mathrm{step}} = hQ_t$. Starting from the same anchor population mean and covariance as every other construction in Proposition 1, freezing $(\rho_t, g_t, Q_t^{\mathrm{step}})$ while interleaving the prescribed frame transports gives

$$\begin{aligned} T_j^{\mathrm{fr}} &= R_j \rho_t, & h_j^{\mathrm{fr}} &= R_j (I - \rho_t) g_t, \\ \bar{c}_{j+1}^{\mathrm{fr}} &= T_j^{\mathrm{fr}} \bar{c}_j^{\mathrm{fr}} + h_j^{\mathrm{fr}}, & \Sigma_{j+1}^{\mathrm{fr}} &= T_j^{\mathrm{fr}} \Sigma_j^{\mathrm{fr}} (T_j^{\mathrm{fr}})^\top + R_j Q_t^{\mathrm{step}} R_j^\top. \end{aligned} \tag{42}$$

Equivalently, at state level $c_{j+1}^{\mathrm{fr}} = T_j^{\mathrm{fr}} c_j^{\mathrm{fr}} + h_j^{\mathrm{fr}} + R_j \xi_j$, where ξ_j is centered, conditionally uncorrelated with the current state, and has covariance Q_t^{step}. If that covariance is stored in post-transport coordinates, the outer R_j factors are omitted. Conditional on the frame path, this is the exact recursion for frozen coefficients with refreshed independent increments. The superscript fr labels these frozen coefficients while retaining the transports R_j; the plotted Frozen baseline additionally replaces every R_j by the identity.

Let $O_j = O_{t_j}$ be the Procrustes maps above. The interleaved objects in the common true chart are

$$\begin{aligned} \mathcal{P}_j^{\mathrm{fr},\sharp} &= O_{j+1} T_j^{\mathrm{fr}} O_j^\top, & h_j^{\mathrm{fr},\sharp} &= O_{j+1} h_j^{\mathrm{fr}}, \\ \mathcal{Q}_j^{\mathrm{fr},\sharp} &= O_{j+1} R_j Q_t^{\mathrm{step}} R_j^\top O_{j+1}^\top, & \bar{c}_j^{\mathrm{fr},\sharp} &= O_j \bar{c}_j^{\mathrm{fr}}, \quad \Sigma_j^{\mathrm{fr},\sharp} = O_j \Sigma_j^{\mathrm{fr}} O_j^\top. \end{aligned} \tag{43}$$

To relate the continuous reference to this discrete recursion, let $\widehat{U}^\sharp$ be the propagator of $\widehat{A}_s^\sharp$ and define its exact one-step homogeneous, affine, and source maps

$$\begin{aligned} \mathcal{P}_j^{\mathrm{cont}} &= \widehat{U}^\sharp(t_{j+1}, t_j), \\ h_j^{\mathrm{cont}} &= \int_{t_j}^{t_{j+1}} \widehat{U}^\sharp(t_{j+1}, r) \widehat{b}_r^\sharp \, dr, \\ \mathcal{Q}_j^{\mathrm{cont}} &= \int_{t_j}^{t_{j+1}} \widehat{U}^\sharp(t_{j+1}, r) \widehat{Q}_r^\sharp \widehat{U}^\sharp(t_{j+1}, r)^\top dr. \end{aligned} \tag{44}$$

The aligned continuous reference moments $(\bar{c}^{\mathrm{cont}}, \Sigma^{\mathrm{cont}})$ use these exact maps and start from the common anchor moments. Define the continuous-to-discrete defects

$$\epsilon_{\mathrm{cd}}^{\mu} = \|\bar{c}_{t+T}^{\mathrm{cont}} - \bar{c}_N^{\mathrm{fr},\sharp}\|_2, \qquad \epsilon_{\mathrm{cd}}^{\Sigma} = \|\Sigma_{t+T}^{\mathrm{cont}} - \Sigma_N^{\mathrm{fr},\sharp}\|_F. \tag{45}$$

The experiments retain the following endpoint-transport, or endpoint-split, approximation. With $R_{0,N} = \widehat{V}_N^\top \widehat{V}_0$,

$$\begin{aligned} \bar{c}_N^{\mathrm{end}} &= R_{0,N} \big[\rho_t^N \bar{c}_t + (I - \rho_t^N) g_t \big], \\ \Sigma_N^{\mathrm{end}} &= T_N^{\mathrm{end}} \Sigma_{\mathrm{mem},t} (T_N^{\mathrm{end}})^\top + \sum_{r=0}^{N-1} R_{0,N} \rho_t^r Q_t^{\mathrm{step}} (\rho_t^r)^\top R_{0,N}^\top, \\ T_N^{\mathrm{end}} &= R_{0,N} \rho_t^N. \end{aligned} \tag{46}$$

Align the endpoint by $\bar{c}_N^{\mathrm{end},\sharp} = O_N \bar{c}_N^{\mathrm{end}}$ and $\Sigma_N^{\mathrm{end},\sharp} = O_N \Sigma_N^{\mathrm{end}} O_N^\top$, and define only the interleaved-to-endpoint defects

$$\epsilon_{\mathrm{split}}^{\mu} = \|\bar{c}_N^{\mathrm{fr},\sharp} - \bar{c}_N^{\mathrm{end},\sharp}\|_2, \qquad \epsilon_{\mathrm{split}}^{\Sigma} = \|\Sigma_N^{\mathrm{fr},\sharp} - \Sigma_N^{\mathrm{end},\sharp}\|_F. \tag{47}$$

With endpoint Lipschitz constants L_μ, L_Σ, the aggregate term used in Proposition 1 is $\epsilon_{\mathrm{disc}} := L_\mu(\epsilon_{\mathrm{cd}}^{\mu} + \epsilon_{\mathrm{split}}^{\mu}) + L_\Sigma(\epsilon_{\mathrm{cd}}^{\Sigma} + \epsilon_{\mathrm{split}}^{\Sigma})$. The split defects collect the homogeneous, affine-forcing, and injected-source differences between interleaved transport and one endpoint transport. They vanish for a fixed frame with the same frozen coefficients and for a single discrete step. The separate ϵ_{cd} terms compare the continuous flow with the optimizer discretization, so they may remain at one step. Ordered products matter because relaxation and rotation generally do not commute: in the constant continuous-time example, $e^{-(K+\Omega)T} \neq e^{-\Omega T} e^{-KT}$, with leading homogeneous splitting error $O(T^2\|[K,\Omega]\|)$. Persistent target shifts are propagated in parallel through their readout–shift cross-covariance as described in Section 2.1.

When the future frame is unknown, Procedure 1 obtains $\widehat{V}_j$ from clean mean propagation; the reported experiments apply the endpoint formula above. Proposition 1 follows the nonoverlapping chain true continuous $\to$ aligned continuous reference $\to$ aligned interleaved discrete $\to$ aligned endpoint implementation.

Ambient implementation of Procedure 1. For a shared fixed feature map, write $H = \Sigma + \lambda I$, $R = I - \eta H$, and $d = \Phi^\top y/P$. If S_r denotes the readout covariance and Q^{step} the one-step readout innovation covariance, the exact ambient updates

$$a_r = R^r a_0 + \eta \sum_{j=0}^{r-1} R^j d, \qquad S_r = R^r S_0 (R^r)^\top + \sum_{j=0}^{r-1} R^j Q^{\mathrm{step}} (R^j)^\top \tag{48}$$

are finite polynomial maps and hence remain continuous through repeated eigenvalues and crossings without selecting individual eigenvectors.

For learned features, first register the checkpoint replicas in a common parameterization and form the empirical anchor $(\bar{\theta}_t^{[M]}, \bar{a}_t^{[M]}, \Sigma_{\mathrm{mem},t}^{[M]})$. This registered mean defines one central parameter path; in settings without a meaningful registration, the analogous construction averages replica predictions directly in function space. Evaluate $\Phi_t = \phi_{\bar{\theta}_t^{[M]}}(X)$ and its Gram matrix $\Sigma_t^{\mathrm{feat}} = \Phi_t^\top \Phi_t / P = V_t \Lambda_t V_t^\top$, and initialize the moment state with $\bar{c}_t = V_t^\top \bar{a}_t^{[M]}$ and $\Sigma_{\mathrm{mem},t}^{[M]}$ in the same frame. Estimate g_t and Q_t^{step}, then apply the full-batch clean update defined by R_{clean} for N optimizer steps to obtain the predicted parameter and frame paths. The endpoint-split update transports the readout moments into that endpoint frame before evaluating the local loss. Ambient operators or grouped projectors make the same calculation continuous through repeated eigenvalues. In the plots, replacing the predicted frame by the measured endpoint frame defines a reference that isolates frame-prediction error.

B.4.2 Retrospective Local Error Bound

Figure 7 is a visual guide to the proof of Proposition 1. The horizontal chain compares four state constructions from the same anchor; the evaluation step adds only feature and subspace discrepancies. Each defect is assigned to one arrow, which prevents double counting.

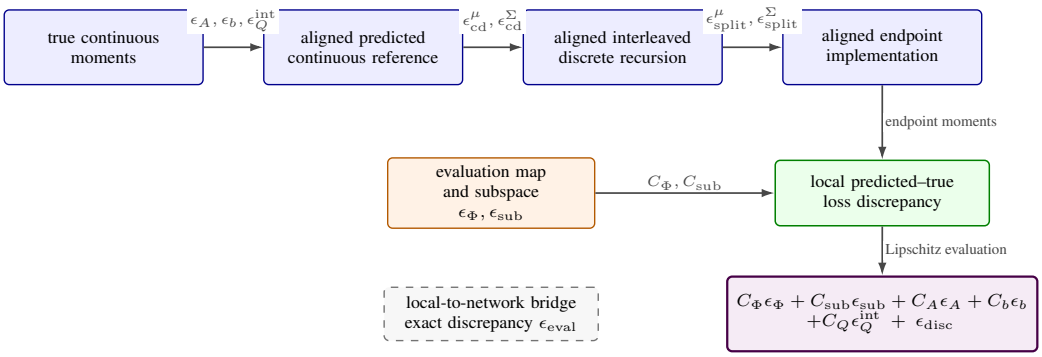


Figure 7: **Guide to the Proposition 1 proof.** Coefficient mismatch, continuous/discrete conversion, endpoint splitting, and evaluation-map mismatch occupy distinct arrows. A final triangle inequality adds ϵ_{eval} when the local quadratic evaluation is connected to the exact network.

Proof of Proposition 1. States use the Euclidean norm, transition maps its subordinate operator norm, and covariance errors the Frobenius norm. All hatted continuous quantities below are the aligned $\sharp$ quantities defined above, and every stage starts from the same anchor moments.

The proof has two layers. Steps 1–3 follow the four-object chain in Figure 7 within one conditioning field and one aligned chart. Step 4 then returns to the coarser anchor field. For the continuous comparison set

$$H_T := M_U e^{\omega_+ T}, \qquad R_\mu := \sup_{s\in[t,t+T]} \|\bar{c}_s^{\mathrm{cont}}\|_2, \qquad R_\Sigma := \sup_{s\in[t,t+T]} \|\Sigma_s^{\mathrm{cont}}\|_F. \tag{49}$$

The proposition's boundedness assumptions make these quantities finite. They expose how local amplification, mean size, and covariance size determine admissible constants; problem-specific estimates can sharpen them.

Step 1 (true continuous to aligned continuous reference). Let $e_c = \bar{c}^{\mathrm{true}} - \bar{c}^{\mathrm{cont}}$. Variation of constants gives the exact identity

$$e_c(t+T) = \int_t^{t+T} U(t+T,s)\{(A_s - \widehat{A}_s^\sharp)\bar{c}_s^{\mathrm{cont}} + (b_s - \widehat{b}_s^\sharp)\}\, ds. \tag{50}$$

Because $\|U(t+T,s)\|_2 \le H_T$ throughout the window,

$$\|e_c(t+T)\|_2 \le H_T\big(R_\mu\epsilon_A + \epsilon_b\big). \tag{51}$$

Using $\omega_+ = \max(\omega, 0)$ covers both stable and locally expanding propagators in one expression. For $D_s = \Sigma_s^{\mathrm{true}} - \Sigma_s^{\mathrm{cont}}$, subtraction of the two Lyapunov equations gives

$$\dot{D}_s = A_s D_s + D_s A_s^\top + (A_s - \widehat{A}_s^\sharp)\Sigma_s^{\mathrm{cont}} + \Sigma_s^{\mathrm{cont}}(A_s - \widehat{A}_s^\sharp)^\top + (Q_s - \widehat{Q}_s^\sharp). \tag{52}$$

The zero initial difference and Duhamel's formula yield

$$\begin{aligned} D_{t+T} = \int_t^{t+T} U(t+T,s)\big[&(A_s - \widehat{A}_s^\sharp)\Sigma_s^{\mathrm{cont}} + \Sigma_s^{\mathrm{cont}}(A_s - \widehat{A}_s^\sharp)^\top \\ &+ (Q_s - \widehat{Q}_s^\sharp)\big]U(t+T,s)^\top ds. \end{aligned} \tag{53}$$

Submultiplicativity of the Frobenius norm therefore gives the explicit estimate

$$\|D_{t+T}\|_F \le H_T^2\big(2R_\Sigma\epsilon_A + \epsilon_Q^{\mathrm{int}}\big). \tag{54}$$

The aligned generator already contains the connection, so $A - \widehat{A}^\sharp$ collects the full frame-rotation mismatch together with spectral relaxation error.

Step 2 (aligned continuous reference to aligned interleaved discrete reference). At t_j, compare the exact-flow maps $(\mathcal{P}_j^{\mathrm{cont}}, h_j^{\mathrm{cont}}, \mathcal{Q}_j^{\mathrm{cont}})$ with $(\mathcal{P}_j^{\mathrm{fr},\sharp}, h_j^{\mathrm{fr},\sharp}, \mathcal{Q}_j^{\mathrm{fr},\sharp})$. With $e_j = \bar{c}_{t_j}^{\mathrm{cont}} - \bar{c}_j^{\mathrm{fr},\sharp}$,

$$e_{j+1} = \mathcal{P}_j^{\mathrm{cont}} e_j + (\mathcal{P}_j^{\mathrm{cont}} - \mathcal{P}_j^{\mathrm{fr},\sharp})\bar{c}_j^{\mathrm{fr},\sharp} + (h_j^{\mathrm{cont}} - h_j^{\mathrm{fr},\sharp}), \tag{55}$$

and, for $D_j = \Sigma_{t_j}^{\mathrm{cont}} - \Sigma_j^{\mathrm{fr},\sharp}$,

$$\begin{aligned} D_{j+1} = \mathcal{P}_j^{\mathrm{cont}} D_j (\mathcal{P}_j^{\mathrm{cont}})^\top + (\mathcal{P}_j^{\mathrm{cont}} - \mathcal{P}_j^{\mathrm{fr},\sharp}) \Sigma_j^{\mathrm{fr},\sharp} (\mathcal{P}_j^{\mathrm{cont}})^\top \\ + \mathcal{P}_j^{\mathrm{fr},\sharp} \Sigma_j^{\mathrm{fr},\sharp} (\mathcal{P}_j^{\mathrm{cont}} - \mathcal{P}_j^{\mathrm{fr},\sharp})^\top + (\mathcal{Q}_j^{\mathrm{cont}} - \mathcal{Q}_j^{\mathrm{fr},\sharp}). \end{aligned} \tag{56}$$

Here $e_0 = 0$ and $D_0 = 0$ because the constructions share the anchor moments. The two displayed recursions are the discrete Duhamel formulas: each new one-step defect is propagated by all later transition maps. In particular, with empty products interpreted as identities,

$$\mathcal{P}_{N-1}^{\mathrm{cont}} \cdots \mathcal{P}_0^{\mathrm{cont}} - \mathcal{P}_{N-1}^{\mathrm{fr},\sharp} \cdots \mathcal{P}_0^{\mathrm{fr},\sharp} = \sum_{k=0}^{N-1} \mathcal{P}_{N-1}^{\mathrm{cont}} \cdots \mathcal{P}_{k+1}^{\mathrm{cont}} (\mathcal{P}_k^{\mathrm{cont}} - \mathcal{P}_k^{\mathrm{fr},\sharp}) \mathcal{P}_{k-1}^{\mathrm{fr},\sharp} \cdots \mathcal{P}_0^{\mathrm{fr},\sharp}. \tag{57}$$

Uniform bounds on the products and reference moments therefore give computable upper bounds on the endpoint defects in terms of

$$\sum_j \|\mathcal{P}_j^{\mathrm{cont}} - \mathcal{P}_j^{\mathrm{fr},\sharp}\|_2, \qquad \sum_j \|h_j^{\mathrm{cont}} - h_j^{\mathrm{fr},\sharp}\|_2, \qquad \sum_j \|\mathcal{Q}_j^{\mathrm{cont}} - \mathcal{Q}_j^{\mathrm{fr},\sharp}\|_F. \tag{58}$$

The endpoint norms themselves are $\epsilon_{\mathrm{cd}}^{\mu}$ and $\epsilon_{\mathrm{cd}}^{\Sigma}$ by definition; the sums above show how homogeneous-map, affine-forcing, source-quadrature, and connection-discretization errors accumulate into them.

Step 3a (endpoint implementation). Insert the four stages in the order shown in Figure 7. The last arrow contributes $L_\mu \epsilon_{\mathrm{split}}^{\mu} + L_\Sigma \epsilon_{\mathrm{split}}^{\Sigma}$, while Step 2 contributes the analogous ϵ_{cd} terms.

Step 3b (subspace alignment and evaluation). If the corresponding-cluster, same-rank, external-gap, and small-perturbation conditions of the classical Davis–Kahan estimate hold for every $s \in [t, t+T]$, with external gap $\gamma_s > 0$, then

$$\epsilon_{\mathrm{sub}} \le 2 \sup_{s \in [t, t+T]} \frac{\|\Sigma_s^{\mathrm{feat}} - \widehat{\Sigma}_s^{\mathrm{feat}}\|_2}{\gamma_s}. \tag{59}$$

In particular, if $\gamma_s \ge \gamma_* > 0$ uniformly, the right-hand side is at most $(2/\gamma_*) \sup_s \|\Sigma_s^{\mathrm{feat}} - \widehat{\Sigma}_s^{\mathrm{feat}}\|_2$. An endpoint-only Gram discrepancy controls the endpoint projector discrepancy, not this window supremum. To make the evaluation step explicit, let ϑ_s be the largest principal angle and $O_s = \mathrm{polar}(V_s^\top \widehat{V}_s)$. Then

$$\|V_s - \widehat{V}_s O_s^\top\|_2 = 2 \sin(\vartheta_s/2) \le \sqrt{2}\, \|V_s V_s^\top - \widehat{V}_s \widehat{V}_s^\top\|_2, \tag{60}$$

and consequently

$$\|\Phi_s V_s - \widehat{\Phi}_s \widehat{V}_s O_s^\top\|_2 \le \|\Phi_s - \widehat{\Phi}_s\|_2 + \sqrt{2} \|\widehat{\Phi}_s\|_2 \|V_s V_s^\top - \widehat{V}_s \widehat{V}_s^\top\|_2. \tag{61}$$

If r is the modeled subspace rank, the Frobenius-frame version carries the corresponding $\sqrt{2r}$ factor. Thus the Procrustes maps place moment and coefficient differences in the common chart, while the aligned evaluation discrepancy is absorbed into $C_\Phi \epsilon_\Phi + C_{\mathrm{sub}} \epsilon_{\mathrm{sub}}$.

Let L_μ, L_Σ be the endpoint Lipschitz constants in the proposition. Applying the evaluation Lipschitz bound only after the three state comparisons gives

$$\begin{aligned} |\widehat{E}_{\mathrm{test}}^{\mathrm{loc}} - E_{\mathrm{test}}^{\mathrm{loc}}| \le C_\Phi \epsilon_\Phi + C_{\mathrm{sub}} \epsilon_{\mathrm{sub}} + L_\mu \big[H_T (R_\mu \epsilon_A + \epsilon_b) + \epsilon_{\mathrm{cd}}^{\mu} + \epsilon_{\mathrm{split}}^{\mu} \big] \\ + L_\Sigma \big[H_T^2 (2 R_\Sigma \epsilon_A + \epsilon_Q^{\mathrm{int}}) + \epsilon_{\mathrm{cd}}^{\Sigma} + \epsilon_{\mathrm{split}}^{\Sigma} \big]. \end{aligned} \tag{62}$$

Hence one admissible choice is

$$C_A = L_\mu H_T R_\mu + 2 L_\Sigma H_T^2 R_\Sigma, \qquad C_b = L_\mu H_T, \qquad C_Q = L_\Sigma H_T^2, \tag{63}$$

and the remaining four endpoint terms are exactly ϵ_{disc}. This proves the proposition's pathwise display. The subsequent triangle inequality with ϵ_{eval} connects this local result to exact-network evaluation.

Step 4 (outer conditioning). When the conditioning field is $\mathcal{H}_{\mathrm{eval}} \supsetneq \mathcal{G}$, the preceding result reads $|\widehat{E}_{\mathcal{H}} - E_{\mathcal{H}}| \le B_{\mathcal{H}}$ pathwise under that field, where $\widehat{E}_{\mathcal{H}}$ and $E_{\mathcal{H}}$ are the predicted and true evaluation

losses conditional on $\mathcal{H}_{\text{eval}}$ and $B_{\mathcal{H}}$ denotes the right-hand side of the bound in Proposition 1. Define $\widehat{E}_{\mathcal{G}} := \mathbb{E}[\widehat{E}_{\mathcal{H}} \mid \mathcal{G}]$ and $E_{\mathcal{G}} := \mathbb{E}[E_{\mathcal{H}} \mid \mathcal{G}]$. Conditional Jensen gives

$$\begin{aligned} |\widehat{E}_{\mathcal{G}} - E_{\mathcal{G}}| &= \big|\mathbb{E}[\widehat{E}_{\mathcal{H}} - E_{\mathcal{H}} \mid \mathcal{G}]\big| \\ &\le \mathbb{E}[|\widehat{E}_{\mathcal{H}} - E_{\mathcal{H}}| \mid \mathcal{G}] \le \mathbb{E}[B_{\mathcal{H}} \mid \mathcal{G}]. \end{aligned} \tag{64}$$

If the fine-field predictor is already $\mathcal{G}$-measurable, its outer conditional expectation equals itself. This proves the coarser-field clause. □

B.4.3 Run-Level Calibration Design

The calibration reported in Figure 3(d) uses $n_{\text{cal}} = 48$ completed real-text runs (four corpora, two context lengths, two widths, and three seeds). Each run contributes the maximum normalized endpoint error over its common windows following the initial scale-estimation window, and the split-conformal quantile at $\alpha = 0.1$ is $q_{\text{conf}} = 0.530292$. The quantile, score formula, target grid, and ablations were fixed before the evaluation continuations were observed. The same-grid evaluation uses five seeds in every grid cell, for 80 complete runs and 960 endpoint forecasts. With complete calibration and target runs exchangeable under this fixed score mapping, the maximum score yields simultaneous run-level coverage (Lei et al., 2018); the window counts additionally show where individual intervals succeed or fail.

B.5 The Predictability Horizon

This subsection converts local drift and diffusion estimates into horizon statements. It separates the intrinsic causal floor from the avoidable Frozen penalty and records the constants needed to evaluate the certificate.

Proof of Theorem 1. Condition on $\mathcal{F}_t$. Variation of constants gives

$$\begin{aligned} c_{t+T} &= \mu_{t+T} + \zeta_{t+T}, \\ \mu_{t+T} &= U(t+T,t)c_t + \int_t^{t+T} U(t+T,s)b_s ds, \\ \zeta_{t+T} &= \int_t^{t+T} U(t+T,s)G_s\, dW_s. \end{aligned} \tag{65}$$

The paths A, b, Q are fixed under this conditioning; predictability and conditional square-integrability of G give $\mathbb{E}_t[\zeta_{t+T}] = 0$. For any $\mathcal{F}_t$-measurable $\widehat{c}_{t+T}$, expanding the squared Euclidean norm removes the conditional cross term, and conditional Itô isometry gives

$$\mathbb{E}_t\|c_{t+T} - \widehat{c}_{t+T}\|_2^2 = \|\mu_{t+T} - \widehat{c}_{t+T}\|_2^2 + \int_t^{t+T} \operatorname{Tr}\big(U(t+T,s)Q_sU(t+T,s)^\top\big)\, ds. \tag{66}$$

The first term is nonnegative and vanishes for $\widehat{c}_{t+T} = \mu_{t+T}$, proving the stated infimum and attainability. The first-crossing definition makes every $T < T_{\text{int}}(\varepsilon)$ feasible. Local boundedness of A gives $U(t+T,s) = I + O(T)$ uniformly for $s \in [t, t+T]$. If Q is continuous at t, the integral equals $T \operatorname{Tr} Q_t + o(T)$; if Q is locally Lipschitz, the remainder is $O(T^2)$. For $\operatorname{Tr} Q_t > 0$, the first-crossing definition and these local upper and lower estimates give $T_{\text{int}}(\varepsilon) \sim \varepsilon / \operatorname{Tr} Q_t$ as $\varepsilon \downarrow 0$. When $Q_t = 0$, a nonzero linear leading term is not implied.

For completeness, if $C \succeq 0$ is $\mathcal{F}_t$-measurable and $\|x\|_C^2 = x^\top Cx$, the same conditional orthogonality gives

$$\mathbb{E}_t\|c_{t+T} - \widehat{c}_{t+T}\|_C^2 = \|\mu_{t+T} - \widehat{c}_{t+T}\|_C^2 + \operatorname{Tr}\Big(C\Sigma^{(t)}(t+T)\Big), \tag{67}$$

$$\Sigma^{(t)}(t+T) = \int_t^{t+T} U(t+T,s)Q_sU(t+T,s)^\top ds. \tag{68}$$

Thus the conditional mean is optimal up to the nullspace of C, and the restarted Lyapunov covariance gives the unavoidable weighted state-MSE. A per-coordinate MSE uses $C = I/d$; centered per-logit MSE similarly includes its centering projection and position/vocabulary normalization. The Euclidean state bound is not itself a lower bound on scalar loss-prediction error. □

Proof of Proposition 2. First fix the comparison chart. For a complete moving frame, put $J_s = V_t^\top V_s$ and $x_s = J_s c_s = V_t^\top a_s$. Its coefficients are

$$A_s^0 = J_s A_s J_s^\top + \dot{J}_s J_s^\top = -J_s K_s J_s^\top, \qquad b_s^0 = J_s b_s, \qquad Q_s^0 = J_s Q_s J_s^\top. \tag{69}$$

Here $\dot{J}_s = J_s \Omega_s$, $J_t = I$, and $\Omega_s = (A_s^\top - A_s)/2$, so J_s is anchor-measurable under the theorem hypotheses. Orthogonality preserves the state-MSE and the trace of the transported diffusion covariance. Frozen holds A_t^0, b_t^0 fixed in this chart. In the remainder of this proof, write c, A, b, Q for these fixed-chart quantities; in particular, the derivative bounds below are fixed-chart bounds. Holding $-K_t - \Omega_t$ constant in an evolving spectral chart, or simply deleting Ω_s, is a different approximation.

Condition throughout on $\mathcal{F}_t$. Assume A_s, b_s are differentiable and bounded on $[t, t + T_{\max}]$, and that $\|U(r,s)\|_2 \le M_U e^{\omega_+(r-s)}$ there, where $M_U \ge 1$ and $\omega_+ := \max(\omega, 0)$. Write

$$L_A := \sup_{s \in [t, t+T_{\max}]} \|\dot{A}_s\|_2, \qquad L_b := \sup_{s \in [t, t+T_{\max}]} \|\dot{b}_s\|_2. \tag{70}$$

Set $H_{\max} = M_U e^{\omega_+ T_{\max}}$ and $R_{\mu,\max} = \sup_{s \in [t, t+T_{\max}]} \|\widehat{\mu}_s^{\mathrm{fr}}\|_2$, which is finite by the stated local boundedness. Fix any $0 \le \Delta \le T_{\max}$. Let $\mu_s = \mathbb{E}_t[c_s]$ and let $\widehat{\mu}_s^{\mathrm{fr}}$ solve $\dot{\widehat{\mu}}_s^{\mathrm{fr}} = A_t \widehat{\mu}_s^{\mathrm{fr}} + b_t$ from $\widehat{\mu}_t^{\mathrm{fr}} = c_t$. For $d_s = \mu_s - \widehat{\mu}_s^{\mathrm{fr}}$, variation of constants gives

$$d_{t+\Delta} = \int_t^{t+\Delta} U(t+\Delta, s)\big[(A_s - A_t)\widehat{\mu}_s^{\mathrm{fr}} + (b_s - b_t)\big] ds. \tag{71}$$

The fundamental theorem of calculus, $\|U(t+\Delta, s)\|_2 \le H_{\max}$, and $\|\widehat{\mu}_s^{\mathrm{fr}}\|_2 \le R_{\mu,\max}$ imply

$$\|d_{t+\Delta}\|_2^2 \le \frac{H_{\max}^2 R_{\mu,\max}^2}{2} \Delta^4 L_A^2 + \frac{H_{\max}^2}{2} \Delta^4 L_b^2. \tag{72}$$

Applying the orthogonal decomposition in the proof of Theorem 1 to $\widehat{c} = \widehat{\mu}^{\mathrm{fr}}$ gives the exact excess identity

$$\mathbb{E}_t \|c_{t+\Delta} - \widehat{\mu}_{t+\Delta}^{\mathrm{fr}}\|_2^2 - q_t(\Delta)\Delta = \|d_{t+\Delta}\|_2^2, \tag{73}$$

and the preceding display proves the $D_t \Delta^4$ bound with $D_t := H_{\max}^2 R_{\mu,\max}^2 L_A^2/2 + H_{\max}^2 L_b^2/2$. Coefficient drift therefore integrates to an $O(\Delta^2)$ mean error and squares to an $O(\Delta^4)$ state-MSE penalty. Since $q_t(T) \le \bar{q}_t$ on the local interval,

$$\mathrm{MSE}_{\mathrm{Frozen}}(T) \le D_t T^4 + \bar{q}_t T. \tag{74}$$

With $D_t + \bar{q}_t > 0$, the polynomial $D_t T^4 + \bar{q}_t T$ is continuous, strictly increasing on $T \ge 0$, and unbounded, so the stated equation has one nonnegative solution and every earlier horizon within the local cap satisfies the certificate. □

B.6 Local Cross-Entropy Geometry

The text experiments optimize cross-entropy rather than squared loss. Locally, the same moving-frame construction applies after replacing the ordinary feature Gram by the softmax Fisher-weighted Gram.

Softmax readout geometry. For C classes with readout $A \in \mathbb{R}^{d \times C}$, sample i has logits $z_i = A^\top \phi_i$, probabilities $p_i = \mathrm{softmax}(z_i)$, and empirical cross-entropy (CE) gradient $\nabla_A L = \frac{1}{P} \sum_i \phi_i (p_i - e_{y_i})^\top$. In this CE subsection, $\mathrm{vec}(A)$ stacks the rows of A (equivalently, it column-vectorizes $A^\top$). With this convention the Hessian is

$$H_A = \frac{1}{P} \sum_{i=1}^{P} (\phi_i \phi_i^\top) \otimes F_i, \qquad F_i = \mathrm{Diag}(p_i) - p_i p_i^\top, \tag{75}$$

the categorical Fisher matrix: CE replaces the MSE feature Gram by a Fisher-weighted feature Gram. (Under the standard column-wise $\mathrm{vec}(A)$ convention the two Kronecker factors are reversed.)

Fix a checkpoint $\bar{A}_t$ over one local window, set $a = \mathrm{vec}(A)$, $\bar{a}_t = \mathrm{vec}(\bar{A}_t)$, and let $g_{\mathrm{CE},t}^{\mathrm{grad}} = \nabla_a L_{\mathrm{CE}}(\bar{a}_t)$ and $H_t^{\mathrm{CE}} = V_t^{\mathrm{CE}} \Lambda_t^{\mathrm{CE}} (V_t^{\mathrm{CE}})^\top$. For $c = (V_t^{\mathrm{CE}})^\top (a - \bar{a}_t)$ define

$$r_{\lambda,t} = (V_t^{\mathrm{CE}})^\top (g_{\mathrm{CE},t}^{\mathrm{grad}} + \lambda \bar{a}_t). \tag{76}$$

Taylor expansion at the fixed checkpoint gives

$$\dot{c} = -\eta\big[(\Lambda_t^{\mathrm{CE}} + \lambda I)c + r_{\lambda,t}\big] + O(\eta\|c\|^2). \tag{77}$$

When $r_{\lambda,t}$ lies in the modeled range, this may equivalently be written as

$$\dot{c} = -\eta(\Lambda_t^{\mathrm{CE}} + \lambda I)(c - g_{\lambda,t}^{CE}) + O(\eta\|c\|^2), \qquad g_{\lambda,t}^{CE} = -(\Lambda_t^{\mathrm{CE}} + \lambda I)^{\dagger} r_{\lambda,t}. \tag{78}$$

Thus weight decay changes both the curvature and the checkpoint gradient; for $\lambda = 0$ this reduces to the usual local Newton displacement on the Fisher range.

For centered local logit perturbations, the second-order contribution to the replica-averaged loss is correspondingly $\frac{1}{2}\operatorname{Tr}(F(\bar{z})\operatorname{Cov}_m[\delta z_m])$. This Fisher geometry supplies the local bridge from the quadratic derivation to the character-level and LLM experiments, which evaluate the actual cross-entropy and negative log-likelihood.

C Experimental Details

This section records the settings and statistics used for the numerical claims. The supplement provides experiment and figure code and deterministic and stochastic Lean 4 projects.

C.1 Forecast Protocol and Conditioning

Table 2 records the architectures and training settings for the learned-feature experiments. The exact grids for the controlled closure experiments are constants in the released scripts.

For finite M, the implementation uses $\bar{c}_s^{[M]} = M^{-1}\sum_m c_{m,s}$ and $\Sigma_{\mathrm{mem},s}^{[M]} = (M-1)^{-1}\sum_m (c_{m,s} - \bar{c}_s^{[M]})(c_{m,s} - \bar{c}_s^{[M]})^{\top}$, or an explicitly declared covariance with $1/M$ normalization. Under conditionally independent and identically distributed replicas, the $1/(M-1)$ covariance is unbiased, whereas the $1/M$ version has expectation $(1-1/M)\Sigma_{\mathrm{mem}}$ and the same factor in its diffusion drift. A realized covariance also has a martingale term. If the common evaluation path is not $\mathcal{G}_{t_0}^{\mathrm{shared}}$-measurable, the identity in (2) is first applied conditionally on $\mathcal{H}_{\mathrm{eval}}$ and then outer-averaged. Feature–readout cross terms, persistent-target cross-covariance not carried in the state, and finite-M anchor estimation remain in the endpoint discrepancy.

C.2 Theory-to-Experiment Correspondence

Table 3 connects each theoretical object to the experiment that makes it observable. A *direct* check evaluates the proposition's specialized event, a *component* check measures one ingredient of a larger bound, and an *intervention/proxy* check changes one mechanism to reveal its effect. Independent problems or complete runs form the replication units, with intervals nested within them.

C.2.1 Learned-Feature Closure Protocol

The first evaluation uses 50 independent width-32 tanh-MLP problems, a rank-8 frame, and four horizons; Procrustes alignment places every measured defect in one chart. The exact-network evaluation uses 64 minibatch replicas at each of 16 full-rank anchors and measures $\epsilon_{\mathrm{eval}} = |E_{\mathrm{test}}^{\mathrm{exact}} - E_{\mathrm{test}}^{\mathrm{loc}}|$. Under the shared-feature quadratic specialization, the maximum numerical residual in identity (2) is 1.67×10^{-16}. Fluctuations from estimating population moments with 64 replicas provide the finite-ensemble uncertainty around this comparison. Table 3 reports the closure and run-level calibration statistics.

C.2.2 Controlled Tests of the Bounds

These controlled tests make every coefficient path and bound constant available. They evaluate the seven Proposition 1 groups—feature map, subspace, drift operator, affine forcing, source covariance, continuous/discrete conversion, and endpoint splitting—on 48 stable two-dimensional affine diffusions. Separate experiments test the horizon, curvature, finite-M, and total covariance. The horizon evaluation uses 40 systems and four fixed tolerances and recomputes the true derivative suprema. It

Table 2: Experimental configurations. "Replicas" are nested Monte Carlo or data-order repetitions; the problem or complete run is the independent unit. All MLPs have one hidden layer and scalar output. Transformer feed-forward width is $4d_{\text{model}}$, with pre-layer normalization, Gaussian error linear-unit activations, and zero dropout.

Experiment	Network and data	Optimization and horizon	Replication and evaluation
Moving-frame MLP	Tanh MLP, input 18, width 28; $360/1{,}500$ Gaussian train/test inputs and a nonlinear single-index target. Weights are Gaussian with fan-in scaling and readouts have scale $0.2/\sqrt{28}$.	Full-batch gradient descent (GD) on replica-fixed Gaussian-label targets ($\sigma = 0.35$), $\eta = 0.08$; hidden/readout weight decay $5\times10^{-4}/2\times10^{-3}$. The local experiment uses 420 updates and the long-horizon experiment 480; their checkpoint grids are fixed in the released code.	50 problems, 48 replicas per problem; 550 local intervals. Test MSE is averaged over replicas; one-shot and rolling forecasts use the same runs.
Predictability crossover (Figure 1)	Tanh MLP, input 14, width 32; $384/1{,}024$ Gaussian train/test inputs and target $\tanh(1.5x^\top u) + 0.25\sin(1.2x^\top v)$ for fixed orthonormal directions u, v. Fan-in Gaussian hidden initialization; readout scale $0.2/\sqrt{32}$.	Full-batch GD for 384 anchor updates, then minibatch SGD for 2,400 updates; $\eta = 0.0025$, batch 48, hidden/readout weight decay $5\times10^{-4}/2\times10^{-3}$; recorded every four updates.	16 problems (seeds 52000–52015), 96 post-anchor replicas each. All hidden and readout parameters are updated.
Proposition 1 closure	The same width-32 tanh family with $384/768$ train/test examples. The primary local state is the leading rank-8 readout frame; the exact-network arm also uses full rank 32.	384 full-batch anchor updates plus 1,200 continuation updates; readout rate 0.0025, hidden-rate multiplier 2, weight decays as above. The affine source uses label-noise scale 0.35 and batch 48; endpoints are 200, 400, 800, 1200.	Local budget: 50 problems (seeds 62000–62049), with moments transported directly. Exact loss: 16 problems (seeds 63000–63015), 64 independent minibatch replicas each.
Real-text moving-frame Transformer	Character next-token prediction on Shakespeare, Austen, code, and BabyLM. One Transformer block; contexts $32/64$, $d_{\text{model}} = 24/48$, heads $2/4$, learned token/position embeddings, and a linear vocabulary head. Each run uses $1{,}200/1{,}200$ fixed train/test windows from an $85/15$ corpus split and at most 64 character symbols.	Minibatch SGD, batch 128; 140 anchor and 140 continuation updates. The learning rate is 0.018 at width 24 and 0.014 at width 48. Thirteen successive forecast windows are evaluated.	Four corpora, two contexts, two widths, and three seeds: 48 independent runs and 624 nested intervals, with four continuation replicas per run. Calibration is evaluated on 80 complete runs over the same structured grid.

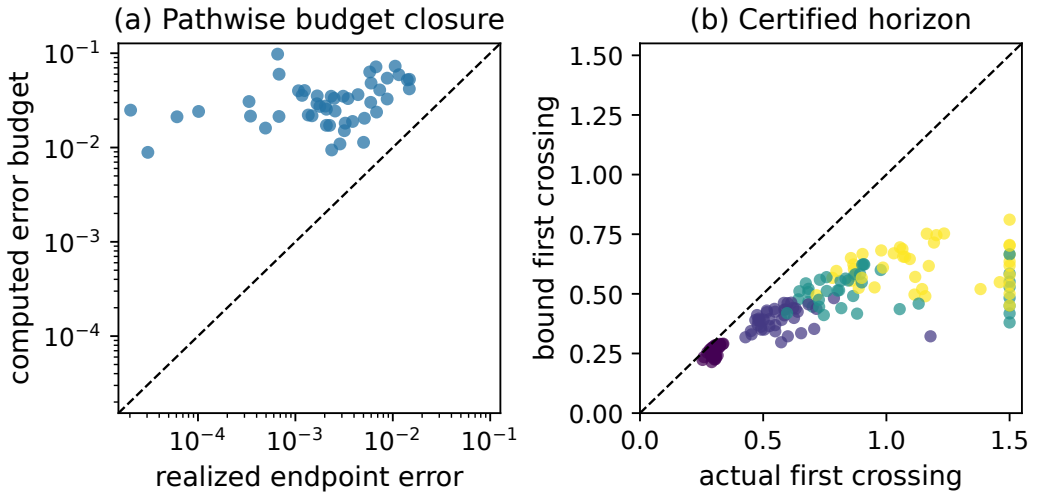


Figure 8: Controlled closure tests. (a) The computed seven-term budget against the realized endpoint error on 48 independent time-varying systems; all points lie above the equality line. (b) First tolerance crossing from the horizon bound against the realized state-MSE crossing for 40 systems and four fixed tolerances; a valid certificate lies on or below the diagonal.

Table 3: Term-by-term empirical evaluation of the reliability theory. Bootstrap intervals resample the independent problem or run units, preserving the nesting of forecast intervals.

Theoretical object	Experimental check and observed statistic	Evidential status
$A = -K - \Omega$ and $b = Kg$; ingredients of ϵ_A, ϵ_b	Across 40 fixed-readout cells, the median relative drift-curve error is 0.00549 (problem-bootstrap 95% interval $[0.00415, 0.00791]$), and the median relative rotation-curve error is 0.00164 $[0.00135, 0.00205]$. The learned-feature rotation law has median relative error 0.006 (MLP–SGD), 0.013 (MLP–Adam), and 0.032 (Transformer–Adam).	Component validation in this evaluation; the complete learned-feature row below evaluates both path integrals directly.
Q and Lyapunov covariance transport; ingredient of $\epsilon_Q^{\mathrm{int}}$	The same evaluation gives median replica-memory error 0.00615 $[0.00509, 0.00696]$; memory variance is linear in the injected σ^2/B source, and the full anisotropic Lyapunov solution has lower error than the scalar-rate fit.	Direct in the fixed-readout specialization and component-level here; the complete learned-feature row below evaluates $\epsilon_Q^{\mathrm{int}}$ directly.
Feature/evaluation defects $\epsilon_\Phi, \epsilon_{\mathrm{sub}}$	In 50 MLP problems (550 nested intervals), frame angle has Spearman rank correlation $\rho_s = 0.899$ with frozen-minus-observed-frame error; a problem-cluster bootstrap gives $[0.881, 0.915]$. The problem-level moving-frame/observed-frame error ratio is 1.005 $[0.987, 1.023]$.	Intervention/proxy support in this comparison; the complete learned-feature row below measures both window suprema directly.
Continuous/discrete and endpoint-splitting term ϵ_{disc}	Multi-step propagation has lower error than a first-order Euler frame update in $47/50$ problems and $475/550$ intervals. At a designed eigenvalue crossing, the per-mode implementation has $2400\times$ larger error than the operator implementation.	Implementation-level support in this contrast; the complete learned-feature row below separates ϵ_{cd} and $\epsilon_{\mathrm{split}}$.
Run-level split-conformal calibration	With $q = 0.5303$ fixed using 48 calibration runs at $\alpha = 0.1$, evaluation on 80 complete curves covers $77/80$ curves (Wilson 95% interval $[0.895, 0.987]$) and $957/960$ windows.	Run-level calibration check under complete-run exchangeability and the fixed score mapping.
Causal diffusion floor and conditional-mean optimum, Theorem 1	In 10 fixed-feature problems with four nested branch groups per anchor and 512 branches per group, the median ratios of conditional-mean MSE and empirical branch variance to the analytic floor are 1.0007 and 1.0012 for minibatch SGD; all estimates agree with the floor within three Monte Carlo standard errors.	Direct state-space check that branch variance realizes the diffusion floor.
Complete learned-feature budget and nonlinear ϵ_{eval}	On 50 learned-MLP paths and four horizons, the rank-8 local budget has $0/200$ violations; the maximum ratio is 0.366 and the problem-level maximum-ratio median is 0.277 $[0.251, 0.307]$. In the full-rank exact-network evaluation (16 problems, 64 replicas each), $\mathcal{B}_{\mathrm{loc}} + \epsilon_{\mathrm{eval}}$ has $0/64$ violations, maximum ratio 0.410, and median $\epsilon_{\mathrm{eval}}/\mathcal{B}_{\mathrm{loc}} = 0.00748$.	Direct learned-feature check using completed-window quantities. The measured ϵ_{eval} connects the local bound to exact-network loss; full rank exposes this connection without rank-8 truncation dominating it.

Table 4: Controlled evaluations and the direct neural-network horizon-scaling check. Counts use independent generated systems except where the row describes an analytic spread sweep.

Statement	Result	Interpretation
Full local budget, Proposition 1	$0/48$ violations after measuring all seven defect groups on each path; median (maximum) realized/budget ratio 0.0927 (0.441).	Direct designed closure with valid local constants; it complements the separate completed-window learned-network check in Table 3.
Frozen drift penalty and certified horizon, Proposition 2	$0/60{,}000$ pointwise bound violations and $0/160$ cases in which the first crossing of the sharper horizon-dependent bound exceeded the realized first crossing; median bound/realized horizon 0.695, rank correlation 0.868.	Direct check of the completed-window inequality using measured derivative suprema; the plotted horizon uses the sharper horizon-dependent envelope.
Joint neural-network predictability crossover	Early-time diffusion and Frozen-drift exponents are 0.960 and 4.017; the anchor predicts the measured median switch within 47 updates ($1{,}055$ vs. $1{,}008$). At $\varepsilon = 3.5 \times 10^{-5}$, all 16 MLP problems cross and the median within-problem moving-frame/Frozen horizon ratio is 1.922.	Same-system component and method comparison in an end-to-end network, with 96 replicas nested within each independent problem.
Curvature and finite M, Lemma 1	Clean mean-drift bias scales as $\sigma^{2.01}$; after the covariance correction the residual scales as $\sigma^{3.00}$. Mean-estimation root-mean-squared error (RMSE) scales as $M^{-0.503}$.	Direct designed rate checks under bounded standardized moments and conditionally iid replicas.
Random coefficient paths and conditional covariance transport	Total-variance prediction/empirical ratio is 1.0015 (reciprocal empirical/theory ratio 0.9985); the between-path term is 64.7% of total variance, and omitting it underestimates variance by $2.83\times$.	Direct check that between-path mean variation completes the coarse-field covariance.

plots the sharper pointwise bound with the exact $q_t(T)T$ term. Exact grids, integrator steps, seeds, and constants accompany the code.

Figure 1 uses 16 width-32 MLP problems and 96 post-anchor replicas per problem. The displayed tolerance makes every problem's crossing visible on a common scale. Figure 9 then varies width, anchor, and batch size with eight independent problems per condition.

C.3 Prediction Experiments

The fixed-feature readout evaluation spans 40 cells varying noise, sample count, width, and weight decay. The median relative errors of the drift, replica-memory, signal-to-noise-ratio (SNR), and rotation curves are 0.00549, 0.00615, 0.00807, and 0.00164. The same records confirm linear

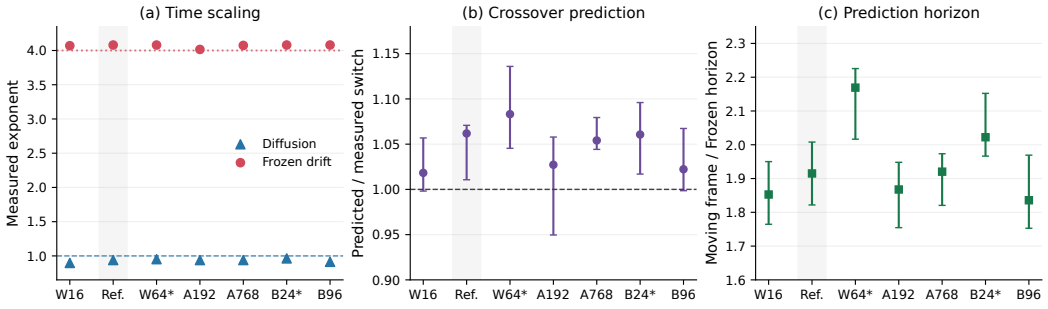


Figure 9: Robustness of the predictability-crossover mechanism. W, A, and B vary width, anchor step, and batch size; the shaded column is the reference setting. (a) Diffusion remains near T^1 and Frozen drift near T^4. (b) The anchor predicts the measured component switch. (c) Moving-frame prediction extends the horizon in every condition. All 56 problem–condition cells exhibit a component switch, and the condition-wise median moving-frame/Frozen horizon ratio ranges from 1.836 to 2.169.

σ^2/B source scaling and show lower covariance error for anisotropic Lyapunov transport than for a scalar-rate approximation.

Across 50 independent MLP problems, causal moving-frame propagation has lower error than the frozen frame in all 50 problem-level comparisons and $546/550$ intervals, with a median paired frozen-to-moving error ratio of 8.75. On character-level text it has lower error in $619/624$ intervals and all 48 runs; repeated re-anchoring maintains low error over the full continuation (median relative endpoint NLL error 1.07×10^{-4} vs. 1.39×10^{-3} for the frozen frame). The measured-endpoint reference isolates the contribution of frame prediction. Separately, the rotation law approximates measured frame motion with median relative error 0.006–0.032, and the operator implementation remains stable at a designed eigenvalue crossing where a per-mode implementation has $2400\times$ larger error.

Matched related-work baselines. Official LC-PFN and FT-PFN checkpoints and the public `pow3` implementation receive the same loss prefix observed through each anchor, without configuration covariates. On the matched 400 MLP intervals, median independent-problem errors are 0.367, 0.0622, 0.0219, and 0.00652 for `pow3`, FT-PFN, LC-PFN, and the moving-frame predictor. On 480 matched char-NTP intervals they are 0.00950, 0.00253, 0.00259, and 1.38×10^{-4}. The moving-frame predictor has the lowest error in all 50 MLP problems and all 48 text runs.

Long-horizon re-anchoring and text prediction. On the same 50 MLP problems, one-shot moving-frame/Frozen median errors are $0.0138/0.0769$, while rolling re-anchoring reduces them to $0.00401/0.0387$; the moving-frame predictor has lower error in every paired comparison. The character-NTP evaluation spans four corpora, two context lengths, two widths, and three seeds (48 runs, 624 intervals), with the full results summarized above and in Table 1.

Run-level calibration check. The 48 calibration runs define one maximum-over-windows score per complete run and give $q = 0.5303$. Evaluation covers $77/80$ complete curves and $957/960$ windows. Under complete-run exchangeability, the maximum score transfers the calibration jointly across all windows of a new run.

C.4 LLM-Scale Same-Anchor Continuations

This experiment tests both the stochastic target of the theory and the moving-frame point predictor at substantially larger parameter count. We start from the public EleutherAI Pythia-1B checkpoint (1.012 billion parameters), perform 32 deterministic pre-anchor updates, and save the model and complete nonzero AdamW state. From this common anchor, 48 continuations differ only in their post-anchor minibatch sequences. Measurements follow $1, 2, 4, 8, 16, 32$, and 64 updates, with 4,096 processed tokens per update. All branches share a fixed 32-example validation set; eight additionally

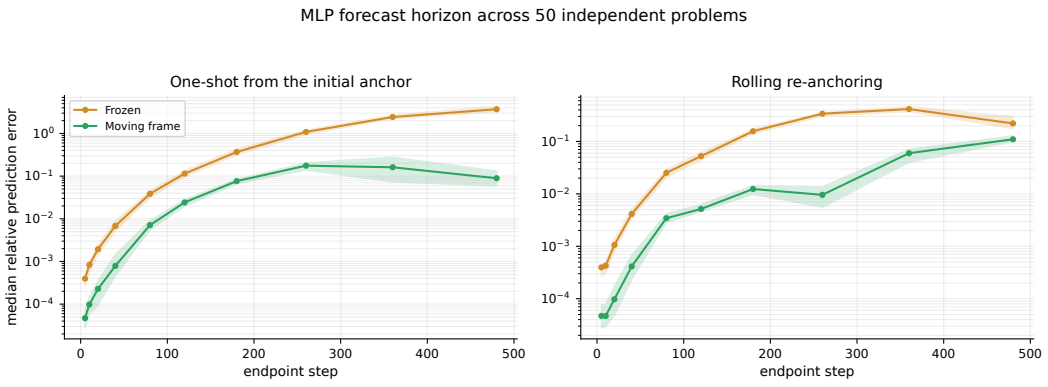


Figure 10: Forecast horizon across 50 independent learned-feature MLP problems. One-shot forecasts accumulate frame-curvature error with distance from the initial anchor (left); rolling re-anchoring repeatedly restores the local approximation (right). Curves are problem medians and bands are interquartile ranges.

retain full logits and final hidden states at the same $N_{\mathrm{probe}} = 256$ token positions. A branch is the independent sampling unit.

Proposed predictor at LLM scale. Procedure 1 predicts endpoint features by deterministic propagation and advances the vocabulary readout to the same endpoint. Frozen pairs that readout with anchor features, isolating the value of feature transport. Both forecasts start at the common anchor and are evaluated against stochastic continuations; AdamW and cross-entropy are propagated in optimizer coordinates.

Eight branches retain logits at $N_{\mathrm{probe}} = 256$ shared positions. Centering over the $|\mathcal{V}| = 50{,}304$ vocabulary entries preserves probabilities and NLL. We average branch variance over positions and vocabulary to obtain $D_{\mathrm{logit}}(T)$, and use the analogous squared distance for prediction error.

Response-masked instruction tuning. On response-masked Dolly-15k, the proposed predictor has lower function-space MSE than Frozen for every branch and horizon: median ratio 0.260, with step-64 MSE 0.0334 versus 0.0912. Final median absolute probe-NLL errors are 0.0163, 0.0269, and 0.0295 nats/token for the proposed predictor, Frozen, and persistence.

Their stochastic spread follows $D_{\mathrm{logit}}(T) \propto T^{0.851}$ over steps 16–64 ($R^2 = 0.968$; 95% interval $[0.763, 0.937]$), consistent with a horizon-dependent rate in $q_t(T)T$. Final full-validation NLL standard deviation is 4.23×10^{-3} across 48 branches.

Code next-token prediction. On disjoint 512-token blocks of public NumPy text, the proposed predictor again has lower MSE than Frozen for every branch and horizon: median MSE ratio 0.400, with step-64 MSE 0.0523 versus 0.1399. Final median absolute probe-NLL errors are 0.0222, 0.0385, and 0.0363 for the proposed predictor, Frozen, and persistence.

Here $D_{\mathrm{logit}}(T) \propto T^{0.792}$ over steps 16–64 ($R^2 = 0.940$; 95% interval $[0.735, 0.849]$), and final full-validation NLL standard deviation is 3.56×10^{-3} across 48 branches.

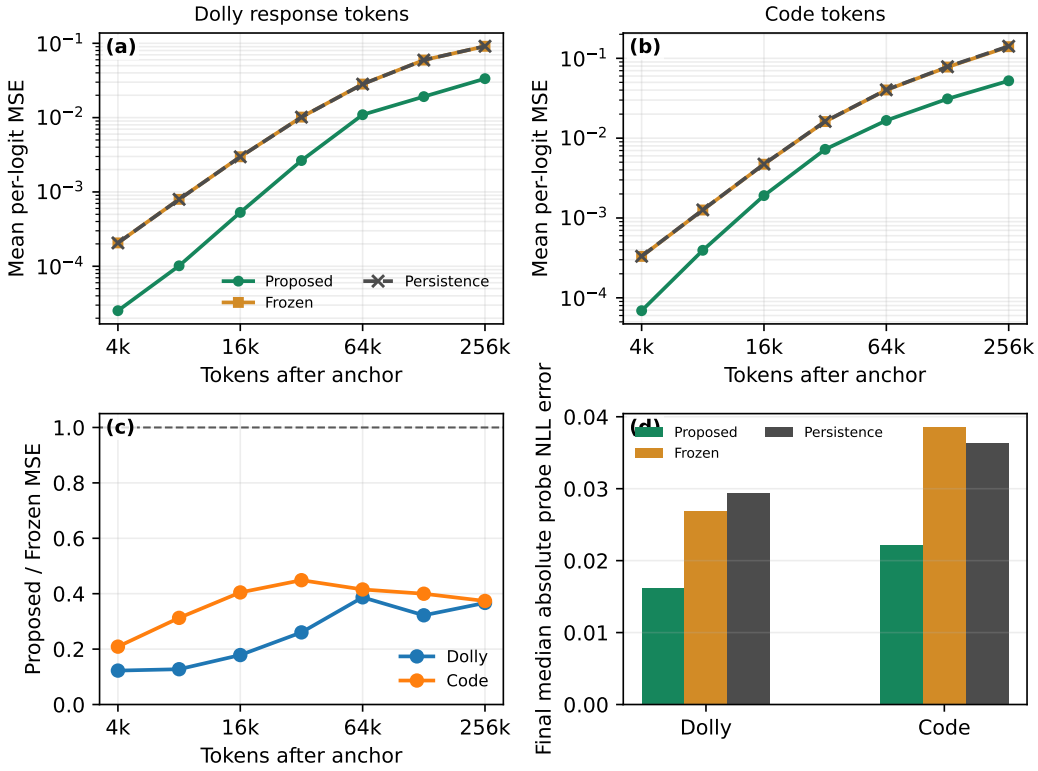


Figure 11: Pythia-1B results. (a,b) Per-logit MSE against eight same-anchor Dolly and code continuations. (c) Proposed/Frozen MSE; values below one favor feature transport. (d) Median absolute probe-NLL error at the final 64-update endpoint.