WORLD MODELS WITH PREDICTABLE LONG-HORIZON MARGINALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate one-step predictions do not ensure that a world model's rollouts retain the data distribution. We make the model's decoded stationary law explicit by learning a decoder of a fixed Gaussian reference and constraining the behaviour-averaged transition to preserve that reference. For controlled systems, a joint transition uses a conditional action chart to preserve behaviour occupancy without requiring invariance at each fixed action. Joint state–action rotations and parallel Gaussian noise give an exactly preserving transition with a tractable conditional density. We derive an absolute convergence bound from finite initialization banks and control departure from the reference through conditional action-space divergence. Across 216 fitted pixel checkpoints on twelve control tasks, the occupancy model with a reference mixture retains every evaluated chain at 10^5 steps in all 36 task–seed cells, with a rollout-minus-reference energy-statistic difference of -0.0002 ± 0.0003 (training-seed standard error). Each of the four nonpreserving comparison arms loses chains, although the Gaussian arm is more accurate at ten steps. An offline DreamerV3 reference also achieves better short-horizon accuracy. These results distinguish three properties of a world model: the distribution it approaches, the rate of approach, and the conditional dynamics it learns.

1 INTRODUCTION

World models support prediction and planning by repeatedly applying a learned transition. Their outputs become future inputs, so accurate one-step predictions are insufficient: a rollout can collapse, expand, or leave the region represented in the data. Conversely, trajectories in a chaotic system can separate even when both obey the correct dynamics. The long-horizon marginal is therefore a distinct object of prediction.

Teacher-forced likelihood scores transitions under the data law, whereas a rollout visits the model's evolving distribution (Appendix A.6). Even in a first-order autoregressive process, a coefficient perturbation can incur second-order conditional divergence while changing the invariant variance to first order (Appendix I.3). Small local error alone need not control the long-run law.

We address this problem by **making the model's stationary law explicit**. A stationary data law satisfies $\mu\bar{K}^\star = \mu$, where $\bar{K}^\star$ averages the true transition under the behaviour policy. In suitable latent coordinates, this identity becomes preservation of a standard Gaussian reference γ. We learn a decoder of this reference and preserve it in the transition; matching the resulting observation law p_θ to data remains a learned approximation. The transition determines the conditional dynamics and how quickly other initial laws approach that reference. For a measured discrepancy from data, this yields

$$\underbrace{D_T}_{\text{rollout discrepancy}} = \underbrace{D_\infty}_{\text{decoded-reference discrepancy}} + \underbrace{r_T}_{\text{signed transient}}, \qquad |r_T| \le \text{an explicit bound.}$$

Independent reference draws estimate this model-implied limiting level before any long rollout. The transition must still learn useful dynamics.

Invariant-measure objectives fit long-run statistics (Schiff et al., 2024), and learned Markov chain Monte Carlo transitions preserve a specified target (Song et al., 2017). A controlled world model must also fit action-dependent dynamics. Our construction combines behaviour occupancy preservation, an exact conditional density, and finite-initialization bounds in one parallel transition.

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.

Requiring invariance at each action can exclude valid controlled dynamics. We instead Gaussianize behaviour actions and preserve the joint state–action reference, enforcing the correct averaged constraint (Figure 1).

Our contributions are:

- **A controlled preserving architecture.** A joint state–action mixer and parallel noise stage enforce behaviour occupancy while retaining an exact conditional density. Representation and affine-coupling analysis identify the structural constraints (§2–3).
- **Computable marginal guarantees.** Applying Gaussian entropy contraction to the joint innovation gives an absolute finite-bank bound and controls policy deviation in physical action-space Kullback–Leibler (KL) divergence (§4).
- **A separation of model properties.** Static-reference comparisons, controlled examples, and 216 pixel checkpoints distinguish marginal predictability from conditional accuracy and planning return (§5).

2 Stationary data as a preserved reference

Let μ be a stationary state law and $K^\star(x, a, \mathrm{d}x')$ its controlled transition under a fixed behaviour policy π_b. The relevant constraint is

$$\iint \mu(\mathrm{d}x)\pi_b(\mathrm{d}a \mid x)K^\star(x, a, \mathrm{d}x') = \mu(\mathrm{d}x'). \tag{1}$$

It does not require every action kernel to preserve μ: independent uniform $a \in \{0, 1\}$ and $x' = a$ give stationary uniform states, although each fixed action collapses the state law.

A stationary representation. Let $\gamma_d = \mathcal{N}(0, I_d)$ and let $E_\#\mu$ denote the pushforward of μ by E. For an invertible chart satisfying $E_\#\mu = \gamma_d$, transport of the kernel gives $\mu\bar{K} = \mu \iff \gamma_d\tilde{K} = \gamma_d$. With arbitrary measurable charts and kernels, this represents precisely the pair laws with equal atomless marginals (Proposition 7). Differentiable density representations require the transport regularity stated in Appendix A.1. Finite networks cover only part of this constraint set; the theorem characterizes pair laws, not the reach of any fixed architecture.

The conditional action chart. In latent coordinates, transform the behaviour action law by

$$\alpha = G(a \mid z), \qquad G(\cdot \mid z)_\#\pi_b(\cdot \mid z) = \gamma_m. \tag{2}$$

Then α is Gaussian independently of z. A transformation preserving the *joint* reference of (z, α) can move information from action to state while preserving occupancy. With a state-independent chart, joint preservation strictly contains fixed-action preservation. With state-dependent charts, these classes need not be nested: fixing a chart coordinate and fixing a physical action are different interventions (Proposition 27, Remark 28).

For the recurrent pixel model, a frozen tanh-Gaussian policy reads the current frame o: $a = \tanh(m_b(o) + s_b \odot \alpha)$, with known $s_b > 0$. Recorded pre-tanh actions provide the exact training chart. During generation the policy reads the decoded observation, so its innovation remains independent of the latent state (Appendix A.4). A learned approximate chart requires additional residual-error control.

3 Gaussian-preserving transitions

For a diffeomorphism T, Gaussian preservation is equivalent to

$$\log|\det J_T(x)| = \tfrac{1}{2}\big(\|T(x)\|^2 - \|x\|^2\big). \tag{†}$$

It couples radial displacement to volume change and determines the admissible form of affine couplings.

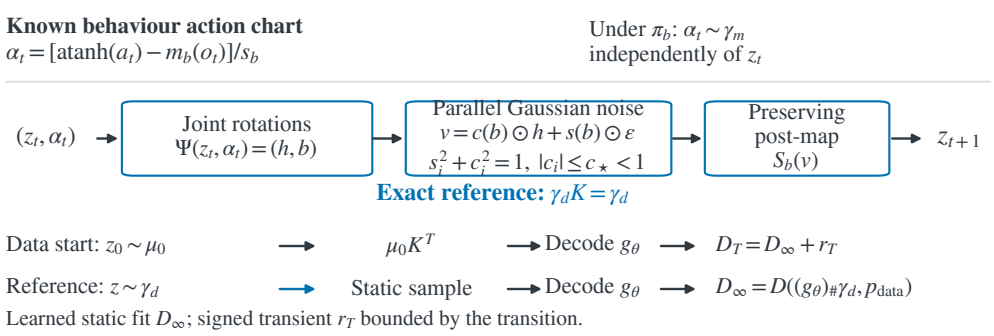


Figure 1: **A transition with an exactly preserved Gaussian reference.** Latent state z and charted action α feed joint rotations under behaviour π_b, a parallel Gaussian noise stage with $s_i^2 = 1 - c_i^2$, and a preserving post-map, so the Gaussian reference holds at every parameter value. Decoding the reference gives the static observation law p_θ. The rollout discrepancy from data equals the static discrepancy plus a bounded signed transient r_T (Section 4).

Theorem 1 (Gaussian-preserving affine couplings)**.** *The block-affine diffeomorphism* $(u, v) \mapsto (u, A(u)v + b(u))$ *preserves a standard Gaussian if and only if* $b(u) = 0$ *and* $A(u)A(u)^\top = I$. *Thus every admissible coupling is*

$$(u, v) \longmapsto (u, Q(u)v), \qquad Q(u) \in \mathrm{O}(n_v). \tag{3}$$

This characterization (Appendix A.2) gives the building blocks (Figure 1). Conditioned Givens rotations implement this identity without projection: we update disjoint coordinate pairs in parallel, change the fixed partition between layers, and with a fixed number of layers the depth no longer grows with latent width, as it would in a scan. These maps preserve each sample's norm. Complementary norm-changing maps take the form $\Phi^{-1} \circ S \circ \Phi$, where Φ applies the standard Gaussian cumulative distribution function coordinatewise and S preserves volume in the quantile cube. Their approximation theorem allows unrestricted smooth profiles and depth (Appendix A.3); finite-network capacity is a separate constraint. The stochastic architecture below uses rotations for both deterministic stages, with radial change supplied by its noise.

Joint mixing and simultaneous refresh. Mix state and action innovations using a norm-preserving Gaussian bijection Ψ:

$$(h, b) = \Psi(z, \alpha), \qquad v_i = c_i(b)h_i + \sqrt{1 - c_i(b)^2}\,\varepsilon_i, \qquad z' = S_b(v), \tag{4}$$

where $\varepsilon \sim \gamma_d$ is fresh, $|c_i(b)| \le c_\star < 1$, and S_b preserves γ_d. At the reference law, h and b are independent Gaussians. Conditional on b, the simultaneous noise stage and the post-map retain γ_d. Every recurrent coordinate receives noise; leaving deterministic memory outside this state can prevent contraction (Appendix F.1). The coefficient network reads b, not the coordinates being refreshed. In code, $c_i = c_\star \tanh \ell_i(b)$.

Given (z, a, o), including any observation context used by the action chart, both h and b are known. For a diffeomorphic post-map S_b, let $v = S_b^{-1}(z')$; the exact conditional log density is

$$\log k_\theta(z' \mid z, a, o) = \sum_i \log \mathcal{N}\big(v_i; c_i(b)h_i, 1 - c_i(b)^2\big) + \log |\det J_{S_b^{-1}}(z')|. \tag{5}$$

The inverse-Jacobian term vanishes for the implemented rotational post-map. The input mixer has no Jacobian term in this *conditional* density because it transforms known conditioning variables. The sequential block scan used in the fixed-latent transition benchmark is a separate preserving parameterization, described in Appendices I.2 and A.7.

An optional reference component. The mixture $k_r = (1-r)k_\theta + r\gamma_d$ preserves the reference and refreshes the entire state with probability r. Its additional conditional loss is at most $-\log(1 - r)$: 0.001001 nats per transition at the recurrent study's default $r = 10^{-3}$. The reference component also bounds extreme log-loss relative to the fitted marginal (Appendix B.4). We examine its density responsibility in frozen-latent checkpoints and evaluate a reset-free recurrent variant.

Training with observations and recurrent inference. Invertible state charts use the pair objective $-\log p_\theta(x' \mid x, a) - w \log p_\theta(x)$, with marginal weight w. For pixels, a convolutional decoder and recursive Gaussian posterior $q_\varphi(z_t \mid z_{t-1}, o_t, a_{t-1})$ are learned jointly with the transition and reward head. Training minimizes the negative evidence lower bound (ELBO) of the joint behaviour-generated sequence model, up to known behaviour-action log densities. It combines observation and reward reconstruction, an initial KL to γ_d, and transition KL terms using the exact prior density. Observation o_t is emitted before z_{t+1} and supplies its action-chart context; the posterior approximates the latent posterior. Appendix C.4.1 gives the factorization and objective; Appendix C.4 specifies the comparisons.

4 Marginal certificates

Write g_θ for the observation map, including an inverse chart and any decoder. The static reference image is $p_\theta = (g_\theta)_\# \gamma_d$. For a stochastic observation model the same argument applies to its observation kernel; pixel-mean metrics below use the deterministic decoder mean. This formulation accommodates noninvertible pixel decoders.

Theorem 2 (Zero marginal drift)**.** *Under the behaviour action chart—including observation-conditioned charts, by Lemma 25—the transition in equation 4, with or without the full-state reference mixture, preserves γ_d for every parameter value. If $z_0 \sim \gamma_d$, then $\mathrm{Law}(z_T) = \gamma_d$ and $\mathrm{Law}(g_\theta(z_T)) = p_\theta$ at every horizon.*

The guarantee concerns the generative prior; recurrent inference learns an approximate posterior.

Static level and signed transient. Let K be the behaviour-averaged latent kernel, μ_0 its initialized law, and $A_T = (g_\theta)_\#(\mu_0 K^T)$. For a metric D satisfying $D(P, Q) \le C_D \, \|P - Q\|_{\mathrm{TV}}$, set $D_T = D(A_T, p_{\mathrm{data}})$ and $D_\infty = D(p_\theta, p_{\mathrm{data}})$, using the total-variation (TV) norm in the $[0, 2]$ convention.

Proposition 3 (Static prediction and transient bound)**.** *For any reference-preserving transition,*

$$D_T = D_\infty + r_T, \qquad |r_T| \le C_D \left\|\mu_0 K^T - \gamma_d\right\|_{\mathrm{TV}} \le C_D \, \|\mu_0 - \gamma_d\|_{\mathrm{TV}} \, . \tag{6}$$

Consequently, total-variation convergence to the reference implies $D_T \to D_\infty$.

Reverse triangle inequality and data processing give the bound (Appendix A.5). For observations of diameter Δ, the squared energy statistic instead satisfies $|\mathcal{E}(A_T, Q) - \mathcal{E}(p_\theta, Q)| \le 2\Delta \left\|\mu_0 K^T - \gamma_d\right\|_{\mathrm{TV}}$ (Appendix B.2). Normalized root-mean-square (RMS) pixel distance has $\Delta \le 1$. Unbounded state-space distances require additional moment control.

Preservation alone permits the identity kernel; the parallel noise stage additionally supplies relaxation.

Theorem 4 (Joint contraction and finite-start convergence)**.** *Let M be the kernel from (z, α) to z' in equation 4, with Gaussian-preserving Ψ and S_b and uniform cap $|c_i(b)| \le c_\star < 1$. For any joint input law ν,*

$$D_{\mathrm{KL}}(\nu M \| \gamma_d) \le c_\star^2 D_{\mathrm{KL}}(\nu \| \gamma_{d+m}). \tag{7}$$

Under an exact behaviour chart, $\nu = \mu \otimes \gamma_m$, giving state-kernel KL contraction by $c_\star^2$. If Ψ also preserves norms pointwise and $M_2 = \mathbb{E}_{\mu_0}\|z\|^2 < \infty$, define

$$B_\star = \tfrac{1}{2}\left[c_\star^2 (M_2 + m) - d c_\star^2 - d \log(1 - c_\star^2)\right]. \tag{8}$$

Then, for $T \ge 1$,

$$\left\|\mu_0 K^T - \gamma_d\right\|_{\mathrm{TV}} \le \min\{2, \sqrt{2B_\star}\, c_\star^{T-1}\}. \tag{9}$$

Proof sketch. After Ψ, the KL chain rule separates the residual action b from the conditional state $h \mid b$. At fixed b, Gaussian entropy contraction and post-map data processing give $c_\star^2$; convexity removes b. Norm preservation bounds $\mathbb{E}\|h\|^2$ by $M_2 + m$. Convexity bounds first-step KL by $B_\star$, even from atomic initial laws. Iteration and Pinsker give equation 9 (Appendices B.1–B.2).

For a specified finite initialization bank, M_2 is exactly computable; a population certificate requires a population moment bound. The optional reference mixture gives the independent bound $2(1 - r)^T$ for *any* initial law and strengthens the entropy contraction to $(1 - r)c_\star^2$. Across the fitted

initialization banks in §5.4, equation 9 becomes informative after 3×10^3–10^4 steps. At 10^5 steps, the reference-mixture bound puts the population marginal gap below 1.5×10^{-43}, far below evaluation sampling error. The terminal comparison therefore tests the predicted static level; intermediate horizons assess the finite-start bound. These are real-arithmetic guarantees, with numerical scope detailed in Appendix B.5.

Policy deviation in physical action space. For a bijective exact chart, divergence is invariant under its action transformation. Let $\lambda_t = \mathrm{Law}(Z_t, O_t)$, with state marginal μ_t, and define $H_t = D_{\mathrm{KL}}(\mu_t \| \gamma_d)$ and $\kappa_t = \mathbb{E}_{\lambda_t} D_{\mathrm{KL}}(\pi_t(\cdot \mid z, o) \| \pi_b(\cdot \mid z, o))$. The transition reads o only through the action innovation. For finite H_0, the joint-input argument gives

$$H_{t+1} \le c_\star^2 (H_t + \kappa_t), \qquad H_T \le c_\star^{2T} H_0 + \sum_{t<T} c_\star^{2(T-t)} \kappa_t. \tag{10}$$

Pinsker gives a marginal TV bound. When both policies are nonsingular tanh-Gaussians with the same tanh transform, their conditional KL is the Gaussian KL before tanh; deterministic action selection has infinite KL against this behaviour law. The bound concerns departure from the behaviour reference, not predictive error. Fitted charts require divergence to their induced policy (Appendix B.3; numerics in Figure 3).

5 Experiments

We evaluate static marginal predictions, finite-start bounds, conditional fit, and planning. A chain *escapes* if its state or decoded observation becomes non-finite; recurrent evaluation also stops a chain when any latent coordinate exceeds $\pm10^6$. *Survival* is the fraction remaining at the stated horizon, a numerical criterion distinct from distributional fidelity. We report negative log-likelihood (NLL), mean-squared error (MSE), and marginal energy statistics; standard errors (SEs) distinguish evaluation sampling from training-seed variation (Appendix C.3).

Pixel studies use the DeepMind Control Suite (Tassa et al., 2018). The marginal statistic is $\mathcal{E}(P, Q) = 2\mathbb{E}\delta(X, Y) - \mathbb{E}\delta(X, X') - \mathbb{E}\delta(Y, Y')$, with independent draws. The distance δ is normalized root-mean-square pixel distance in the recurrent/native study; fixed-latent studies use Euclidean distances in the stated latent or pixel coordinates (Appendix C.3). Finite unbiased estimates can be negative. D_{ref} estimates the static level D_∞; pixel scores use decoder means.

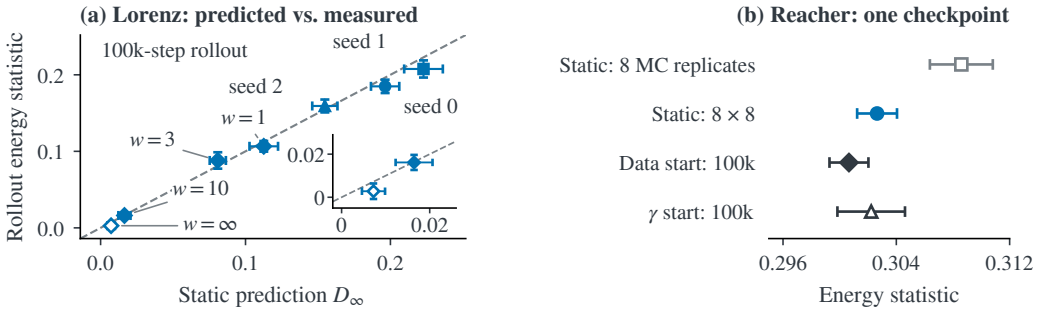
5.1 Static prediction of the long-horizon marginal

Can a model's limiting discrepancy be predicted before running it to a long horizon? Figure 2 compares static reference estimates with 10^5-step rollouts.

Model-specific levels. Seven constrained Lorenz-63 models—three training seeds and four marginal-weight fits—span a $31\times$ range of static energy statistics. Each 10^5-step measurement is within one nominal combined SE of its prediction; this SE omits the unavailable prediction–rollout covariance. Every constrained chain remains finite. An unconstrained flow improves held-out pair likelihood by 0.83–0.99 nats per dimension, yet only 14%–94% of its chains survive. Survivor distances describe the retained subpopulation (Appendix D).

Marginal fidelity and conditional fit. At fixed architecture and training budget, raising the marginal-loss weight from 1 to 10 reduces long-horizon error by $6.6\times$ while worsening pair NLL by 0.19 nats per dimension (Appendix D.1). Conditional and marginal objectives select different trade-offs. Soft multistep, Jacobian, and invariant-measure penalties (Schiff et al., 2024) are compared in Appendix D.2.

Fixed-latent pixel models. The sequential preserving transition retains every chain on twelve tasks with frozen 256-dimensional autoencoder latents. Flow matching retains every chain on ten tasks, with smaller pixel distance on five and larger on five. The sequential model preserves fixed-action kernels (data and fitting: Appendices C.1–C.2; outcomes and alternate metrics: Appendices E.1 and E.3). For the capped Reacher checkpoint in Figure 2b, resampling and independent rollout replication support sampling variability as the explanation of the static offset (Appendices G.1 and G.1.1).



Bars: ±1 SE of repeated evaluation, conditional on each fitted model.

Figure 2: **Static reference samples predict model-specific long-horizon discrepancies. (a)** Seven Lorenz models span a $31\times$ range of predicted energy statistics; w is the marginal-loss weight. The inset enlarges the two smallest values; an open marker denotes a value below evaluation resolution. **(b)** One capped Reacher checkpoint, using Euclidean distance in 256-dimensional frozen-autoencoder coordinates: the static estimate (8 Monte Carlo replicates), resampling (eight 2,048-point clouds per reference block; SE over eight blocks), and data- and prior-initialized 10^5-step rollouts. The offset is consistent with static-estimate sampling variability.

5.2 Finite-start bounds and conditional likelihood

Do the computable bounds hold for singular initial laws and known policy shifts, and what do they leave unconstrained? We first check the implemented inequalities, then examine fitted conditional densities.

Atomic starts and policy shift. A finite initialization bank has infinite KL to a Gaussian but a known second moment. Across contraction caps, reset probabilities, and tanh-Gaussian policy shifts, all 114 bounded-observable comparisons satisfy the population bounds with the simultaneous sampling allowance (Figure 3). The diagnostics in Appendix G.3 check the implementation of the proved inequalities through bounded observables; predictive accuracy is evaluated separately below. Panel (c) isolates prior computation: sampling takes 2.50–4.57 ms per batch of 32 as state width increases from 256 to 1024. These CPU timings exclude the decoder and simulator (Table 28).

Conditional fit in frozen-latent models. The frozen-latent sequential-scan checkpoints illustrate what preservation leaves unconstrained. On 479,968 held-out Reacher transitions, the cap reduces maximum conditional NLL from 1.41×10^8 to 193 nats per dimension (Appendix E.2). The capped Swimmer model's median remains 294. Adding the reference mixture without retraining reduces that median to 6.79, but the reference branch dominates 98.3% of transition densities. Most of this reduction in scored loss comes from the static reference component (Table 26, Appendix G.2).

5.3 Controlled occupancy and planning

Does preserving occupancy avoid the restrictions of preserving every fixed-action kernel? With independent Gaussian behaviour actions ($\tau = 0$), the stationary action-driven system has a fixed-action-preserving conditional-loss floor $-\frac{1}{2}\log(1-\varrho^2)$, where ϱ controls the action-to-state correlation. Across four $\tau = 0$ settings spanning a $25\times$ range of floors, its learned excess loss is 0.99–1.00 times the prediction, while the occupancy model's excess is 0.0011 nats (Appendix H.1). Control also depends on the fitted dynamics. On a controlled Duffing oscillator, occupancy-based cross-entropy-method (CEM) planning approaches the oracle at short horizons but incurs mean excess cost 0.242 at horizon 500; the Gaussian head's excess is 0.001. The per-action comparison includes differences in fit and transition size (Appendix H.2).

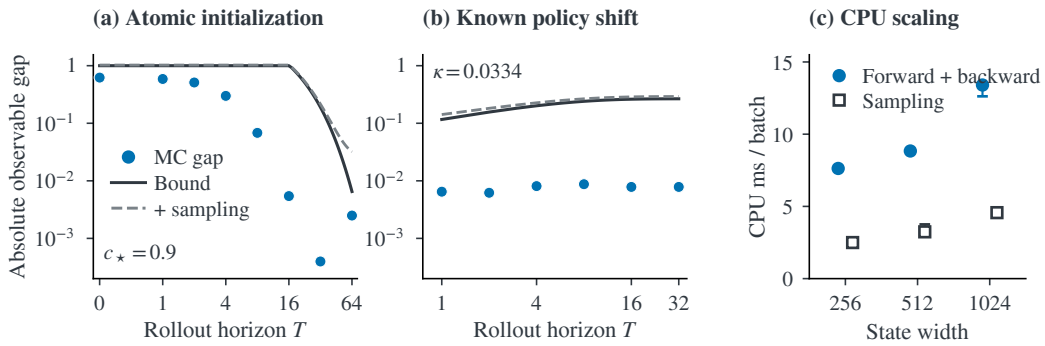


Figure 3: **Finite-start and policy-shift checks for the parallel transition.** **(a)** A 16-point initialization bank supplies the moment in equation 9. **(b)** A known policy shift has conditional action KL $\kappa = 0.0334$. Dots are Monte Carlo gaps for $f(z) = \exp(-\|z\|^2/16)$, with 8,192 independent chains per horizon. Solid curves are population bounds; dashed curves add a simultaneous 99% Hoeffding allowance. **(c)** Median forward/backward and sampling times per batch of 32, on four CPU threads, with min–max whiskers over five repetitions (prior only). These are randomized-kernel diagnostics without dynamics fitting (Appendix G).

5.4 Jointly trained recurrent pixel models

Does the separation persist when the representation, inference, and dynamics are learned jointly? We train six arms on twelve screened control tasks with three paired training seeds, yielding 216 checkpoints. All share the observation, recursive posterior, and reward networks. The occupancy transition uses joint state–action mixing; a conditioner-only ablation removes that mixing. Four additional arms test a Gaussian transition, conditional RealNVP, latent overshooting, and an invariant-measure penalty. The latter two test multistep and marginal regularization against structural preservation. Collection uses a fixed frame-conditioned tanh-Gaussian behaviour policy (Appendix C.4).

Comparison protocol. The arms use common data and paired seeds, but differ in achieved conditional fit. None of the five alternative families meets the suite-wide validation matching criteria (prediction-error, likelihood, parameter, and compute gates in Appendix C.4.2). We therefore report fitted-model contrasts under these common budgets. A same-checkpoint output-gain intervention meets these criteria, with gain 1.001 selected on validation data. Table 1 reports equal-task means; the nominal 95% percentile intervals resample three paired seeds within each fixed task, assuming independent task–seed cells. SEs follow the same convention; Appendix C.3 examines the reused seed IDs and small-sample calibration.

Long-horizon survival and marginal agreement. The occupancy arm retains all 294,912 evaluated chains across its 36 task–seed cells. Its task-mean static discrepancies span 0.0033–0.0758, with mean rollout-minus-reference gap -0.0002 ± 0.0003 (SE; Figure 4). The Gaussian, conditional RealNVP, latent-overshooting, and invariant-penalty arms have survival fractions 0.51, 0.24, 0.34, and 0.44. All four paired survival contrasts against occupancy have intervals excluding zero (Figure 5). All three hopper stand Gaussian cells survive with marginal gaps of 1.10, 1.06, and -0.16. The conditioner-only arm reaches survival 0.99993, with 21 threshold crossings in one Swimmer cell; its extreme posterior initialization is examined below.

Conditional accuracy remains distinct. The Gaussian arm has the smallest measured mean ten-step MSE among the six arms (contrast against occupancy -0.0563 $[-0.1067, -0.0024]$), despite losing almost half its chains. Conversely, ten occupancy checkpoints have observation NLL above 1, reaching 17.37 on pendulum swingup, while retaining every chain with marginal gaps within ± 0.007. The conditioner-only variant also exceeds NLL 1 in the same ten task–seed cells. Its 21 threshold crossings occur at the first step from an initialization bank with second moment 7.4×10^{10}, compared with 256 for the Gaussian reference; its reference-initialized control retains

Table 1: **Conditional accuracy and long-horizon stability across recurrent models.** Equal-task means, twelve tasks × three paired seeds; NLL, MSE, and gaps show training-seed SE. NLL is a posterior-initialized particle score under recorded actions, in nats per pixel channel. D_{ref} scores decoded Gaussian reference states, a limiting level only for preserving models under their convergence assumptions. "Escape": no gap estimated after censoring. The last row scales the fitted occupancy model's output. Bold marks best primary-arm means; the signed gap has no directional optimum.

Transition / objective	Survival	Obs. NLL ↓	10-step MSE ↓	$D_{10^5} - D_{\mathrm{ref}}$
Occupancy	**1.0000**	0.6843 ± 0.6832	0.1271 ± 0.0211	-0.0002 ± 0.0003
Conditioner only	0.9999	$\mathbf{0.5648 \pm 0.5422}$	0.1416 ± 0.0247	Escape
Gaussian	0.5149	1.7013 ± 1.1747	$\mathbf{0.0708 \pm 0.0218}$	Escape
Conditional RealNVP	0.2394	3.1391 ± 1.5809	0.1119 ± 0.0264	Escape
Latent overshooting	0.3380	7.0333 ± 1.0566	0.1692 ± 0.0212	Escape
Invariant penalty	0.4399	4.0644 ± 1.6440	0.1120 ± 0.0327	Escape
Occupancy $\times 1.001$	1.0000	0.6845 ± 0.6832	0.1273 ± 0.0211	-0.0003 ± 0.0003

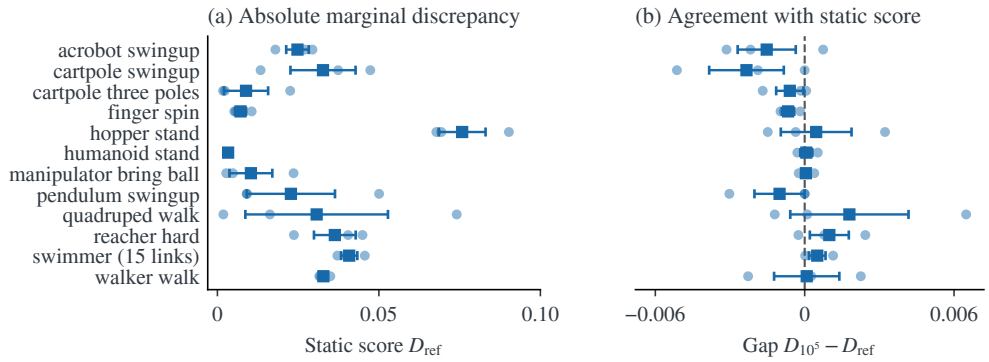


Figure 4: **Absolute marginal discrepancy and static-prediction agreement.** **(a)** Occupancy reference energy statistics against data, using normalized RMS pixel distance. **(b)** Rollout-minus-reference discrepancies against the same data. Points: three training seeds; squares: task means ± SE. The equal-task gap is -0.0002, with SE 0.0003 for independent task–seed cells and 0.0006 when grouping the reused seeds (Appendix C.3).

every chain. These measurements distinguish posterior initialization from reference-law preservation (Appendix C.4.3). The conditioner-only model also has higher ten-step MSE than occupancy (contrast $+0.0145$ $[0.003, 0.026]$).

Output gain and architectural ablations. Multiplying each prior output by 1.001 removes exact preservation but leaves survival unchanged at 10^5 steps. The changes are small: observation NLL $+0.00019$ $[0.00012, 0.00028]$, ten-step MSE $+0.00015$ $[-0.00015, 0.00041]$, and marginal gap -0.00016 $[-0.00029, -0.00004]$. An exploratory gain of 1.003 also retains every chain in all 36 cells, while the worst-cell second-moment deviation grows from 0.15 to 0.54 to 6.9 for gains 1, 1.001, and 1.003. Moment diagnostics thus distinguish these models even when finite-horizon survival is identical (Table 3). The selected gain covers all 36 primary cells; six retrained cap, reset, and width variants cover four priority tasks (12 cells each). All have survival 1.0; across these scopes, mean gaps range from -0.00052 to $+0.00047$. The default reset probability is $r = 10^{-3}$; the reset-free variant retains all chains in its 12 evaluated cells with gap -0.00025 ± 0.00035 (Figure 6, Appendices C.4.4 and C.4.7).

Planning tests a stronger requirement. Of 348 planning evaluations, 172 fail because imagination produces fewer finite candidates than required CEM elites. Primary-arm counts range from 14 to 19 failures among 36 cells; occupancy and conditioner-only each have 19. Five tasks account for 77% of all failures (Figure 8). On commonly completed cells, equal-task mean return contrasts

against occupancy range from -0.0128 to $+0.0089$; these descriptive comparisons condition on completion. The native reference completes all 12 evaluations and leads in three of the four cells where both it and occupancy complete. Completion and return measure the additional requirements of action-selected planning (Appendix C.4.5).

A native DreamerV3 reference. An offline DreamerV3 baseline (Hafner et al., 2025) retains the authors' categorical recurrent state and observation networks on four priority tasks. Exposure is matched at 25.6M frames; capacity and context differ (8.85M versus 3.26–3.29M parameters; 64 versus 16 frames per window), and wall time is not matched. With a common external Gaussian image score, it achieves lower observation NLL and ten-step MSE (0.0024 ± 0.00004 versus 0.1206 ± 0.0244). Both models retain every chain at the shared 10^3-step horizon. The native model's 10^5-step marginal was not measured, and our Gaussian certificate does not apply to its categorical state. It supplies a strong conditional-prediction reference alongside the long-horizon property studied here (Table 4, Figure 7, Appendix C.4.6).

6 Related work

World models and rollout error. Latent world models support prediction and planning (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2021; 2025). Multistep objectives and distribution-shift mitigation address errors along generated trajectories (Bengio et al., 2015; Ross et al., 2011; Talvitie, 2017; Chen et al., 2024), while control-oriented analyses distinguish likelihood from task-relevant fidelity (Lambert et al., 2020). Our constraint acts on the marginal law itself, complementing objectives that improve conditional predictions. Occupancy methods estimate stationary-distribution corrections for off-policy evaluation (Nachum et al., 2019; Zhang et al., 2020); here a conditional action chart makes occupancy preservation an identity of the learned generative model.

Invariant measures and structured dynamics. Hamiltonian, Lagrangian, symplectic, and equivariant networks encode structure in learned dynamics (Greydanus et al., 2019; Cranmer et al., 2020; Jin et al., 2020; Cohen & Welling, 2016; Garcia Satorras et al., 2021). Invariant-measure objectives fit long-run statistics (Li et al., 2022; Jiang et al., 2023; Schiff et al., 2024; Cheng et al., 2025). We instead learn coordinates with an exactly preserved latent reference and bound relaxation to it. Target-preserving transitions also appear in learned Markov chain Monte Carlo samplers (Song et al., 2017; Hoffman et al., 2019); a world model additionally fits the observed conditional dynamics. Koopman-operator projections (Koopman, 1931; Williams et al., 2015) provide finite-basis diagnostics (Appendix J.3); our certificate follows from the architecture directly.

Flows, conditional transport, and mixing. Normalizing flows provide invertible density representations (Dinh et al., 2015; 2017; Kingma & Dhariwal, 2018; Papamakarios et al., 2017; Kingma et al., 2016). We identify which block-affine couplings preserve a Gaussian reference exactly; density checks, finite-network capacity, and copula examples appear in Appendices I.1, J.1, and J.2. Quantile constructions connect this constraint to measure-preserving geometry and copulas (Arnold & Khesin, 1998; Nelsen, 2006), and the action chart uses conditional Rosenblatt transport (Rosenblatt, 1952). Gaussian entropy contraction (Gentil, 2010) applied to the joint state–action innovation yields our finite-start and policy-deviation bounds.

7 Conclusion and limitations

An explicitly preserved latent reference makes a model's long-horizon marginal available by static sampling. Joint state–action mixing fits controlled dynamics under this constraint, while parallel noise supplies finite-start bounds. The experiments distinguish marginal predictability, conditional accuracy, and planning.

Limitations. Stationarity and the action-chart conditions delimit the occupancy interpretation; approximate charts and policy shifts require error control (§4). Finite-start bounds use bank-specific moments and real arithmetic. Architecture contrasts combine preservation with achieved fit; the matched intervention isolates output scaling. The measured planning completion rates and posterior excursions motivate improved inference and finite-candidate planning (Appendix C.4.8).

AI USE STATEMENT

Generative AI assisted code development, manuscript writing, mathematical and reference checks, the authors reviewed and verified all AI-assisted work.

REPRODUCIBILITY STATEMENT

Appendix C specifies the systems, data splits, training budgets, and evaluation procedures. The theoretical supplement states assumptions and provides proofs. Detailed experimental sections report per-seed results, estimator conventions, and diagnostic controls. Experiment and analysis scripts generate the reported statistics and tables from saved outputs.

REFERENCES

Vladimir I. Arnold and Boris A. Khesin. *Topological Methods in Hydrodynamics*, volume 125 of *Applied Mathematical Sciences*. Springer, 1998.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xiaoyuan Cheng, Yi He, Yiming Yang, Xiao Xue, Sibo Cheng, Daniel Giles, Xiaohang Tang, and Yukun Hu. Learning chaos in a linear way. In *International Conference on Learning Representations (ICLR)*, 2025.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, 2016.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. In *ICLR Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. In *ICLR Workshop*, 2015.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Ivan Gentil. Logarithmic Sobolev inequality for diffusion semigroups, 2010. URL `https://arxiv.org/abs/1009.3421`.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations (ICLR)*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Matthew D. Hoffman, Pavel Sountsov, Joshua V. Dillon, Ian Langmore, Dustin Tran, and Srinivas Vasudevan. NeuTra-lizing bad geometry in Hamiltonian Monte Carlo using neural transport. *arXiv:1903.03704*, 2019.

Ruoxi Jiang, Peter Y. Lu, Elena Orlova, and Rebecca Willett. Training neural operators to preserve invariant measures of chaotic attractors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Pengzhan Jin, Zhen Zhang, Aiqing Zhu, Yifa Tang, and George Em Karniadakis. SympNets: Intrinsic structure-preserving symplectic networks for identifying hamiltonian systems. *Neural Networks*, 132:166–179, 2020.

Alexander S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer, New York, 1995. doi: 10.1007/978-1-4612-4190-4.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1×1 convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

B. O. Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931. doi: 10.1073/pnas.17.5.315.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *PMLR*, pp. 761–770, 2020.

Zongyi Li, Miguel Liu-Schiaffini, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Learning chaotic dynamics in dissipative systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 16768–16781, 2022.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Roger B. Nelsen. *An Introduction to Copulas*. Springer, 2nd edition, 2006.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Yair Schiff, Zhong Yi Wan, Jeffrey B. Parker, Stephan Hoyer, Volodymyr Kuleshov, Fei Sha, and Leonardo Zepeda-Núñez. DySLIM: Dynamics stable learning by invariant measure for chaotic systems. In *International Conference on Machine Learning (ICML)*, 2024.

Jiaming Song, Shengjia Zhao, and Stefano Ermon. A-NICE-MC: Adversarial training for MCMC. In *Advances in Neural Information Processing Systems*, 2017.

Erik Talvitie. Self-correcting models for model-based reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pp. 2597–2603, 2017.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015. doi: 10.1007/s00332-015-9258-5.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. GenDICE: Generalized offline estimation of stationary values. In *International Conference on Learning Representations (ICLR)*, 2020.

A REPRESENTATION, PRESERVATION, AND RELAXATION PROOFS

Throughout, γ_d is the standard Gaussian law on $\mathbb{R}^d$; φ and Φ are its scalar density and cumulative distribution function (CDF). A diffeomorphism is C^1 with a C^1 inverse. State, action, and observation spaces are standard Borel, so the conditional laws used below exist. Total variation uses the signed-measure $[0, 2]$ convention of Section 4. Reference preservation is a property of a specified kernel, while accuracy of its decoded reference is a separate statistical question.

A.1 TRANSPORT OF STATIONARY PAIR LAWS

Definition 5 (Reference-preserving kernel)**.** A state kernel $\tilde{K}$ preserves the reference if $\gamma_d \tilde{K} = \gamma_d$. For a controlled kernel, we distinguish this property after behaviour-policy averaging from the distinct requirement that every fixed-action kernel preserve γ_d.

Proposition 6 (Transport of the law)**.** *Let T be a bimeasurable bijection with $T_{\#}\mu = \gamma$, and let $\tilde{K}$ be the pushforward of K under T. Then $\mu K = \mu \iff \gamma \tilde{K} = \gamma$.*

Proof of Proposition 6. For measurable B,

$$(\gamma \tilde{K})(B) = \int \tilde{K}(w, B)\, \gamma(\mathrm{d}w) = \int K(z, T^{-1}B)\, \mu(\mathrm{d}z) = (\mu K)(T^{-1}B) = T_{\#}(\mu K)(B),$$

using $T_{\#}\mu = \gamma$ and the definition $\tilde{K}(Tz, \cdot) := T_{\#}K(z, \cdot)$. Since T is a bijection, $T_{\#}(\mu K) = \gamma = T_{\#}\mu$ iff $\mu K = \mu$. □

Proposition 7 (Representation of stationary pair laws)**.** *Let $\mathcal{X}$ be a standard Borel space and let $\mathcal{P}_=$ be the probability laws on $\mathcal{X} \times \mathcal{X}$ with equal atomless marginals. As E ranges over measure-space isomorphisms to the Gaussian reference, modulo null sets, and $\tilde{K}$ over Gaussian-preserving kernels, the induced laws of $(E^{-1}(z), E^{-1}(z'))$ range over precisely $\mathcal{P}_=$.*

For a smooth realization, suppose $\mathcal{X} = \mathbb{R}^d$ and the common marginal has strictly positive conditional densities on the full real coordinate line. Assume each conditional CDF $F_i(x_i \mid x_{<i})$ is C^k jointly in its arguments, $k \geq 1$, has positive derivative in x_i, and tends to 0 and 1 at the two ends of each conditional support. Then $E_i(x) = \Phi^{-1}(F_i(x_i \mid x_{<i}))$ defines a C^k diffeomorphism onto $\mathbb{R}^d$. When the pair law has a density, its representation is

$$p(x, x') = \gamma(Ex) |\det J_E(x)|\, \tilde{p}(Ex' \mid Ex) |\det J_E(x')|.$$

Proof. A Gaussian-preserving kernel gives a pair law whose two marginals are γ. Applying E^{-1} to each coordinate therefore gives equal marginals. Conversely, an atomless probability measure on a standard Borel space is isomorphic, modulo null sets, to the Gaussian probability space (Kechris, 1995, Theorem 17.41). Transporting a stationary pair through this isomorphism gives two Gaussian marginals. Disintegrating that pair yields a Gaussian-preserving kernel, and transporting back recovers the original pair law.

Under the smooth assumptions, the conditional probability integral transform (Rosenblatt, 1952) maps the common marginal to independent uniforms, and Φ^{-1} maps these to γ_d. Each coordinate is strictly increasing in x_i and onto $\mathbb{R}$ for fixed $x_{<i}$. Coordinates can therefore be inverted successively for every point of $\mathbb{R}^d$. The Jacobian is triangular with positive diagonal, and the inverse function theorem gives a C^k inverse locally at every point. Together with global bijectivity, this establishes the claimed diffeomorphism. Applying the change-of-variables formula to both coordinates gives the density expression. □

Remark 8 (Transport regularity and finite architectures). The measurable statement characterizes pair laws; the differentiable statement supplies their likelihood formula. Joint smoothness of conditional CDFs is an explicit hypothesis. Locally uniform integrable bounds on density derivatives suffice to justify differentiation of the marginal integrals; smoothness of the joint density alone need not suffice. Global bounds on the Jacobian or its inverse are additional assumptions only where stated.

These are statements about unrestricted transports and kernels. The frozen feature experiments and the recurrent pixel model use different finite representations; neither is asserted to be an exact chart of pixel space.

Corollary 9 (Exact representation with a perfect chart)**.** *For a stationary pair law with common marginal μ and a joint density, if E is a diffeomorphism with $E_{\#}\mu = \gamma$ exactly, the class of arbitrary Gaussian-preserving transition kernels represents that density, with zero marginal and conditional excess loss. Finite parameterizations retain their approximation restrictions.*

Proof of Corollary 9. Transporting the true stationary pair through an exact chart gives its true Gaussian-preserving conditional (Proposition 7). The unrestricted preserving-kernel class therefore attains zero excess marginal and conditional loss. Restricting either transport or transition to a finite parameterization can introduce approximation or optimization error. □

A.2 Rigidity and preserving diffeomorphisms

Proof of the identity equation †. Since T is a diffeomorphism and both densities are continuous, $T_{\#}\gamma = \gamma$ holds if and only if the change-of-variables identity $\gamma(T(x))\,|\det J_T(x)| = \gamma(x)$ holds at every x. Taking logarithms of both sides and cancelling the common constant $-\frac{d}{2}\log 2\pi$ leaves $-\frac{1}{2}|T(x)|^2 + \log|\det J_T(x)| = -\frac{1}{2}|x|^2$, which is exactly equation †. □

Corollary 10 (Volume and norm preservation)**.** *For a γ-preserving diffeomorphism, $|\det J_T| \equiv 1 \iff |T(x)| = |x|$ for all x.*

Proof of Corollary 10. Both directions are immediate from equation †: the left side vanishes iff the right side does. □

Proof of the identity equation ‡. The density of $T_{\#}\gamma$ at w is $\gamma(T^{-1}w)|\det J_{T^{-1}}(w)|$. Substituting $w = T(z)$ with $z \sim \gamma$,

$$\log \frac{\mathrm{d}T_{\#}\gamma}{\mathrm{d}\gamma}(T(z)) = \log\gamma(z) - \log|\det J_T(z)| - \log\gamma(T(z)) = \tfrac{1}{2}\big(|T(z)|^2 - |z|^2\big) - \log|\det J_T(z)|,$$

and $\mathrm{KL}(T_{\#}\gamma\|\gamma) = \mathbb{E}_{w\sim T_{\#}\gamma}[\log \frac{\mathrm{d}T_{\#}\gamma}{\mathrm{d}\gamma}(w)] = \mathbb{E}_{z\sim\gamma}[\cdots]$. Gibbs' inequality gives non-negativity and equality exactly when the two laws agree. Continuity then gives the pointwise identity equation †. □

Lemma 11 (Relative entropy of volume-preserving maps)**.** *If $|\det J_T| \equiv 1$ then $\mathrm{KL}(T_{\#}\gamma\|\gamma) = \frac{1}{2}(\mathbb{E}_\gamma|T(z)|^2 - d)$.*

Proof. Substitute $\log|\det J_T| = 0$ into equation ‡ and use $\mathbb{E}_\gamma|z|^2 = d$. □

For an exactly volume-preserving diffeomorphism, equality of the *population* second moment to d makes this KL zero and therefore implies Gaussian preservation. A near-zero Monte Carlo estimate does not establish that equality: cancellation and sampling can conceal nonzero pointwise residuals. Appendix I.1 reports both diagnostics.

Lemma 12 (Scalar rigidity)**.** *A γ_1-preserving diffeomorphism of $\mathbb{R}$ is $x \mapsto x$ or $x \mapsto -x$.*

Proof. T is a diffeomorphism of $\mathbb{R}$, hence strictly monotone. If increasing, $T_{\#}\gamma_1 = \gamma_1$ forces $\Phi(T(x)) = \Phi(x)$ for all x (equality of CDFs), so $T = \mathrm{id}$. If decreasing, $\Phi(T(x)) = 1 - \Phi(x) = \Phi(-x)$, so $T(x) = -x$. □

Theorem 13 (Triangular rigidity)**.** *Let T be a γ_d-preserving diffeomorphism whose Jacobian is triangular for a fixed coordinate order. Then $T_i(x) = \sigma_i x_i$ with $\sigma_i \in \{\pm 1\}$. The conclusion also applies to compositions that retain this common triangular order (Corollary 14).*

Proof. Order coordinates so that J_T is lower triangular; then T_i depends only on $x_{1:i}$, and $\partial_i T_i \neq 0$ everywhere (else $\det J_T = 0$), so $T_i(x_{1:i-1}, \cdot)$ is strictly monotone. We show $T_i(x) = \sigma_i x_i$ by induction on i.

$i = 1$: the first coordinate of $T_{\#}\gamma_d = \gamma_d$ is γ_1, and T_1 is a function of x_1 alone with $x_1 \sim \gamma_1$, so $(T_1)_{\#}\gamma_1 = \gamma_1$ and Lemma 12 gives $T_1 = \sigma_1 x_1$.

Inductive step: assume $T_j = \sigma_j x_j$ for $j < i$. The map $x_{1:i} \mapsto T_{1:i}(x_{1:i})$ pushes γ_i to γ_i (a marginal of $T_{\#}\gamma_d = \gamma_d$). Condition on $x_{1:i-1} = \xi$; then $T_{1:i-1} = (\sigma_1\xi_1, \ldots, \sigma_{i-1}\xi_{i-1})$ is determined, and under γ_i the conditional law of the i-th output coordinate given the first $i-1$ outputs is γ_1 by independence. For almost every ξ, $T_i(\xi, \cdot)$ therefore pushes γ_1 forward to γ_1 and is a monotone diffeomorphism of $\mathbb{R}$, so by Lemma 12 $T_i(\xi, x_i) = \sigma_i(\xi)x_i$ with $\sigma_i(\xi) \in \{\pm 1\}$. The derivative $\partial_i T_i$ is continuous and nonzero on a connected domain, so its sign is constant. Continuity extends the resulting identity from almost every ξ to every ξ. □

Corollary 14. *A composition of triangular diffeomorphisms in a common fixed coordinate order can preserve γ_d only by a coordinate-wise sign flip. Depth and conditioner expressiveness do not alter this conclusion.*

Proof. A composition in one fixed triangular order is again triangular, so Theorem 13 applies to the complete map. Changing the order between layers removes this hypothesis. □

Theorem 15 (Block reduction)**.** *Let $T(u, v) = (u, g(u, v))$ be a C^1 diffeomorphism with $g(u, \cdot)$ a diffeomorphism of $\mathbb{R}^{n_v}$ for each $u \in \mathbb{R}^{n_u}$. Then T is $\gamma_{n_u+n_v}$-preserving iff $g(u, \cdot)$ is γ_{n_v}-preserving for every u. In particular for $n_v = 1$, $g(u, v) = \sigma v$ with $\sigma \in \{\pm 1\}$ constant, so a coupling layer with scalar blocks is trivial.*

Proof. The density of $T_{\#}\gamma$ at (u, w) is $\gamma_{n_u}(u)\,\gamma_{n_v}(g_u^{-1}(w))\,|\det J_{g_u^{-1}}(w)|$, since the first block is carried through unchanged. Equating to $\gamma_{n_u}(u)\gamma_{n_v}(w)$ and cancelling the strictly positive factor $\gamma_{n_u}(u)$ gives, for every u, $\gamma_{n_v}(g_u^{-1}(w))|\det J_{g_u^{-1}}(w)| = \gamma_{n_v}(w)$, which is exactly $(g_u)_{\#}\gamma_{n_v} = \gamma_{n_v}$. For $n_v = 1$, Lemma 12 gives $g(u, v) = \sigma(u)v$, and continuity plus connectedness makes σ constant. □

Proof of Theorem 1. By Theorem 15, T is γ-preserving iff for every u the affine map $v \mapsto A(u)v + b(u)$ pushes γ_{n_v} to γ_{n_v}. The pushforward of γ_{n_v} under an affine map is the Gaussian with mean $b(u)$ and covariance $A(u)A(u)^\top$, and a Gaussian is determined by its mean and covariance, so this holds iff $b(u) = 0$ and $A(u)A(u)^\top = I$. Requiring it for *every* value of the conditioner makes these identities in u. Then $\log|\det J_T| = \log|\det A(u)| = 0$ since $|\det A| = 1$ for orthogonal A. Scalar-block rigidity follows from Theorem 15. □

Proof of the quantile-swirl identity. Let $T = \Phi^{-1} \circ S \circ \Phi$ with S Lebesgue-preserving on $(0, 1)^d$, and write $v' = T(v)$, $\Phi(v) = q$, $S(q) = q'$. By the chain rule, $J_T = J_{\Phi^{-1}}(q')\,J_S(q)\,J_\Phi(v)$, with J_Φ and $J_{\Phi^{-1}}$ diagonal. Hence

$$\log|\det J_T| = \sum_i \log\varphi(v_i) + \log|\det J_S| - \sum_i \log\varphi(v_i') = \tfrac{1}{2}\big(|v'|^2 - |v|^2\big),$$

using $|\det J_S| = 1$ and $\log\varphi(a) - \log\varphi(b) = (b^2 - a^2)/2$. This is equation †, so T is γ_d-preserving; equivalently, Φ conjugates the γ_d-preserving diffeomorphisms onto the Lebesgue-preserving diffeomorphisms of the cube. For the swirl itself: in polar coordinates (ϱ, α) about the learned center, the map is $(\varrho, \alpha) \mapsto (\varrho, \alpha + w(\varrho)\theta)$, whose Jacobian is unit lower-triangular, so $|\det| = 1$; conjugating by a diagonal scaling D multiplies the determinant by $|\det D| \cdot |\det D|^{-1} = 1$, turning the disc into an axis-aligned ellipse at no cost. □

A.3 Ideal approximation and augmented representation

Theorem 16 (Approximation by quantile swirls)**.** *Let $d \geq 2$ and let T be a smooth Gaussian-preserving diffeomorphism joined to the identity by a smooth Gaussian-preserving isotopy T_t, $0 \leq t \leq 1$, whose support lies in one compact set. Allow pairwise quantile swirls with arbitrary interior centers, radii, smooth compactly supported radial profiles, and smooth coefficients depending on the untouched coordinates. Finite compositions of these swirls approximate T in C^1 on compact sets. This is an approximation result for the stated ideal layer family, with no fixed bound on its depth or on the number of profiles.*

Proof. Conjugate by the coordinatewise Gaussian CDF. The isotopy becomes a smooth volume-preserving isotopy S_t of $(0,1)^d$ with common compact support. Its velocity $X_t = \dot{S}_t \circ S_t^{-1}$ is smooth, compactly supported, and divergence free, by differentiating the volume-preservation identity. We give three steps for approximating its time-one flow.

Pairwise decomposition. Every smooth divergence-free field X supported in an interior rectangle has a compactly supported antisymmetric potential A satisfying $X_i = \sum_j \partial_j A_{ij}$. An elementary induction establishes this without imposing boundary conditions on a noncompact potential. The one-dimensional base case is zero: a compactly supported divergence-free scalar field is constant and hence vanishes. For $d \geq 2$, write $x = (x_1, u)$ and define

$$Y_j(u) = \int_{\mathbb{R}} X_j(t,u)\,\mathrm{d}t \quad (j > 1).$$

Then $\sum_{j>1} \partial_j Y_j = 0$. Choose a smooth bump χ in the first-coordinate interval with integral one, and set $Z_1 = X_1$, $Z_j = X_j - \chi(x_1)Y_j(u)$ for $j > 1$. The field Z is divergence free and its components Z_j, $j > 1$, integrate to zero along x_1. Consequently

$$A_{j1}(x) = \int_{-\infty}^{x_1} Z_j(t,u)\,\mathrm{d}t, \qquad A_{1j} = -A_{j1}$$

have compact support in the rectangle, with $\partial_1 A_{j1} = Z_j$ and $\sum_{j>1} \partial_j A_{1j} = Z_1$. By induction, the field Y has an antisymmetric potential B in the remaining coordinates. Adding $A_{jk} = \chi(x_1)B_{jk}(u)$ for $j, k > 1$ represents the subtracted field and proves the claim. Thus X is a finite sum of pairwise Hamiltonian fields $X^{(ij)} = \partial_j A_{ij}\, e_i - \partial_i A_{ij}\, e_j$, with untouched coordinates acting as parameters.

Approximation by radial Hamiltonians. Fix a pair $v = (x_i, x_j)$ and write its Hamiltonian as $H(v,u)$, with u the untouched coordinates. Convolve only in v with a smooth radial mollifier φ_ϵ, choosing ϵ small enough to keep its support inside the pair's square. The convolution converges to H in the joint $C^2(v,u)$ norm: all derivatives of total order at most two commute with convolution and converge uniformly on the common compact support. For fixed ϵ, a Riemann sum gives

$$H_\eta(v,u) = \sum_{m=1}^{M} |Q_m|\, H(v_m,u)\, \varphi_\epsilon(v - v_m).$$

The integrand and its derivatives of total order at most two are uniformly continuous, so this finite sum converges to the convolution in C^2 as the mesh tends to zero. Each term is a radial bump in the updated pair, with a smooth compactly supported coefficient in u. Its Hamiltonian flow preserves radius and rotates the pair by a radius-dependent angle; u remains fixed. It is therefore an admissible pairwise swirl. Smooth radial bumps give a smooth angular profile at the center, since their radial derivative divided by radius extends smoothly. The associated finite sum of swirl vector fields approximates the pairwise field in C^1.

Approximation of the flow. First freeze X_t on sufficiently short time intervals. Smoothness and common compact support give convergence of the resulting piecewise constant flows in C^1, by the flow and variational equations and Grönwall's inequality. On each interval, approximate its frozen field by the finite sum above. Continuous dependence of those same equations gives C^1 convergence of the flows as the fields converge in C^1. Finally, Lie–Trotter splitting approximates the flow of each finite sum by repeated compositions of the individual swirl flows: for these smooth compactly supported fields the one-step C^1 error is $O(h^2)$ and the accumulated error is $O(h)$ over a fixed interval. Each approximation is a finite composition. Choosing the three errors successively small yields the required C^1 approximation of S_1. Conjugating back preserves C^1 convergence on the corresponding compact sets, proving the claim. □

Theorem 16 concerns the ideal family with a compactly supported isotopy, arbitrary smooth profiles, and unrestricted composition length. Density of a fixed profile set or neural conditioner family, parameter counts, and approximation rates fall outside its scope. Each implemented layer's Gaussian preservation follows directly from the change-of-variables identity.

Theorem 17 (Measurable augmented representation)**.** *Let $\tilde{K}$ be a Gaussian-preserving Markov kernel on $\mathbb{R}^d$ with positive conditional densities. There is a Gaussian-preserving Borel isomorphism*

Ψ between full-measure subsets of $\mathbb{R}^{2d}$ such that, for γ_d-almost every z, $\Pi_{1:d}\Psi(z,\varepsilon) \sim \tilde{K}(z,\cdot)$ when $\varepsilon \sim \gamma_d$. Thus d auxiliary Gaussian coordinates suffice for this representation modulo null sets.

Theorem 18 (Smooth augmented representation)**.** *In addition, suppose all scalar conditional CDFs in the forward Knothe–Rosenblatt recursion for $\tilde{K}(z,\cdot)$ and in the reverse recursion for $\mathrm{Law}(z \mid y)$ are C^1 jointly in the scalar argument and all conditioning variables. Assume each has positive derivative in the scalar argument and limits zero and one at $-\infty$ and $+\infty$, for every conditioning value. Then the construction equation 11 defines a C^1 diffeomorphism of $\mathbb{R}^{2d}$ preserving γ_{2d}.*

Proof of Theorem 17. Conditional probability integral transforms and their triangular inverses provide measurable transports R_z with $(R_z)_{\#}\gamma_d = \tilde{K}(z,\cdot)$, invertible modulo conditional null sets. Positive densities ensure the scalar conditional laws used by the recursion are atomless with strictly increasing CDFs on their supports. These constructions can be chosen jointly measurable, using generalized inverses of the measurable conditional CDFs.

Set $y = R_z(\varepsilon)$. Under independent Gaussian z, ε, the pair (z, y) has joint law $\gamma(\mathrm{d}z)\tilde{K}(z, \mathrm{d}y)$ and both marginals are Gaussian. Its reverse conditional also has a density and is atomless for almost every y. Let S_y be its conditional triangular transport to γ_d. Then $\eta = S_y(z)$ is Gaussian independently of y. Define

$$\Psi(z,\varepsilon) = \big(R_z(\varepsilon), S_{R_z(\varepsilon)}(z)\big). \tag{11}$$

Its output has law γ_{2d}. The inverse recovers $z = S_y^{-1}(\eta)$ and then $\varepsilon = R_z^{-1}(y)$, with all identities understood on the full-measure sets where the conditional transforms are invertible. The first output retains the desired conditional law, because the second transport changes only the auxiliary output coordinate. □

Proof of Theorem 18. For each scalar equation $F(y_j \mid z, y_{<j}) = \Phi(\varepsilon_j)$, the assumed joint C^1 regularity and positive derivative in y_j give a jointly C^1 solution by the implicit function theorem. The tail limits make it unique and defined for every finite input. Applying this recursion successively gives a jointly C^1 map $(z,\varepsilon) \mapsto R_z(\varepsilon)$ with a jointly C^1 inverse in its second variable. The same argument applies to $(y, z) \mapsto S_y(z)$ and its inverse.

The construction is the composition of two global triangular changes of variables, $(z,\varepsilon) \mapsto (z, R_z(\varepsilon))$ and $(z, y) \mapsto (y, S_y(z))$. Each is a C^1 diffeomorphism, so their composition is one as well. The measurable argument establishes Gaussian preservation. No global bound on the Jacobian or its inverse is required. □

Remark 19 (Regularity and representation scope). The conditional-CDF hypotheses ensure joint smoothness of the forward and reverse transports. Locally uniform integrable envelopes for the density derivatives provide sufficient conditions, as in Remark 8. The measurable theorem describes attainable laws, and the smooth theorem supplies differentiable representations. Both concern unrestricted transports; capacity and optimization of finite architectures are separate questions.

Corollary 20 (Controlled conditional representation)**.** *Let $u = (z,\alpha) \in \mathbb{R}^{d+m}$ and let $K(u, \mathrm{d}y)$ have positive conditional densities and satisfy $\int K(u,\cdot)\gamma_{d+m}(\mathrm{d}u) = \gamma_d$. There is a Gaussian-preserving Borel isomorphism of $\mathbb{R}^{2d+m}$, modulo null sets, whose first d outputs, on input (z,α,ε) with fresh $\varepsilon \sim \gamma_d$, have conditional law $K(z,\alpha,\cdot)$ for almost every (z,α). Under the corresponding forward and reverse conditional-CDF hypotheses of Theorem 18, the construction is a C^1 diffeomorphism.*

Proof. The forward conditional transport gives $y = R_u(\varepsilon)$. The pair (u, y) has a positive joint density and Gaussian y marginal. Gaussianize its reverse conditional by $\eta = S_y(u) \in \mathbb{R}^{d+m}$. Then (y,η) is product Gaussian and the inverse recovers $u = S_y^{-1}(\eta)$, followed by $\varepsilon = R_u^{-1}(y)$. This is the proof of Theorem 17 with unequal input and output state dimensions. The same scalar implicit-function argument proves the smooth statement. In particular it retains the entire controlled conditional, not only its action-averaged state kernel. □

A.4 THE CONTROLLED CONSERVATION LAW

Stationarity constrains the behaviour-averaged kernel. Preservation at each fixed action is a stronger requirement under independent actions, and need not hold for the true controlled dynamics.

Proposition 21 (Stationarity does not imply per-action invariance)**.** *There is a controlled Markov chain whose data are exactly stationary under its behaviour policy and for which none of its fixed-action transition kernels K_a preserves the state marginal.*

Proof. Let $z \in \{0,1\}$, let $a \in \{0,1\}$ be drawn uniformly and independently of the state, and let the environment be $z_{t+1} = a_t$. If $z_0 \sim \mu := \mathrm{Unif}\{0,1\}$ then $z_t \sim \mu$ for every t, so the data are stationary and equation 1 holds: $\mu\bar{K}^\star = \mu$ with $\bar{K}^\star = \frac{1}{2}(K_0 + K_1)$. But K_0 maps every state to 0 and K_1 maps every state to 1, so $\mu K_0 = \delta_0 \neq \mu$ and $\mu K_1 = \delta_1 \neq \mu$. □

Fact 1 (Controlled stationarity is occupancy invariance)**.** If the data are stationary under a fixed behaviour policy $\pi_b(\mathrm{d}a \mid z)$, then

$$\iint \mu(\mathrm{d}z)\, \pi_b(\mathrm{d}a \mid z)\, K^\star(z, a, \mathrm{d}z') \;=\; \mu(\mathrm{d}z'). \tag{12}$$

The identity is the one-step marginal law under the stationary state–action occupancy.

Definition 22 (Action chart)**.** An *action chart* is a family of measurable maps $G(\cdot \mid z) : \mathcal{A} \to \mathbb{R}^m$ with

$$G(\cdot \mid z)_{\#}\pi_b(\cdot \mid z) \;=\; \gamma_m \qquad \text{for every } z, \tag{13}$$

The family is assumed jointly measurable in (z, a), and we write $\alpha := G(a \mid z)$. A transition is chart-driven if it reads the action only through α. For a differentiable action likelihood, assume the conditional-CDF regularity of Proposition 7, so each $G(\cdot \mid z)$ is a diffeomorphism. A positive joint density alone does not supply the required joint regularity of the conditional transport.

Remark 23 (Randomized charts for atomic actions). A deterministic measurable map sends an atom to an atom, so it cannot map a discrete action law to an atomless Gaussian. For ordered discrete actions, augment the action with an independent $u \sim \mathrm{Unif}(0,1)$ and set

$$\alpha = \Phi^{-1}\big(F(a^- \mid z) + u\, \pi_b(a \mid z)\big)\,.$$

This gives a scalar Gaussian chart exactly under the behaviour law. It is a map of (a, u), extending Definition 22's deterministic-action form. The invariance induction still applies because it uses only the conditional output law. Discrete-action likelihoods require probability masses and integration over the auxiliary variable, rather than the continuous chart's pointwise Jacobian formula. The experiments do not evaluate this extension.

Lemma 24 (Conditional Gaussianization)**.** *Under equation 13, if $z \sim \mu$ and $a \sim \pi_b(\cdot \mid z)$, then $\alpha \sim \gamma_m$ independently of z. In particular, if $\mu = \gamma_d$, then $(z, \alpha) \sim \gamma_d \otimes \gamma_m$. More generally, an exact state transport $E_{\#}\mu = \gamma_d$ gives $(E(z), \alpha) \sim \gamma_d \otimes \gamma_m$.*

Proof. The conditional law of α given z is the same Gaussian at every state, which establishes both its marginal and its independence from z. Applying a deterministic state transport preserves that independence. □

Lemma 25 (Observation-conditioned action charts)**.** *Let $\lambda(\mathrm{d}z, \mathrm{d}o) = \mu(\mathrm{d}z)L(\mathrm{d}o \mid z)$ be any joint state–context law. Suppose the jointly measurable chart $G_{z,o}$ obeys $(G_{z,o})_{\#}\pi_b(\cdot \mid z, o) = \gamma_m$ for every relevant (z, o). Under this behaviour policy, $\alpha = G_{Z,O}(A)$ is independent of (Z, O) and has law γ_m. Consequently, if a transition reads o only through α, its behaviour-averaged state kernel is*

$$K(z, B) = \int M(z, \alpha, B)\, \gamma_m(\mathrm{d}\alpha),$$

independently of L. The context may be a stochastic decoder output; it need not itself have a Gaussian distribution.

If each $G_{z,o}$ is a bimeasurable bijection modulo common null sets, a deployed policy π_t instead gives a joint input law $\nu_t = \mathrm{Law}(Z, \alpha)$ satisfying

$$D_{\mathrm{KL}}(\nu_t \| \gamma_d \otimes \gamma_m) \le D_{\mathrm{KL}}(\mu \| \gamma_d) + \kappa_t, \qquad \kappa_t := \int D_{\mathrm{KL}}(\pi_t(\cdot \mid z, o) \| \pi_b(\cdot \mid z, o))\, \lambda(\mathrm{d}z, \mathrm{d}o). \tag{14}$$

The divergence may be infinite. Thus Theorem 49 implies $H_{t+1} \le c_\star^2(H_t + \kappa_t)$ for this observation-conditioned setting, with $H_t = D_{\mathrm{KL}}(\mu_t \| \gamma_d)$.

Proof. For measurable $C \subseteq \mathcal{Z} \times \mathcal{O}$ and $B \subseteq \mathbb{R}^m$, the behaviour joint probability is

$$\Pr((Z, O) \in C, \alpha \in B) = \int_C (G_{z,o})_\# \pi_b(B \mid z, o)\, \lambda(\mathrm{d}z, \mathrm{d}o) = \lambda(C)\gamma_m(B).$$

This proves independence and the averaged-kernel statement. For the policy deviation, let $q_{z,o} = (G_{z,o})_\# \pi_t(\cdot \mid z, o)$ and $q_z = \int q_{z,o} L(\mathrm{d}o \mid z)$. Divergence invariance under the chart, convexity, and the chain rule give

$$\int D_{\mathrm{KL}}(q_z \| \gamma_m)\, \mu(\mathrm{d}z) \le \int D_{\mathrm{KL}}(q_{z,o} \| \gamma_m)\, \lambda(\mathrm{d}z, \mathrm{d}o) = \kappa_t,$$

followed by equation 14. Averaging out the context may make this inequality strict; equality holds when the context is a deterministic function of the state. Applying the joint-input entropy theorem proves the last claim. □

For the known tanh-Gaussian policy, $G_{z,o}(a) = (\operatorname{atanh} a - m_b(o)) \oslash s_b(o)$, with $s_b(o) > 0$, is exact on the open action cube. When the deployed policy is also a nonsingular tanh-Gaussian with the same tanh transformation, their conditional KL equals the Gaussian KL before tanh. Additional direct observation-dependent transition parameters require a separate preservation argument: the lemma covers context entering through the chart alone. Conditional likelihoods then condition on (z, a, o), unless o is determined by z.

Theorem 26 (Zero marginal drift, controlled)**.** *Let $\Psi : \mathbb{R}^{d+m+k} \to \mathbb{R}^{d+m+k}$ satisfy $\Psi_\# \gamma_{d+m+k} = \gamma_{d+m+k}$, let $\varepsilon_t \sim \gamma_k$ be fresh at each step, and set*

$$z_{t+1} \;=\; \Pi_{1:d}\, \Psi(z_t, \alpha_t, \varepsilon_t), \qquad \alpha_t = G(a_t \mid z_t). \tag{15}$$

If $z_0 \sim \gamma_d$ and the deployed policy's chart image is γ_m at every state — in particular under the behaviour policy itself, by Lemma 24, and under observation-conditioned policies by Lemma 25 — then $\mathrm{Law}(z_T) = \gamma_d$ for every $T \ge 0$, at every parameter value of Ψ, including at initialization.

Proof. Suppose $z_t \sim \gamma_d$. The deployed policy's chart image is γ_m at every state, so $\alpha_t \sim \gamma_m$ independently of z_t (Lemma 24), and $\varepsilon_t \sim \gamma_k$ is fresh, so $(z_t, \alpha_t, \varepsilon_t) \sim \gamma_{d+m+k}$. Hence $\Psi(z_t, \alpha_t, \varepsilon_t) \sim \gamma_{d+m+k}$, whose first d coordinates are γ_d. Nothing beyond measure preservation of Ψ was used. □

Proposition 27 (Strict inclusion under a state-independent chart)**.** *Suppose the behaviour actions are independent of the Gaussian reference state and have a state-independent invertible chart $\alpha = G(a)$. Within the measurable augmented representation, positive-density kernels preserving γ_d for each action form a subclass of kernels preserving the behaviour-averaged Gaussian law. The inclusion is strict, even for smooth positive-density transitions realized by orthogonal augmented maps.*

Proof. Every per-action-preserving kernel preserves the state law after averaging over independent actions. For positive-density kernels, the construction of Theorem 17 can be applied conditionally on α, with measurable dependence on this parameter, and the augmented map leaves the α block fixed. Thus this subclass has a joint preserving representation.

For strictness, take $d = m = k = 1$, $\alpha = a \sim \gamma_1$, and $z' = \varrho a + \sqrt{1 - \varrho^2}\, \xi$ with $0 < |\varrho| < 1$. The map

$$(z, \alpha, \xi) \longmapsto \big(\varrho\alpha + \sqrt{1 - \varrho^2}\, \xi,\; z, -\sqrt{1 - \varrho^2}\, \alpha + \varrho\xi\big)$$

is orthogonal and has this first coordinate. Its state output is Gaussian after averaging over actions, while its law at fixed a is $\mathcal{N}(\varrho a, 1 - \varrho^2) \ne \gamma_1$. □

Remark 28 (Physical actions and state-dependent charts). If G depends on z, preservation at fixed α need not imply preservation at fixed physical action a, because $G(a \mid z)$ then varies across input states. Conversely, per-action preservation alone need not preserve occupancy under a state-dependent policy, as Remark 32 shows. The two classes therefore need not be nested for such a policy. The occupancy construction guarantees the behaviour reference under its conditional-chart hypothesis; it does not assert that every physically per-action-preserving kernel satisfies that occupancy constraint.

Proposition 29 (Per-action likelihood floor for independent actions)**.** *Let $z_{t+1} = \varrho a_t + \sqrt{1-\varrho^2}\,\xi_t$ with $|\varrho| < 1$, independent standard Gaussian actions and innovations, and $z_0 \sim \gamma_1$. Among kernels satisfying $\gamma_1 \hat{K}_a = \gamma_1$ for every a, the minimum expected excess conditional negative log-likelihood (NLL) is $-\frac{1}{2}\log(1-\varrho^2)$ nats. The occupancy-preserving augmented class contains the truth and has zero excess loss.*

Proof. Write $p_a(y)$ for the true Gaussian conditional and $q_a(y \mid z)$ for a candidate density. Under the truth, z and a are independent and p_a does not depend on z. Per-action preservation gives $\int q_a(y \mid z)\gamma(\mathrm{d}z) = \varphi(y)$ for almost every y. Convexity of $-\log$ therefore yields

$$\mathbb{E}_z[-\log q_a(y \mid z)] \geq -\log \varphi(y).$$

Integrating over a and $y \sim p_a$ gives the lower bound $H(\gamma_1) = \frac{1}{2}\log(2\pi e)$ on expected NLL, attained by the independent-refresh kernel $q_a(y \mid z) = \varphi(y)$. Subtracting the truth's entropy $\frac{1}{2}\log(2\pi e(1-\varrho^2))$ proves the floor. The orthogonal map in Proposition 27 realizes the truth in the augmented class. □

Theorem 30 (Policy deviation in conditional total variation)**.** *Let $M(z, \alpha, \cdot)$ be a state kernel whose Gaussian-action average K preserves γ_d, and write $q_t(\cdot \mid z) = \mathrm{Law}(\alpha_t \mid Z_t = z)$ under deployment. Define*

$$\eta_t = \int \|q_t(\cdot \mid z) - \gamma_m\|_{\mathrm{TV}}\ \mu_t(\mathrm{d}z), \qquad d_t = \|\mu_t - \gamma_d\|_{\mathrm{TV}}\,. \tag{16}$$

If $\rho = \frac{1}{2}\sup_{z,w}\|K(z,\cdot) - K(w,\cdot)\|_{\mathrm{TV}}$ is its Dobrushin coefficient, then

$$d_{t+1} \leq \rho d_t + \eta_t, \qquad d_T \leq \rho^T d_0 + \sum_{t<T} \rho^{T-1-t}\eta_t. \tag{17}$$

In particular $d_T \leq d_0 + \sum_{t<T}\eta_t$, while $\rho < 1$ and $\eta_t \leq \bar{\eta}$ imply $\limsup_T d_T \leq \bar{\eta}/(1-\rho)$. For the observation-conditioned chart of Lemma 25, $\eta_t \leq \min\{2, \sqrt{2\kappa_t}\}$.

Proof. Data processing through $M(z,\cdot,\cdot)$ at each fixed z bounds $\|\mu_{t+1} - \mu_t K\|_{\mathrm{TV}}$ by η_t. Reference invariance and the Dobrushin contraction give $\|\mu_t K - \gamma_d\|_{\mathrm{TV}} \leq \rho d_t$. The triangle inequality and induction yield equation 17. Pinsker, Jensen, and the conditional KL bound in the observation lemma give $\eta_t \leq \sqrt{2\kappa_t}$. □

Remark 31 (Invariance and uniform contraction). A preserving deterministic bijection has Dobrushin coefficient one. So does a nontrivial scalar Ornstein–Uhlenbeck (OU) kernel on the unbounded line: for $0 < |c| < 1$, its transition laws have distance $4\Phi(|c|\,|z-w|/(2\sqrt{1-c^2})) - 2 \to 2$ as $|z-w| \to \infty$. Thus a noise floor does not establish $\rho < 1$.

Remark 32 (Matching action marginals is insufficient). Let the state and behaviour action both be independent uniform bits, with K_0 the identity and K_1 the flip. Both kernels preserve uniform states. Deploying $a = z$ from that reference still gives uniform actions, but sends both states to zero. The action marginal is unchanged while the state's TV deviation is one. Its average conditional action deviation is also one.

Remark 33 (Fitted-chart error). Suppose $\|(G_{z,o})_{\#}\pi_b(\cdot \mid z, o) - \gamma_m\|_{\mathrm{TV}} \leq \varepsilon_G$ uniformly over (z, o). This gives $d_T \leq \min\{2, T\varepsilon_G\}$ from the reference by Theorem 30. For a fitted invertible chart, KL comparisons instead use the chart-induced policy $(G_{z,o}^{-1})_{\#}\gamma_m$. A TV residual or a sampled correlation alone does not bound that KL.

A.5 Static marginals and the signed transient

Proof of Theorem 2. Lemma 25 includes the generated observation context and gives $(z_t, \alpha_t) \sim \gamma_{d+m}$. The preserving joint map therefore yields independent Gaussian blocks (h, b). Conditional on b, every refreshed coordinate has variance $c_i(b)^2 + 1 - c_i(b)^2 = 1$, with independent coordinates. The preserving post-map S_b retains γ_d for each b; integrating over b and inducting over time proves the state claim. The Gaussian reference component has the same output law, so its fixed-probability mixture also preserves it. Applying the fixed observation map or observation kernel gives the decoded statement. The uncontrolled case sets $m = 0$. □

Proof of Proposition 3. Reference preservation gives $\mu_0 K^T - \gamma_d = (\mu_0 - \gamma_d)K^T$. Every Markov kernel contracts total variation, as does the observation map or a fixed observation kernel. The reverse triangle inequality therefore implies

$$\begin{aligned} |D_T - D_\infty| &\le D(A_T, p_\theta) \le C_D \, \|A_T - p_\theta\|_{\mathrm{TV}} \\ &\le C_D \, \left\|\mu_0 K^T - \gamma_d\right\|_{\mathrm{TV}} \le C_D \, \|\mu_0 - \gamma_d\|_{\mathrm{TV}} \, . \end{aligned}$$

Define $r_T = D_T - D_\infty$. This proves the identity and both bounds for any initial law μ_0. Total-variation convergence makes the middle bound tend to zero; reference invariance alone does not imply convergence. □

Remark 34 (Energy-distance convergence). For Euclidean observations with finite first moments, write $\mathcal{E}(P, Q) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|$, with independent draws. Its square root is a metric; the squared statistic need not satisfy the metric triangle inequality, and its unbiased sample estimate can be negative. On a space of diameter Δ, the direct signed-measure argument in equation 31 gives $|\mathcal{E}(P, R) - \mathcal{E}(Q, R)| \le 2\Delta \, \|P - Q\|_{\mathrm{TV}}$. Thus the transient bound holds with $C = 2\Delta$ for this statistic. Unbounded observation distances require additional moment control.

Corollary 35 (Marginal weighting at the optimum)**.** *Let $M(\theta)$ and $C(\theta)$ be fixed marginal and conditional cross-entropy objectives, and suppose θ_w is a global minimizer of $J_w = wM + C$. For $w' > w \ge 1$, $M(\theta_{w'}) \le M(\theta_w)$, $C(\theta_{w'}) \ge C(\theta_w)$, and $J_1(\theta_{w'}) \ge J_1(\theta_w)$. Adding N_{unpaired} single-state observations to N_{paired} pairs motivates the population composite-likelihood weight $w = 1 + N_{\mathrm{unpaired}}/N_{\mathrm{paired}}$.*

Proof. Optimality at both weights gives

$$w'\Delta M + \Delta C \le 0, \qquad w\Delta M + \Delta C \ge 0,$$

where $\Delta M = M(\theta_{w'}) - M(\theta_w)$ and similarly for C. Subtracting gives $(w' - w)\Delta M \le 0$, hence $\Delta M \le 0$. The second inequality implies $\Delta C \ge -w\Delta M \ge 0$ and $\Delta J_1 \ge -(w - 1)\Delta M \ge 0$. For the weight, divide the population composite negative log-likelihood $N_{\mathrm{paired}}(M + C) + N_{\mathrm{unpaired}}M$ by the number of pairs. □

The ordering applies to population optima of these cross-entropy objectives; held-out energy distances and finite optimization results have no such ordering.

A.6 One-step accuracy and path-law divergence

The next identities concern state–action Markov path laws with a shared behaviour rule. Their state variable contains the information needed for the Markov property; a pixel observation alone need not suffice.

Proposition 36 (AR(1) sensitivity near the stability boundary)**.** *Let the truth be $z' = \lambda z + \sqrt{1 - \lambda^2}\,\xi$ with $|\lambda| < 1$, $\xi \sim \gamma_1$, and let a free model $z' = mz + s\xi$ have $s^2 = 1 - \lambda^2$ and $m = \lambda + \delta$. Assume fresh independent innovations and a centered initial law of finite variance. Then the stationary variance is $v_\infty = (1 - \lambda^2)/(1 - m^2)$ for $|m| < 1$, so $v_\infty - 1 = 2\lambda\delta/(1 - \lambda^2) + O(\delta^2)$, while $v_T \to \infty$ for $|m| \ge 1$: the transition is at $\delta = 1 - \lambda$ on the positive-memory boundary (the other boundary is $\delta = -1 - \lambda$). The reverse one-step KL at marginal variance v_t is $\delta^2 v_t/(2(1 - \lambda^2))$.*

Proof. The model's marginal variance obeys $v_{t+1} = m^2 v_t + s^2$ with $s^2 = 1 - \lambda^2$. If $|m| < 1$ the unique fixed point is $v_\infty = s^2/(1 - m^2)$; if $|m| \ge 1$ then $v_t \to \infty$ since $v_{t+1} - v_t = (m^2 - 1)v_t + s^2 \ge s^2 > 0$. Substituting $m = \lambda + \delta$,

$$v_\infty - 1 = \frac{(1 - \lambda^2) - (1 - m^2)}{1 - m^2} = \frac{2\lambda\delta + \delta^2}{1 - \lambda^2 - 2\lambda\delta - \delta^2} = \frac{2\lambda\delta}{1 - \lambda^2} + O(\delta^2),$$

the expansion being local to $\delta = 0$, with $|2\lambda\delta + \delta^2| < 1 - \lambda^2$. The denominator vanishes at $2\lambda\delta + \delta^2 = 1 - \lambda^2$, i.e. at $\delta^\star = -\lambda + \sqrt{\lambda^2 + 1 - \lambda^2} = 1 - \lambda = (1 - \lambda^2)/(1 + \lambda)$. For the KL, both kernels are Gaussian with the same variance $s^2 = 1 - \lambda^2$ and means λz and mz, so $\mathrm{KL}(\mathcal{N}(mz, s^2)\|\mathcal{N}(\lambda z, s^2)) = (mz - \lambda z)^2/(2s^2) = \delta^2 z^2/(2(1 - \lambda^2))$; taking expectation under the centered model marginal, for which $\mathbb{E}z^2 = v_t$, gives the claim. Finally, under $m = \cos\phi$, $s = \sin\phi$ we have $1 - m^2 = s^2$ identically, so $v_\infty = s^2/(1 - m^2) = 1$ for every ϕ with $\sin\phi \ne 0$, and $v_{t+1} = \cos^2\phi \, v_t + \sin^2\phi$ has $v_t \equiv 1$ whenever $v_0 = 1$. □

Lemma 37 (Horizon decomposition)**.** *Let $P^{\theta}_{1:T}$ and $P^{\star}_{1:T}$ be time-homogeneous Markov state–action path laws with the same conditional policy $\pi_b(\mathrm{d}a \mid z)$, containing T states and $T-1$ actions. Let the true initial law be μ. Then*

$$D_{\mathrm{KL}}(P^{\theta}_{1:T} \| P^{\star}_{1:T}) = D_{\mathrm{KL}}(\mu^{\theta}_1 \| \mu) + \sum_{t=1}^{T-1} \mathbb{E}_{z\sim\mu^{\theta}_t} \mathbb{E}_{a\sim\pi_b(\cdot|z)} \ell_\theta(z,a), \tag{18}$$

where $\ell_\theta(z,a) = D_{\mathrm{KL}}(K_\theta(z,a,\cdot) \| K^{\star}(z,a,\cdot))$. Relative entropies may take the value $+\infty$. When the terms are finite, time dependence of the incremental reverse KL enters through μ^{θ}_t.

Proof. The joint path law factors as the initial state law times $\prod_{t=1}^{T-1} \pi_b(\mathrm{d}a_t \mid z_t) K(z_t, a_t, \mathrm{d}z_{t+1})$. The chain rule for relative entropy gives the initial KL plus the expected sum of conditional transition KLs; the common action factors contribute zero. The policy factors agree at each history, so each expectation is $\mathbb{E}_{z\sim\mu^{\theta}_t} \mathbb{E}_{a\sim\pi_b(\cdot|z)} \ell_\theta(z,a)$. If actions are marginalized out of the path law, data processing instead bounds that state-only KL above by the joint-path expression; equality need not hold. □

Theorem 38 (Constant reverse relative-entropy rate)**.** *Under Lemma 37, suppose $\mu^{\theta}_t = \mu$ for all t and $\ell^{\mathrm{rev}}_0 := \mathbb{E}_{z\sim\mu} \mathbb{E}_{a\sim\pi_b(\cdot|z)} \ell_\theta(z,a) < \infty$. Then $D_{\mathrm{KL}}(P^{\theta}_{1:T} \| P^{\star}_{1:T}) = (T-1)\ell^{\mathrm{rev}}_0$. The rate may be zero; otherwise the divergence grows linearly with horizon.*

Remark 39 (Forward sequence likelihood)*.* For stationary true data and the same conditional behaviour policy,

$$D_{\mathrm{KL}}(P^{\star}_{1:T} \| P^{\theta}_{1:T}) = D_{\mathrm{KL}}(\mu \| \mu^{\theta}_1) + (T-1)\ell^{\mathrm{fwd}}_0,$$
$$\ell^{\mathrm{fwd}}_0 := \mathbb{E}_{z\sim\mu} \mathbb{E}_{a\sim\pi_b(\cdot|z)} D_{\mathrm{KL}}(K^{\star}(z,a,\cdot) \| K_\theta(z,a,\cdot)).$$

When these terms are finite, forward sequence KL is affine in T without an invariance assumption on the model. Teacher-forced negative log-likelihood estimates the corresponding cross-entropy, differing from KL by the model-independent data entropy when that entropy is well defined.

Proof of Theorem 38 and Remark 39. For the reverse direction, every summand in Lemma 37 equals ℓ^{rev}_0 and the initial KL is zero. For the forward direction, the chain rule takes expectations under the true path law. Its state marginal is μ at every step, giving the stated constant conditional contribution irrespective of the model's rollout marginals. □

A.7 Sequential scan, copulas, and operator certificates

This parameterization is distinct from the parallel joint architecture in Section 3:

$$y_i = \cos\phi_i \odot x_i + \sin\phi_i \odot \varepsilon_i, \qquad \phi_i = \phi_i(y_{<i}, x_{>i}, \alpha), \quad \varepsilon_i \sim \gamma. \tag{19}$$

Here i indexes blocks, and α is an independent conditioning variable.

Exactness of the sequential scan. Fix the independent action coordinate α and induct on i. Write $C_i := (y_{1:i-1}, x_{i+1:B}, \alpha)$ for the conditioning set. The inductive hypothesis is $(y_{1:i-1}, x_{i:B}) \mid \alpha \sim \gamma_d$. Given C_i, the block x_i is conditionally γ — the inductive hypothesis is a product Gaussian law conditional on α. Since $\varepsilon_i \sim \gamma$ is fresh and independent of everything, and ϕ_i is a function of C_i only, $y_i = \cos\phi_i x_i + \sin\phi_i \varepsilon_i$ is, conditionally on C_i, Gaussian with mean 0 and variance $\cos^2\phi_i + \sin^2\phi_i = 1$, and is independent of C_i. Hence $(y_{1:i}, x_{i+1:B}) \sim \gamma_d$, closing the induction. Including an earlier *input* x_j with $j < i$ in the arguments of ϕ_i can invalidate this conditioning argument: the resulting angle need not be C_i-measurable, and the conditional Gaussian calculation no longer applies. Appendix I exhibits a scalar counterexample in which each branch is measure preserving and the mixture is not. □

Lemma 40 (Invariance under noise-only reparameterization)**.** *Let a transition use a fresh independent Gaussian noise vector ε. Replacing it by $W(\varepsilon)$ for a fixed Gaussian-preserving map W leaves the conditional transition kernel unchanged. Parameters affecting only this replacement are therefore unidentifiable from that kernel.*

Proof. For each fixed state and action, the noise argument still has law $W_{\#}\gamma = \gamma$. Integrating the same transition function against this unchanged law gives the same conditional output law. State-dependent variability must enter the transition's dependence on state, rather than an independent preserving transformation of its noise alone. □

Theorem 41 (Gaussian-preserving scalar kernels and copulas)**.** *Let $k(z, \cdot)$ be a Markov kernel on $\mathbb{R}$ with density. Substituting $p = \Phi(z)$, $u = \Phi(y)$ and setting $c(p, u) := k(\Phi^{-1}p, \Phi^{-1}u)/\varphi(\Phi^{-1}u)$ defines a bijection, modulo null sets, between*

1. *Markov kernels k with $\gamma_1 k = \gamma_1$, and*
2. *copula densities c on $(0, 1)^2$, i.e. non-negative c with $\int_0^1 c(p, u)\,\mathrm{d}u = 1$ for a.e. p and $\int_0^1 c(p, u)\,\mathrm{d}p = 1$ for a.e. u (doubly stochastic kernels on the unit square).*

The OU rotation $y = \rho z + \sqrt{1-\rho^2}\,\xi$ with $|\rho| < 1$ corresponds to the Gaussian *copula with correlation ρ. The set in (2) is convex and infinite-dimensional; the Gaussian copulas form a one-parameter curve inside it.*

Proof. Normalization of k is $\int k(z, y)\,\mathrm{d}y = 1$; substituting $y = \Phi^{-1}u$, $\mathrm{d}y = \mathrm{d}u/\varphi(\Phi^{-1}u)$ gives $\int_0^1 c(p, u)\,\mathrm{d}u = 1$. Prior preservation is $\int k(z, y)\varphi(z)\,\mathrm{d}z = \varphi(y)$; substituting $z = \Phi^{-1}p$ and dividing by $\varphi(y) = \varphi(\Phi^{-1}u)$ gives $\int_0^1 c(p, u)\,\mathrm{d}p = 1$. Both substitutions are invertible, so the correspondence is a bijection. For the OU case, (z, y) is a centered bivariate Gaussian with correlation ρ and unit marginals, whose copula is by definition the Gaussian copula with that correlation. Convexity of (2) is immediate since both constraints are linear; infinite-dimensionality follows because for any mean-zero $h \in L^\infty((0, 1)^2)$ with $\int h(p, u)\mathrm{d}u = \int h(p, u)\mathrm{d}p = 0$, $1 + \epsilon h$ is a copula density for small ϵ, and such h span an infinite-dimensional space (e.g. the products $\cos(2\pi jp)\cos(2\pi ku)$, $j, k \geq 1$). □

The scalar correspondence identifies a concrete stochastic-family restriction. Conditional coefficients can vary with admissible untouched coordinates in a multivariate scan, so the full architecture is not a single global one-parameter family. The example in Appendix J.2 isolates the constant-correlation scalar case.

Definition 42 (Markov operator and mean-zero norm)**.** For a γ-preserving kernel $\tilde{K}$ define $P_\theta : L^2(\gamma) \to L^2(\gamma)$ by $(P_\theta f)(z) := \int f(z')\,\tilde{K}(z, \mathrm{d}z')$, and let

$$\sigma_2 \;:=\; \|P_\theta\|_{L^2_0(\gamma)} \;=\; \sup\Big\{\, \|P_\theta f\|_{L^2(\gamma)} : \mathbb{E}_\gamma f = 0,\ \|f\|_{L^2(\gamma)} = 1\Big\}, \tag{20}$$

the operator norm on the mean-zero subspace $L^2_0(\gamma) = \mathbf{1}^\perp$.

Theorem 43 (One-step contraction rate)**.** *Let $\tilde{K}$ be γ-preserving. Then: **(i)** $P_\theta \mathbf{1} = \mathbf{1}$, $\|P_\theta\|_{L^2(\gamma)} = 1$, and P_θ maps $L^2_0(\gamma)$ into itself, so equation 20 is well posed and $\sigma_2 \leq 1$. **(ii)** For any initial law $\mu_0 = (1 + h)\gamma$ with $h \in L^2_0(\gamma)$ — i.e. any μ_0 of finite χ^2 divergence from γ — the χ^2 divergence contracts geometrically,*

$$\chi^2(\mu_0 \tilde{K}^T \,\|\, \gamma)^{1/2} \;\leq\; \sigma_2^T\, \chi^2(\mu_0 \,\|\, \gamma)^{1/2}, \tag{21}$$

*and hence $\left\|\mu_0 \tilde{K}^T - \gamma\right\| \leq \sigma_2^T \chi^2(\mu_0\|\gamma)^{1/2}$ in the total-variation norm Proposition 3 uses. **(iii)** σ_2 is a* one-step *functional: it is determined by the population joint law of (z, z') with $z \sim \gamma$. Its definition is independent of the rollout horizon. **(iv)** Restricting the supremum in equation 20 to any finite-dimensional $V \subset L^2_0(\gamma)$ gives a lower bound $\sigma_2(V) \leq \sigma_2$, and $\sup_n \sigma_2(V_n) = \sigma_2$ along any nested family $V_1 \subset V_2 \subset \cdots$ whose union is dense in $L^2_0(\gamma)$. The operator need not be compact, so the supremum in equation 20 need not be attained and a fixed V need not achieve equality.*

Proof. (i) $P_\theta \mathbf{1} = \mathbf{1}$ is $\tilde{K}$'s normalization. By Jensen applied to the conditional expectation, $\mathbb{E}_\gamma[(P_\theta f)^2] \leq \mathbb{E}_\gamma[P_\theta(f^2)] = \mathbb{E}_\gamma[f^2]$, the last equality being exactly γ-preservation; so $\|P_\theta\| \leq 1$, with equality at $f = \mathbf{1}$. For invariance of the mean-zero subspace, $\mathbb{E}_\gamma[P_\theta f] = \mathbb{E}_\gamma[f] = 0$, again by γ-preservation.

(ii) Write $\mu_0 = (1+h)\gamma$ with $\mathbb{E}_\gamma h = 0$, so $\chi^2(\mu_0\|\gamma) = \|h\|^2$. The density of $\mu_0\tilde{K}^T$ against γ is $1 + P_\theta^{*T}h$, where P_θ^* is the adjoint in $L^2(\gamma)$ — the time reversal of $\tilde{K}$ — and $\|P_\theta^*\|_{L_0^2} = \sigma_2$ because an adjoint has the same norm. Iterating gives $\left\|P_\theta^{*T}h\right\| \le \sigma_2^T \|h\|$, which is equation 21; the total-variation consequence is $\left\|\mu_0\tilde{K}^T - \gamma\right\| = \mathbb{E}_\gamma|P_\theta^{*T}h| \le \left\|P_\theta^{*T}h\right\|$ by Cauchy–Schwarz.

(iii) By disintegration the one-step joint law $\gamma(\mathrm{d}z)\tilde{K}(z,\mathrm{d}z')$ determines $\tilde{K}(z,\cdot)$ for γ-almost every z, hence determines P_θ as an operator on $L^2(\gamma)$ and therefore σ_2.

(iv) A supremum over a subset is no larger, which gives $\sigma_2(V) \le \sigma_2$ for every V. For the nested statement, let $\bigcup_n V_n$ be dense in $L_0^2(\gamma)$ and fix $\epsilon > 0$. By definition of the supremum there is $f \in L_0^2(\gamma)$, $\|f\| = 1$, with $\|P_\theta f\| > \sigma_2 - \epsilon$; by density there is $g \in V_n$ for some n with $\|f-g\|$ small, and $\|P_\theta g\| \ge \|P_\theta f\| - \|P_\theta\|\,\|f-g\| \ge \|P_\theta f\| - \|f-g\|$ since $\|P_\theta\| \le 1$ by (i). Normalising g changes its Rayleigh quotient by $O(\|f-g\|)$, so $\sigma_2(V_n) > \sigma_2 - 2\epsilon$ for n large, and the supremum over the family is σ_2. Note this argument needs only $\|P_\theta\| \le 1$ and never that a maximiser exists. □

Remark 44 (Comparing contraction criteria). For scalar OU with $0 < |c| < 1$, the Dobrushin coefficient is one but $\sigma_2 = |c|$. A deterministic preserving bijection is unitary and has $\sigma_2 = 1$; this excludes strict one-step density contraction but does not exclude decay of correlations of individual observables.

Corollary 45 (Static level and convergence rate)**.** *Let* $\mathrm{Law}(\hat{x}_T) = (g_\theta)_\#(\mu_0\tilde{K}^T)$. *Suppose* $\mu_0 \ll \gamma$ *and* $\beta_\chi = \chi^2(\mu_0\|\gamma)^{1/2} < \infty$. *For a total-variation-dominated metric, or for energy distance on a bounded observation space,*

$$\left|D(\mathrm{Law}(\hat{x}_T), p_{\mathrm{data}}) - D(p_\theta, p_{\mathrm{data}})\right| \le C\sigma_2^T\beta_\chi, \tag{22}$$

where C *is the corresponding domination constant; for energy distance one may take* $C = 2\,\mathrm{diam}(\mathcal{X})$ *as in Remark 34. If* $0 < \sigma_2 < 1$ *and* $\beta_\chi > 0$*, a sufficient integer horizon for error at most* ϵ *is*

$$T \ge \max\left\{0, \left\lceil \frac{\log(\epsilon/(C\beta_\chi))}{\log\sigma_2} \right\rceil\right\}.$$

This is a population upper bound on the required horizon, not an equality for any particular observable. Substituting a finite-basis estimate for σ_2 *does not preserve the guarantee and does not yield a lower bound on the actual stopping time.*

Proof. Theorem 43 bounds the total-variation deviation by $\sigma_2^T\beta_\chi$. Apply Proposition 3's metric argument, or Remark 34 for bounded-space energy distance, and solve the resulting inequality for T. If $\beta_\chi = 0$, the deviation is zero already; if $\sigma_2 = 0$, it is zero after one step. □

Remark 46 (Finite-basis estimation). For an orthonormal finite basis $\{h_i\}$ of $V \subset L_0^2(\gamma)$, one-step samples estimate $M_{ij} = \mathbb{E}[h_i(z)h_j(z')]$. Its largest singular value is the norm of the compression $\Pi_V P_\theta \Pi_V$, and

$$\|\Pi_V P_\theta \Pi_V\| \le \|P_\theta|_V\| \le \sigma_2.$$

Compression and restriction agree when V is invariant under P_θ. For dense nested V_n, their projections Π_n converge strongly to the identity. For fixed unit f, $\Pi_n P \Pi_n f \to Pf$ in L^2; taking f arbitrarily close to a norm-maximizing direction proves $\|\Pi_n P \Pi_n\| \to \|P\|$. No maximizing function or compactness of P is required.

Sample singular values can exceed the population norm because of sampling and selection error; cross-fitting does not convert them to certified upper bounds. Coordinatewise Hermites without interaction terms are not a dense multivariate family. The absolute bounds in Appendix B use architectural caps and a specified initial second moment instead.

Theorem 47 (Architectural contraction bound)**.** *Let the transition be* $\Psi_{\mathrm{post}} \circ N \circ \Psi_{\mathrm{pre}}$*, where both deterministic maps are Gaussian-preserving diffeomorphisms and* N *is the full* B*-block scan equation 19. Suppose a uniform cap* $c_\star \in (0,1)$ *satisfies*

$$\operatorname*{ess\,sup}_{i,j,\,\mathrm{context}} |\cos\phi_{i,j}| \le c_\star. \tag{23}$$

Then ***(i)*** $\sigma_2 = \|N\|_{L^2_0(\gamma)}$ *and* ***(ii)*** $\|N\|_{L^2_0(\gamma)} \le c_\star$. *For example, an architectural range* $\phi_{i,j} \in [\phi_{\min}, \pi - \phi_{\min}]$, *with* $\phi_{\min} \in (0, \pi/2)$, *permits* $c_\star = \cos\phi_{\min}$. ***(iii)*** *For* $0 < \beta_\chi = \chi^2(\mu_0\|\gamma)^{1/2} < \infty$,

$$T \ge \max\left\{0, \left\lceil \frac{\log(\epsilon/\beta_\chi)}{\log c_\star} \right\rceil \right\} \quad \Longrightarrow \quad \chi^2(\mu_0 \tilde{K}^T \|\gamma)^{1/2} \le \epsilon. \tag{24}$$

The averaged controlled operator has the same upper bound if, for every fixed independent chart input α, *its state transition has this factorization and the cap holds uniformly in* α. *Invariance of an arbitrary augmented map alone does not imply this contraction bound.*

Proof. (i) $U_B f := f \circ B$ for a γ-preserving bijection B is a unitary of $L^2(\gamma)$ fixing $\mathbf{1}$, hence a unitary of $L^2_0(\gamma)$. Writing $g = f \circ \Psi_{\text{post}}$, $(P_\theta f)(z) = \mathbb{E}_\varepsilon\big[g\big(N(\Psi_{\text{pre}}(z), \varepsilon)\big)\big] = \big(U_{\Psi_{\text{pre}}} N g\big)(z)$, so $P_\theta = U_{\Psi_{\text{pre}}}\, N\, U_{\Psi_{\text{post}}}$ and the outer factors, being unitary, leave the norm unchanged.

(ii) Write $N = K_1 K_2 \cdots K_B$, where K_i replaces block i by $\cos\phi_i \odot (\cdot) + \sin\phi_i \odot \varepsilon_i$ with ϕ_i a function of the *other* blocks. Each K_i preserves γ. Let $A_i := \mathbb{E}[\,\cdot \mid z_{-i}]$ be the projection averaging out block i, and $\Pi_{<i} := \mathbb{E}[\,\cdot \mid z_1, \ldots, z_{i-1}]$; note $A_B = \Pi_{<B}$ and $\Pi_{<1} = \mathbb{E}_\gamma$, which is 0 on L^2_0. Two facts:

(a) One step contracts exactly the part that depends on the block it refreshes. For fixed context z_{-i}, K_i acts on block i as a tensor product of scalar Ornstein–Uhlenbeck kernels with coefficients $\cos\phi_{i,j}(z_{-i})$, which fixes constants and contracts the mean-zero part by $\max_j |\cos\phi_{i,j}(z_{-i})| \le c_\star$. Integrating over z_{-i},

$$\|K_i h\|^2 \;\le\; \|A_i h\|^2 + c_\star^2\big(\,\|h\|^2 - \|A_i h\|^2\,\big). \tag{25}$$

(b) A partial scan preserves zero conditional mean over its updated blocks. Let $N_{i+1} := K_{i+1} \cdots K_B$ and suppose $\Pi_{\le i} h = 0$. Starting from $z \sim \gamma$, the output W of N_{i+1} agrees with z on blocks $\le i$, and the exactness induction below equation 19 gives $(z_{\le i}, W_{>i}) \sim \gamma$ with the two groups independent. Hence $\mathbb{E}[h(W) \mid z_{\le i}] = \int h(z_{\le i}, u)\,\gamma(\mathrm{d}u) = (\Pi_{\le i} h)(z_{\le i}) = 0$, i.e.

$$\Pi_{\le i} h = 0 \implies \Pi_{\le i} N_{i+1} h = 0. \tag{26}$$

Now prove by downward induction on i that for all $f \in L^2(\gamma)$,

$$\|N_i f\|^2 \;\le\; \|\Pi_{<i} f\|^2 + c_\star^2\big(\,\|f\|^2 - \|\Pi_{<i} f\|^2\,\big), \qquad N_i := K_i \cdots K_B. \tag{27}$$

For $i = B$ this is equation 25 with $A_B = \Pi_{<B}$. Assume equation 27 at $i+1$ and set $g := N_{i+1} f$, $P := \|\Pi_{<i} f\|^2$, $Q := \|\Pi_{\le i} f\|^2$, $F := \|f\|^2$, so $P \le Q \le F$. Split $f = f_1 + f_2$ with $f_1 := \Pi_{\le i} f$. Since f_1 does not depend on blocks $> i$ and each K_j, $j > i$, alters only block j, we have $N_{i+1} f_1 = f_1$ and therefore $A_i N_{i+1} f_1 = A_i \Pi_{\le i} f = \Pi_{<i} f$. For f_2, $\Pi_{\le i} f_2 = 0$, so the induction hypothesis gives $\|N_{i+1} f_2\|^2 \le c_\star^2 \|f_2\|^2$ and equation 26 gives $\Pi_{<i} A_i N_{i+1} f_2 = \Pi_{<i} N_{i+1} f_2 = 0$. The two pieces of $A_i g$ are thus orthogonal — one is $\Pi_{<i}$-measurable, the other is annihilated by $\Pi_{<i}$ — and

$$\|A_i g\|^2 = \|\Pi_{<i} f\|^2 + \|A_i N_{i+1} f_2\|^2 \;\le\; P + c_\star^2 (F - Q). \tag{28}$$

Applying equation 25 to $N_i = K_i g$ and then equation 27 at $i+1$ and equation 28,

$$\begin{aligned} \|N_i f\|^2 \;&\le\; c_\star^2 \|g\|^2 + (1 - c_\star^2)\,\|A_i g\|^2 \\ &\le\; c_\star^2\big[Q + c_\star^2(F - Q)\big] + (1 - c_\star^2)\big[P + c_\star^2(F - Q)\big] \;=\; P + c_\star^2 (F - P), \end{aligned}$$

which is equation 27 at i. At $i = 1$, $\Pi_{<1} f = 0$ for $f \in L^2_0$, giving $\|Nf\| \le c_\star \|f\|$.

(iii) Combine (i), (ii) and Theorem 43(ii) and solve for T.

For the controlled extension, the theorem assumes the state factorization and cap for every fixed chart input α. Parts (i)–(ii) therefore give $\|P_{\theta,\alpha}\|_{L^2_0(\gamma)} \le c_\star$ uniformly in α. This is an additional condition on the state kernel; it does not follow merely from preservation of an augmented state–action law.

Now average. With $\alpha \sim \gamma_m$ independent of z (Lemma 24), the joint law is $\gamma_d \otimes \gamma_m$, so by Jensen and then Fubini

$$\|P_\theta f\|^2 = \mathbb{E}_z\big[(\mathbb{E}_\alpha P_{\theta,\alpha} f(z))^2\big] \le \mathbb{E}_z \mathbb{E}_\alpha\big[(P_{\theta,\alpha} f(z))^2\big] = \mathbb{E}_\alpha \|P_{\theta,\alpha} f\|^2 \le c_\star^2 \|f\|^2.$$

This averaging requires the same Gaussian action law at every state. □

Remark 48 (Sharpness and scope). Constant coefficients equal to $c_\star$ give a tensor product OU kernel with norm $c_\star$, so the class bound is sharp. The cap must control the whole conditioner range; a sampled maximum does not establish it. The frozen-latent initial data banks are atomic and have infinite initial χ^2. This theorem supplies a relative L^2 rate, while the finite-bank absolute bound in Appendix B applies to the separate parallel architecture and its stated hypotheses.

B AN EXPLICIT CONVERGENCE CERTIFICATE FOR PARALLEL CONTROLLED TRANSITIONS

The following construction combines joint Gaussian-preserving maps with a parallel Gaussian noise stage. Its convergence bound uses a second moment of the specified initialization law, including a finite empirical law. The analytic ingredients are Gaussian entropy contraction and relative-entropy data processing; the contribution here is their application to a transition with an exact conditional likelihood and a controlled occupancy constraint. Theorem 4 combines the joint-input result and finite-start corollary proved below.

B.1 JOINT MIXING AND PARALLEL CONDITIONAL NOISE

Let $\Psi : \mathbb{R}^{d+m} \to \mathbb{R}^{d+m}$ be a Gaussian-preserving bijection. The construction is the parallel transition equation 4 of Section 3, whose notation we reuse: given a state z and action chart coordinate α, the joint mixer produces $(h, b) = \Psi(z, \alpha)$, the noise-stage coefficients satisfy $|c_i(b)| \le c_\star < 1$ uniformly with fresh $\varepsilon \sim \gamma_d$, and each post-map S_b preserves γ_d. The families are jointly measurable. For conditional likelihoods, we use diffeomorphisms S_b with computable inverses and determinants. Joint orthogonal couplings provide Ψ; preserving couplings provide S_b. The parallel implementation uses rotations in both positions. Norm-changing preserving post-maps are also admissible. For $m = 0$, the action and residual-action blocks are absent, with γ_0 the one-point probability space.

Under an exact behaviour action chart, α is Gaussian independently of z, including for observation-conditioned policies by Lemma 25. Observation context enters this transition only through that chart. At the joint reference, (h, b) is therefore product Gaussian. Conditional on b, the noise stage and S_b retain γ_d. The state output consequently preserves the reference, even though its law at a fixed physical action need not do so. Every state coordinate receives noise in one parallel stage; the coefficients read b, not the coordinates being refreshed.

Theorem 49 (Joint-input entropy contraction)**.** *Let M be the kernel from (z, α) to z' defined by equation 4. For any joint input law ν,*

$$D_{\mathrm{KL}}(\nu M \| \gamma_d) \le c_\star^2 D_{\mathrm{KL}}(\nu \| \gamma_{d+m}). \tag{29}$$

In particular, under an exact behaviour action chart — including observation-conditioned charts, by Lemma 25 — the joint input is $\mu \otimes \gamma_m$ and the behaviour-averaged state kernel K obeys $D_{\mathrm{KL}}(\mu K \| \gamma_d) \le c_\star^2 D_{\mathrm{KL}}(\mu \| \gamma_d)$.

Proof. If $c_\star = 0$, the conditional noise stage is a complete refresh, which proves the claim directly. Otherwise, assume the input divergence is finite; the infinite-divergence case is immediate. Write $\rho = \Psi_\# \nu$. Divergence invariance under a preserving bijection and the chain rule give

$$D_{\mathrm{KL}}(\nu \| \gamma_{d+m}) = D_{\mathrm{KL}}(\rho_b \| \gamma_m) + \mathbb{E}_{\rho_b} D_{\mathrm{KL}}(\rho_{h|b} \| \gamma_d).$$

For a constant Ornstein–Uhlenbeck (OU) coefficient $c = e^{-s}$, Gaussian logarithmic Sobolev inequality and entropy dissipation yield $D_{\mathrm{KL}}(\lambda Q_c \| \gamma_d) \le c^2 D_{\mathrm{KL}}(\lambda \| \gamma_d)$ (Gentil, 2010). Indeed, $s \mapsto Q_{e^{-s}}$ is the Ornstein–Uhlenbeck flow, and along $\mu_s = \lambda Q_{e^{-s}}$ the de Bruijn identity gives $\frac{d}{ds} D_{\mathrm{KL}}(\mu_s \| \gamma_d) = -I(\mu_s \| \gamma_d)$, with $I(\cdot \| \gamma_d)$ the relative Fisher information. The Gaussian logarithmic Sobolev inequality with constant one states $I(\mu_s \| \gamma_d) \ge 2 D_{\mathrm{KL}}(\mu_s \| \gamma_d)$, so integration over s multiplies the entropy by at most $e^{-2s} = c^2$. The inequality extends from smooth densities to finite-entropy laws by approximation and lower semicontinuity.

At fixed b, factor the diagonal OU stage into an OU kernel with scalar coefficient $c_\star$ followed by coordinatewise OU kernels with coefficients $c_i(b)/c_\star$. The latter preserve the reference, including a reflection for a negative coefficient. Data processing therefore gives the factor $c_\star^2$. Apply data processing through S_b and convexity over the actual law ρ_b to obtain

$$D_{\mathrm{KL}}(\nu M \| \gamma_d) \le c_\star^2 \mathbb{E}_{\rho_b} D_{\mathrm{KL}}(\rho_{h|b} \| \gamma_d),$$

which proves equation 29. The chain-rule term includes the dependence between h and b away from the reference. Under the behaviour chart the joint input is $\mu \otimes \gamma_m$, giving the state-kernel statement. □

B.2 A FINITE WARM START AND A DECODED PLATEAU BOUND

Corollary 50 (Absolute convergence from a finite-second-moment law)**.** *Under the exact behaviour action chart, let Ψ also preserve the Euclidean norm pointwise, and assume $M_2 = \mathbb{E}_{\mu_0}\|Z\|^2 < \infty$. With $B_\star$ as defined in equation 8, for every integer $T \geq 1$,*

$$D_{\mathrm{KL}}(\mu_0 K^T \| \gamma_d) \leq c_\star^{2(T-1)} B_\star, \tag{30}$$

and the total-variation bound equation 9 applies with the same $B_\star$. Initial absolute continuity is unnecessary. For a specified uniform finite bank $\{z_j\}_{j=1}^n$, one may use its exact moment $M_2 = n^{-1}\sum_j \|z_j\|^2$.

Proof. Condition on (h, b) at the first transition. Convexity of relative entropy and data processing through S_b bound its output divergence by

$$\frac{1}{2}\mathbb{E}\sum_i \left[c_i(b)^2 h_i^2 - c_i(b)^2 - \log(1 - c_i(b)^2)\right].$$

The function $s \mapsto -s - \log(1-s)$ is increasing on $[0, 1)$. Furthermore, norm preservation and the fresh Gaussian action coordinate give $\mathbb{E}\|h\|^2 \leq M_2 + m$. The expression is consequently at most $B_\star$. This Gaussian-mixture argument applies to atomic initial laws as well. Theorem 49 contracts the finite first-step divergence at all later transitions. Pinsker's inequality in the total-variation norm used here is $\|P - Q\|_{\mathrm{TV}} \leq \sqrt{2 D_{\mathrm{KL}}(P\|Q)}$. □

For $c_\star \in (0, 1)$, a sufficient horizon for total-variation error at most $\epsilon > 0$ is

$$T \geq 1 + \max\left\{0, \left\lceil \frac{\log(\sqrt{2B_\star}/\epsilon)}{-\log c_\star} \right\rceil \right\}.$$

At $c_\star = 0$ the noise stage is complete refresh and the output is exactly Gaussian after one step. A finite bank gives a certificate for that specified initialization law. Replacing M_2 by a sample estimate for an unknown population requires a valid population moment upper bound.

Let F be a measurable decoder with output in a metric space of diameter Δ. For the population energy statistic $\mathcal{E}$ defined in Remark 34,

$$|\mathcal{E}(P, R) - \mathcal{E}(Q, R)| \leq 2\Delta \, \|P - Q\|_{\mathrm{TV}}. \tag{31}$$

To see this, put $\xi = P - Q$ and $f_S(x) = \int \mathrm{dist}(x, y) S(\mathrm{d}y)$. The difference equals $2\int f_R \mathrm{d}\xi - \int (f_P + f_Q)\mathrm{d}\xi$. For a zero-mass signed measure, $|\int f \mathrm{d}\xi| \leq \mathrm{osc}(f)\, \|\xi\|_{\mathrm{TV}}/2$; the two oscillations are bounded by Δ and 2Δ. Since decoder pushforward contracts total variation, the static reference $p_\theta = F_\# \gamma_d$ thus satisfies

$$|\mathcal{E}(F_\#(\mu_0 K^T), p_{\mathrm{data}}) - \mathcal{E}(p_\theta, p_{\mathrm{data}})| \leq 2\Delta \min\{2, \sqrt{2B_\star}\, c_\star^{T-1}\}.$$

The distance must be bounded in the units actually used by the evaluator; an unbounded feature or probe distance needs separate moment control.

B.3 POLICY DEVIATION IN THE ORIGINAL ACTION SPACE

Let O_t denote the policy context, including a generated observation when the policy reads frames, and let $\lambda_t(\mathrm{d}z, \mathrm{d}o) = \mu_t(\mathrm{d}z) L_t(\mathrm{d}o \mid z)$ be its joint law. Assume the exact chart $G_{z,o}$ is a bimeasurable bijection modulo common null sets, maps $\pi_b(\cdot \mid z, o)$ to γ_m, and supplies the transition's only dependence on o. At each fixed (z, o),

$$D_{\mathrm{KL}}((G_{z,o})_\# \pi_t(\cdot \mid z, o) \| \gamma_m) = D_{\mathrm{KL}}(\pi_t(\cdot \mid z, o) \| \pi_b(\cdot \mid z, o)).$$

The same equality holds for total variation. Define

$$H_t = D_{\mathrm{KL}}(\mu_t \| \gamma_d), \qquad \kappa_t = \mathbb{E}_{(Z,O)\sim\lambda_t} D_{\mathrm{KL}}(\pi_t(\cdot \mid Z, O) \| \pi_b(\cdot \mid Z, O)).$$

Lemma 25 gives a joint (Z_t, α_t) input divergence at most $H_t + \kappa_t$; marginalizing a stochastic observation can make the inequality strict. If the context is a function of the state, this reduces to the state-conditioned equality. Theorem 49 therefore gives the policy recurrence equation 10; its second expression assumes $H_0 < \infty$. From Gaussian initialization and $\kappa_t \le \kappa$, it implies

$$\|\mu_T - \gamma_d\|_{\mathrm{TV}} \le \min\left\{2, \sqrt{\frac{2c_\star^2(1-c_\star^{2T})\kappa}{1-c_\star^2}}\right\}.$$

An atomic state initialization may instead receive a behaviour-policy warm start before introducing the policy shift. For a fitted invertible chart, the same formulas use its induced policy $\widehat{\pi}_b(\cdot \mid z, o) = (G_{z,o}^{-1})_{\#}\gamma_m$; identifying this with the true behaviour policy requires checking the fitted conditional model. A small total-variation chart residual alone does not upper-bound a KL residual. The policy expectation is over the joint state and context law visited by the deployed model. For two nonsingular tanh-Gaussian policies with the same tanh transformation, the conditional divergence is the Gaussian KL before tanh. Deterministic action selection, including a deterministic cross-entropy-method (CEM) choice, has infinite KL against a nonsingular tanh-Gaussian behaviour law; the resulting KL bound is vacuous for that policy.

For the more general controlled kernel of Theorem 30, Pinsker and Jensen still give $\eta_t \le \min\{2, \sqrt{2\kappa_t}\}$. This generic perturbation result and the entropy contraction of equation 10 have different architectural hypotheses.

B.4 An optional reference component

Let $\Gamma(z, \cdot) = \gamma_d$ and choose a constant $0 < r < 1$. For any reference-preserving base kernel K_0, define $K_r = (1-r)K_0 + r\Gamma$. This standard independent-refresh construction gives, for every initialization,

$$\left\|\mu_0 K_r^T - \gamma_d\right\|_{\mathrm{TV}} \le 2(1-r)^T. \tag{32}$$

Indeed, every zero-mass signed measure ξ satisfies $\xi\Gamma = 0$, and hence $\xi K_r = (1-r)\xi K_0$. This argument requires a reset of the entire state appearing in the bound, including any recurrent memory.

For a conditional density the mixture has $k_r = (1-r)k_0 + r\gamma_d$, so pointwise

$$-\log k_r \le \min\{-\log k_0 - \log(1-r),\ -\log\gamma_d - \log r\}. \tag{33}$$

With an invertible observation chart, the second bound becomes the model's marginal negative log likelihood minus $\log r$. It bounds conditional loss relative to the fitted marginal, whose accuracy remains determined by the chart. Sampling uses an actual reference draw with probability r, and the density is evaluated by a log-sum-exp, without clipping likelihoods.

For the parallel base kernel, convexity further replaces the entropy contraction coefficient by $q = (1-r)c_\star^2$ and the warm-start bound by $(1-r)B_\star$. Thus its absolute KL bound is $(1-r)B_\star q^{T-1}$, and its policy recurrence is $H_{t+1} \le q(H_t + \kappa_t)$. Under the behaviour policy, one may take the smaller of the resulting Pinsker bound and equation 32. Independent refresh can interrupt paths; its effect on short-horizon predictions and control must be evaluated separately from its marginal guarantee.

B.5 Conditional likelihood and numerical scope

For fixed (z, a, o), the chart coordinate and the values (h, b) in equation 4 are deterministic. The exact conditional log density is equation 5 of Section 3. The bounds assume the stated mathematical kernel, exact behaviour innovations, and a uniform architectural coefficient cap. Gaussian diagnostics and floating-point comparisons assess the implementation under these hypotheses; a global roundoff certificate is outside their scope.

C Reproducible experimental methods

Four studies connect the construction to model behaviour. Low-dimensional systems isolate conditional representation and preservation. The twelve-task fixed-latent pixel study evaluates transitions

under an external OU action process. CPU diagnostics check conditional loss and the parallel kernel's finite-start bound. The recurrent study jointly trains the inference network, decoder, reward head, and transition from pixels.

C.1 Data and occupancy screening

Lorenz-63. We integrate $\dot{x} = 10(y - x)$, $\dot{y} = x(28 - z) - y$, and $\dot{z} = xy - (8/3)z$ by Euler–Maruyama with independent additive Wiener forcing of amplitude 0.5 per coordinate and step size 0.01. After 10,000 integration steps of burn-in, we retain every fifth state. The 200,000-state trajectory supplies 199,999 consecutive pairs. A contiguous 90/10 split discards the pair crossing the boundary and leaves 19,999 held-out pairs. Consecutive held-out observations remain dependent. The nonstationary control ramps ρ from 28 to 45 over the recorded window, with $\rho = 28$ during burn-in.

Sequential-scan pixel collection. DeepMind Control (Tassa et al., 2018) observations are 64×64 RGB images, collected with action repeat two. The driver is an OU process with time scale 8 and noise scale 0.8, clipped to the action box; episodes last 60,000 steps. The simulator together with the driver is time homogeneous. The observed frame alone need not be Markov, and clipped OU actions are not exact Gaussian innovations. These experiments use the sequential-scan parameterization, which preserves the reference for every fixed external action. A reference start independent of the entire external action sequence and transition noise then retains its reference marginal conditional on that sequence. This is a narrower experiment than the state-dependent occupancy construction in Section 3.

The screen covers 22 tasks with 96 independent chains each. Eighteen pass its temporal and coverage checks. Finger turn hard and fish upright have unresolved transients; cheetah run and stacker stack 2 develop concentrated occupancies. The twelve evaluated tasks are the headline subset of the eighteen, spanning state dimensions 3–53 and 1–21 actuators; the six other healthy tasks (walker stand, ball in cup catch, cartpole two poles, point mass easy, swimmer with six links, and fish swim) remain screened and available but are not fitted. The twelve evaluated tasks are all reported in Table 16. Passing this finite-data screen does not prove stationarity or a latent Markov representation. The screen compares first and second moments in 2,000-step windows against late windows beginning after 30,000 steps, using chain-level averages and Bonferroni adjustment across windows. A matched leave-one-out window-to-window energy-distance baseline additionally checks distributional homogeneity. Autocorrelation times and effective sample sizes describe temporal dependence. Late-to-widest coordinate spread and frozen-coordinate counts check coverage separately from temporal stability.

Swimmer (15 links) is retained as an ensemble occupancy with a metastability flag: its between-chain distance is 0.0186 against a matched late-window baseline of 0.00065. Individual chains need not explore that ensemble. A separate cheetah run with 128 chains and 200,000 steps still concentrates its torso pose while its limbs move. Its exclusion concerns this exploration policy's occupancy, rather than the task's value as a control benchmark. For a homogeneous process, the $2/N$ invariance-defect bound for a length-N time-averaged population law is separate from these diagnostics; repeated chains do not increase N in that statement.

C.2 Sequential-scan fitting objectives and budgets

For low-dimensional systems, the state chart E has eight flow layers with hidden width 256. Four deterministic layers precede and follow a block-scan noise stage. Each deterministic layer combines rotation couplings and quantile swirls with three rounds, three profiles per round, and radial levels $(2, 3, 4)$. The sequential block scan is distinct from the parallel construction of Section 3; its conditioning and operator bound are described in Appendix I.2.

The pair-likelihood experiments minimize

$$-\frac{1}{d}\,\mathbb{E}_{(x,a,x')\sim\widehat{\pi}}\left[w \log p_E(x) + \log p_\theta(x' \mid x, a)\right].$$

Here d is the modeled state dimension, including $d = 256$ for the frozen pixel latent, and $w = 1$ unless swept. Marginal and conditional density changes include the appropriate learned-chart

Jacobians. The low-dimensional pair trainer and pixel trainer use AdamW with zero weight decay, initial learning rate 10^{-3}, cosine decay, and gradient clipping at 10. Nonfinite updates are recorded and training aborts if their allowed fraction is exceeded. The default batch sizes are 16,384 and 2,048, respectively. Raw-latent Gaussian heads minimize only the conditional NLL; conditional flow matching minimizes its regression objective. Thus a flow-matching loss is not a comparable likelihood value.

Table 2: Training budgets. Counts are gradient updates per configuration. A training seed is a fitted-model replication; repeated rollout samples are not additional training seeds.

Comparison	Updates	Replication and scope
Lorenz prediction–rollout calibration	40,000	Three training seeds; two noise blocks
Lorenz marginal-weight sweep	30,000	Four constrained weights and two free arms; three blocks
Lorenz soft stabilization	8,000	Nine penalty/weight cells at seed 0; selected arm at three seeds
Frozen-latent pixel transitions	30,000	Twelve tasks; one fitted seed per task/arm
Capped pixel transitions	30,000	Twelve tasks at .999; three-task cap sweep and retraining controls
Controlled Gaussian and Duffing	20,000	One seed per Gaussian configuration; three planning seeds
Finite-divergence memory comparison	4,000	Six synthetic configurations
Non-Gaussian copula example	12,000	Three training seeds

The seven-model Lorenz calibration combines three 40,000-update replicates and four 30,000-update weight settings. It tests each fitted model's own static prediction, rather than comparing seven models at identical capacity and budget. The soft-stabilization objective instead uses total three-dimensional pair NLL, without division by d, plus its weighted penalty. Its table therefore reports nats per pair, and its numerical weights cannot be transferred directly to the per-dimension objective above.

C.3 Distances, uncertainty, and reference dependence

The recurrent and native studies use root-mean-square (RMS) distance over normalized RGB channels. The fixed-latent studies use unnormalized Euclidean distance, either in 256-dimensional frozen-autoencoder coordinates (including the Reacher static calibration) or on flattened, downsampled decoded images, as labeled. These scores have different units; the diameter-one bound applies only to the normalized pixel geometry.

The population energy functional is estimated by a complete two-sample U-statistic in the synthetic and fixed-latent studies. The recurrent study instead uses an unbiased linear estimator: each repetition splits 1,024 samples into 512 disjoint pair blocks, averaging their cross- and within-sample distance terms; eight repetitions provide paired evaluation SEs. Unbiasedness in each case assumes independent draws. Finite estimates can be negative. At extreme latent scales, subtraction of large distance terms can also become numerically unreliable. All rollout comparisons report the retained finite fraction; a distance computed after discarding failed chains describes their surviving subpopulation. The alternate-metric study additionally reports the fraction inside its specified box. These quantities cannot be read as unconditional rollout accuracy.

Monte Carlo SEs condition on the fitted model and collected dataset. Where both levels share a reference repetition, the paired SE is computed from their within-repetition differences. The Lorenz seed table has only separate saved level summaries, so its quadrature-combined SE omits the unknown covariance and is labeled nominal. The twelve-task paired comparison uses eight shared reference repetitions; detailed Walker rollout summaries use four. Training-seed variation is reported separately.

Recurrent equal-task SEs and paired percentile intervals condition on the fixed task suite and assume independent task–training-seed cells. The same three numerical seed IDs initialize the random streams across tasks, so cross-task independence is an approximation. Grouping each shared seed

across all tasks instead gives an occupancy-gap SE of 0.000602, compared with the reported within-task SE of 0.000325; the mean remains -0.000182. Both summaries are compatible with zero, but three shared-seed groups provide limited uncertainty calibration. Neither summary covers collection uncertainty or a population of tasks.

Temporal dependence in the reference also matters. Within-reference terms in either estimator can be biased by correlated observations. That common term cancels in a paired difference using the same reference pool, but the cross terms still have conditional sampling variation. The Lorenz disjoint-subset baseline is 0.0007 ± 0.0020 (SE over eight splits), with single-split SD 0.0057. A separate 64-record calibration estimates a correlation-induced offset of 0.01032 ± 0.00022 at lag 128. Pixel reference diagnostics find offsets as large as 0.196 on Swimmer. These are estimand-specific diagnostics, not universal corrections to be subtracted from other scores. Reference sampling uncertainty, fitted-seed variation, and collection-to-collection uncertainty remain different quantities.

C.4 Jointly trained recurrent pixel model and GPU protocol

The model and objective are specified in Appendix C.4.1, and data separation and evaluation in Appendix C.4.2. Initialization, gain, and planning diagnostics appear in Appendices C.4.3–C.4.5; the native baseline and complete results follow in Appendices C.4.6 and C.4.7.

C.4.1 Model and sequential objective

The implemented model jointly learns a four-stage convolutional neural network (CNN) observation encoder, a recursive diagonal-Gaussian posterior $q_\varphi(z_t \mid z_{t-1}, o_t, a_{t-1})$, a convolutional decoder, a reward head, and a transition prior. Every persistent recurrent coordinate belongs to the sampled stochastic vector z_t. There is no additional deterministic recurrent memory or privileged simulator state supplied to the model. The study uses latent, feature, and shared hidden widths of 256, with CNN base width 32 and 64×64 RGB images. Only the prior-network hidden width is adjusted to match transition parameter budgets; the shared backbone remains identical. Occupancy and conditioner-only priors use width 256. Gaussian, overshooting, and invariant-penalty priors use 600, or 608 on Humanoid, Quadruped, and Swimmer. Conditional RealNVP uses 168, or 176 on Hopper, Humanoid, Manipulator, Quadruped, Swimmer, and Walker.

Collection uses a saved observation-conditioned tanh-Gaussian behaviour policy,

$$v_t = m_b(o_t) + s_b(o_t) \odot \alpha_t, \qquad a_t = \tanh v_t, \qquad \alpha_t \sim \gamma.$$

The collection shards store $m_b(o_t)$, $s_b(o_t) > 0$, and v_t, so the training innovation is the known quantity $\alpha_t = (v_t - m_b(o_t))/s_b(o_t)$. The loader checks these fields and the consistency of $\tanh v_t$ with the recorded action. Missing policy context is an error for occupancy training. The fixed policy and its content hash are recorded with the dataset. This chart uses the recorded policy parameters; clipped OU actions in the fixed-latent dataset do not supply these innovations.

Before collection, reward-weighted imitation fits the policy on separate calibration trajectories and freezes the validation-selected checkpoint. Boundary action targets use a recorded inverse-tanh clipping convention; they are supervised targets, not inferred behaviour innovations. Fitting alone does not establish improved return. Fresh independent chains supply the benchmark, with return and within-chain motion reported alongside temporal moment and spread diagnostics. Failed tasks remain in the attempted-task denominator and are not replaced.

All six arms share the same CNN, posterior, decoder, and reward architecture, initialized identically under a common seed. Their priors are the parallel occupancy kernel; the same kernel without joint state–action mixing (*conditioner only*); a diagonal Gaussian; conditional RealNVP; the Gaussian with overshooting; and the Gaussian with an invariant-measure penalty. All use the same innovation input by default. The conditioner-only arm is not fixed-physical-action preservation when the chart depends on the observation. The occupancy prior uses four joint layers, $c_\star = 0.999$, and reset probability 10^{-3} by default. Its certificate concerns the generative prior, not the aggregate variational posterior or an invertible chart of real images.

Sequential objective and units. For L observed frames, the negative evidence lower bound (ELBO) of the joint behaviour-generated sequence model, up to known behaviour-action log densi-

ties, is

$$\mathcal{L}_{\mathrm{ELBO}} = \frac{1}{L}\left(\mathbb{E}_q\left[-\sum_{t=0}^{L-1}\log p_\theta(o_t \mid z_t) - \sum_{t=0}^{L-2}\log p_\theta(R_t \mid z_t, a_t, z_{t+1})\right] + D_{\mathrm{KL}}(q_\varphi(z_0 \mid o_0)\|\gamma_d) + \sum_{t=0}^{L-2}\mathbb{E}_q\left[\log\frac{q_\varphi(z_{t+1} \mid z_t, o_{t+1}, a_t)}{k_r(z_{t+1} \mid z_t, a_t, o_t)}\right]\right]. \quad (34)$$

Under the known behaviour chart this is the implemented conditioning: the prior network reads (z_t, α_t) with the chart coordinate $\alpha_t = G_{z_t,o_t}(a_t)$, an equivalent conditioning to (a_t, o_t) given z_t. The generative ordering supplies the observation context: o_t is emitted from z_t before z_{t+1} is drawn, so $p(o_{t+1}, z_{t+1} \mid z_t, a_t, o_t) = p_\theta(o_{t+1} \mid z_{t+1})\, k_r(z_{t+1} \mid z_t, a_t, o_t)$. The initial KL is analytic; transition terms use reparameterized Monte Carlo samples and the exact one-step density of the selected prior. A sampled KL contribution can be negative and is not clipped. Pixel and latent coordinates are summed; division by L occurs after summation. Unit observation, KL, and reward weights give this ELBO and are used in all primary fits. The Gaussian emission has fixed standard deviation 0.1 on normalized RGB intensities; its sigmoid decoder output is the mean. This is a continuous density, not an exact discrete 8-bit image likelihood. Rewards use a scalar Gaussian with standard deviation one. The outgoing reward is R_t, with timing $o_t \xrightarrow{a_t, R_t} o_{t+1}$.

Overshooting samples intermediate prior states and uses detached posterior targets; the resulting expected final conditional KL is an auxiliary bound, not an exactly evaluated multistep mixture KL. The invariant penalty is biased empirical RBF MMD squared between prior outputs driven by independent reference states/innovations and fresh reference samples; it uses 128 samples and bandwidths $(0.5, 1, 2)$ with squared distances divided by latent dimension. Both auxiliary terms and their configured weights are logged separately. The overshooting horizon is three; the primary overshooting and invariant-penalty weights are 0.1 and 1: these are fixed-configuration mechanism controls, not a best-tuned comparison of all regularization objectives.

C.4.2 Data separation and evaluation

Each task has 32 independently initialized collection trajectories. After 10,000 burn-in decisions per trajectory, we retain 20,000 transitions with action repeat two and camera zero: 640,000 transitions per task. Shards store uint8 frames, outgoing actions/rewards, reset flags, and policy fields. A fixed split seed (9041) assigns 25/3/4 shards to training/validation/test. Training windows never cross reset or shard boundaries and use no additional burn-in; filtering restarts at each window. A training minibatch samples a shard proportional to its valid-window count, then draws 16 windows within it. Held-out sampling uses independent windows. We use Adam at constant learning rate 3×10^{-4}, zero weight decay, and gradient clipping at 100. Validation runs every 1,000 updates on 16 batches; reported evaluations use the final 100,000-update checkpoint.

Evaluation protocol. The recurrent study contains twelve tasks, six primary arms, and three paired training seeds (1101–1103). Each fit uses 100,000 updates of 16-frame windows. The protocol separates held-out forecasts, paired static-versus-rollout distances, interventions on preservation, and pixel-based planning. Policy shift is evaluated only for randomized kernels, without fitted dynamics (Figure 3b); fitted-model policy-shift performance is unmeasured. Posterior initialization, reference sampling, actions, transitions, emissions, and planning have distinct recorded random streams. Empirical initialization banks are frozen before evaluation; their actual second moment enters the finite-start bound. The architectural bound applies in real arithmetic under its stated behaviour innovations. At the evaluation horizon $T = 10^5$ with the default reset probability $r = 10^{-3}$, the population bound on the marginal gap supplied by equation 32 and equation 31 is at most $2\Delta \cdot 2(1-r)^T \le 1.5 \times 10^{-43}$, far below the measurement's $\pm 3 \times 10^{-4}$ equal-task training-seed standard error. This SE includes variation in checkpoint-level estimates after averaging evaluation repetitions; it does not separately resolve their Monte Carlo error. The terminal gap checks numerical consistency with the static reference, whereas intermediate horizons assess bound sharpness (Figure 3, Appendix G). Observation-space diagnostics and finite-particle likelihood estimates remain separate from exact latent conditional densities and analytic bounds.

A physical-action forecast or planner must recompute the known policy's location and scale from its current generated observation. Using a future true frame's policy context would reveal future

observations. Replaying recorded innovations instead defines a distinct behaviour-noise replay experiment. CEM selection uses only model predictions; the simulator executes chosen actions for evaluation. Emission means and samples define different evaluated laws, as do physical-action interventions and behaviour-innovation rollouts. Reported metrics use completed evaluations, with failed cells retained in the relevant denominator.

Short forecasts impose the recorded future physical actions as interventions. They do not condition on the information those actions reveal under the behaviour policy. We evaluate 32 batches of 16 held-out windows, with prefix length four, horizons 1/5/10, and nested 16/64 particles for the approximate filtered observation scores. Long evaluation uses 1,024 chains in each of eight repetitions, a 4,096-state finite initialization bank, and horizons through 10^5. The metric compares bounded decoder means with held-out pixels using RMS pixel distance; the Gaussian emission samples are a distinct unbounded law. Shared references and all raw paired energy terms are retained.

Matching uses validation observation NLL within 0.02 nats and each 1/5/10-step mean squared error (MSE) within $10^{-4} + 0.05$ times the reference error. Parameters must be within 5% and measured training compute within 10%. These are eligibility gates, not equivalence tests. All failed matches remain in the study. A frozen output-gain intervention estimates that specific scaling effect; it is not a general causal effect of stationarity.

C.4.3 Posterior initialization and guard crossings

The conditioner-only arm's single escape cell (swimmer, seed 1103; 21 of $8{,}192$ chains) crosses the latent guard at the first step from large posterior-bank initializations. Every escape is present at horizon one, with per-repetition surviving fractions constant thereafter, and the γ_d-initialized control cloud of the same checkpoint survives $8{,}192$ of $8{,}192$ chains to 10^5 steps. The escaped chains remain finite: their retained pre-censoring states have norms 3.05–7.87×10^6, just above the $\pm 10^6$ guard, and they map onto the eight largest rows of the cell's $4{,}096$-state initialization bank, of which 29 rows have norms 1.06–7.98×10^6 (bank second moment 7.4×10^{10} against $d = 256$ under γ_d). Re-encoding the recorded source windows with the frozen checkpoint reproduces posterior means of norm 1.3–3.7×10^6 at cosine 1.0000 to the bank rows; all 29 rows come from one contiguous segment of a single held-out test episode. This traces the large initial states to the fitted posterior mean on that segment. Fitted transition coefficients stay strictly inside the 0.999 cap, the rotation stack preserves norms to float32 accuracy, and guard-free replays from the same initial states re-enter the typical set within $\approx 3 \times 10^3$ steps. Theorem 2 covers the conditioner-only variant as the specialization of equation 4 whose joint mixer keeps the action block fixed. The bank is not a Gaussian initialization: its second moment exceeds the reference moment by eight orders of magnitude. The finite-start bound permits this early transient (total variation ≤ 1.998 at $T = 1$). The Gaussian arm's escapes on the same task and seed show the opposite signature—progressive leakage appearing between 10^4 and 10^5 steps in both clouds.

C.4.4 Output-gain controls

Table 3 compares the primary model with two output gains. Gain 1.001 is the validation-selected intervention; gain 1.003 is exploratory. Both use the same checkpoints, banks, initialization draws, reference pools, random streams, and evaluator as the unscaled model, verified by per-job input hashes. The gain multiplies the sampled output, including the reference-reset branch, and therefore changes the Gaussian reference law. All cells retain their chains at the measured horizons, while second-moment deviations increase with gain. With reset probability 10^{-3} retained, these measurements characterize finite-horizon survival and marginal sensitivity.

Table 3: **Output scaling and finite-horizon marginal sensitivity.** All gains share checkpoints, initialization banks, and random streams; scaling applies to the sampled prior output, including the reset branch ($r = 10^{-3}$). Survival counts cells retaining all chains at 10^5 steps (36 cells, 1,024 chains $\times$ 8 repetitions each). Marginal gap reports the cell mean [minimum, maximum]. The last column is the largest coordinate second-moment deviation from one at that horizon.

Output gain	Survival	Marginal gap at 10^5	2nd-moment dev.
1.000 (primary)	36/36	-0.00018 [-0.0051, $+0.0065$]	0.15
1.001 (selected)	36/36	-0.00034 [-0.0048, $+0.0072$]	0.54
1.003 (exploratory)	36/36	-0.00081 [-0.0121, $+0.0086$]	6.92

C.4.5 Planning protocol and common-cell contrasts

The primary CEM comparison uses horizon 20, 64 candidates, eight elites, three iterations and four particles, over ten paired 200-decision episodes with four warm-up decisions. The discount is 0.99; the pre-tanh proposal starts at standard deviation 1 and has minimum standard deviation 0.05. It entails 30.72 million imagined transitions per fitted model. Horizons 5 and 50 form a separate prespecified sweep on the four priority tasks with occupancy and Gaussian models. Full imagined-step cost, GPU memory, training throughput and prior sampling are profiled separately.

On the intersection of completed primary plan cells, the equal-task mean normalized-return difference against occupancy (candidate minus occupancy, with the cross-task standard error) is -0.0000 ± 0.0007 for the conditioner-only variant (17 common cells), -0.0053 ± 0.0068 for the Gaussian (13), $+0.0089 \pm 0.0080$ for conditional RealNVP (11), -0.0128 ± 0.0135 for latent overshooting (12), and $+0.0051 \pm 0.0066$ for the invariant penalty (12). These descriptive contrasts condition on completion by both planners; their cross-task SEs do not measure training-seed uncertainty on the full task–seed grid. Against the native reference, four priority-suite cells have both planners complete (the three swimmer cells and quadruped walk seed 1103); the native model returns more on three of them (0.1312 and 0.1246 on swimmer seeds 1101–1102 and 0.1098 on quadruped, against 0.1109, 0.1188, 0.1041) and less on swimmer seed 1103 (0.1152 against 0.1211). The selected common cells do not establish a suite-wide planning comparison.

C.4.6 Native recurrent baseline

The offline DreamerV3 comparison (Hafner et al., 2025) retains the original CNN encoder/decoder, categorical recurrent state-space model (RSSM) with deterministic memory, KL balancing and free nats, reward and continuation heads, and world-model optimizer. We use `danijar/dreamerv3`, revision `e3f0224`, with offline window sampling and no actor–critic updates. The size12m configuration uses 64-frame training windows on the four priority tasks, with the same three training seeds. Its world model has 8.85 million parameters, against 3.26–3.29 million in the primary models; it trains for 25,000 updates of 16 windows in bfloat16, using the native optimizer at learning rate 4×10^{-5} with 1,000 warm-up updates and adaptive gradient clipping 0.3. Evaluation uses the final checkpoint. The primary fits use 100,000 updates of 16 windows of length 16, so both protocols observe exactly 25.6 million frames per model; wall time is not equated. Table 4 reports the corresponding model sizes and measured outcomes.

The native image head uses the original bounded MSE objective. For common observation scoring we explicitly evaluate a Gaussian of standard deviation 0.1 around its predicted image mean; this external score is distinct from the native training objective. Outgoing rewards are shifted to the incoming state convention, the first unobserved reward is masked, and shard boundaries are not labeled as environment terminations.

Native forecasts and CEM use the same physical actions and evaluation budgets; Table 4 reports the measured comparison. Under the frame-conditioned behaviour policy, native long rollouts decode each step to obtain the next action context. This cost differs from the preserving model's direct innovation rollout. Native rollouts start from filtered data and extend through 10^3 steps, the common comparison horizon with the primary models; there is no Gaussian latent reference, static-limit identity, or convergence certificate for that baseline.

C.4.7 DETAILED RECURRENT AND NATIVE-MODEL RESULTS

Figure 5 reports paired survival contrasts.

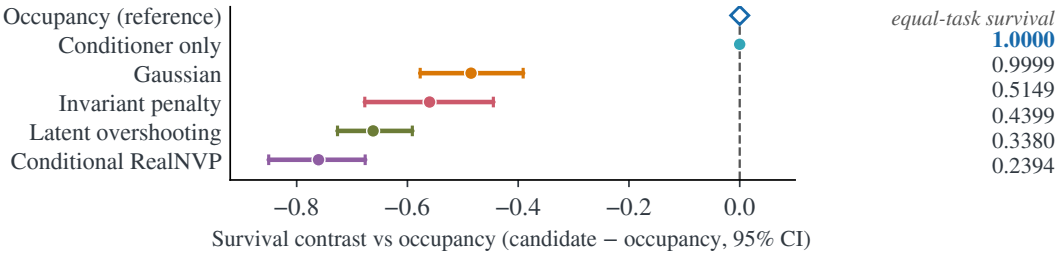


Figure 5: **All four nonpreserving arms lose chains at 10^5 steps.** Paired survival contrasts (candidate minus occupancy), with nominal 95% paired training-seed percentile-bootstrap intervals over twelve tasks and three seeds (10,000 draws, fixed task suite). The right column gives equal-task survival. Occupancy retains every evaluated chain.

Recurrent evaluation coverage and uncertainty. The primary comparison contains 216 fitted task/seed/arm checkpoints; 168 have at least one unavailable estimate because a rollout is censored or planning fails. Of 648 primary evaluation stages, 545 complete and 103 have deterministic planning failures. The corresponding counts are 178/216 completed stages for ablations, 12/24 for the planning sweep, and 89/108 for the selected gain, with 38, 12, and 19 planning failures, respectively. There are 172 failures in all. Each primary, ablation, and selected-gain checkpoint has a short-horizon, long-horizon, and planning stage; the sweep contains 24 task–arm–seed cells. The protocol does not estimate unconditional distances from censored rollouts or full-grid paired returns when a required planner fails.

Gaussian minus occupancy one-step observation NLL is 1.017 nats per pixel channel (paired training-seed interval $[-1.324, 3.413]$). Returns are divided by the planned physics-step budget, including early termination. Uncertainty uses training seeds as replicates within the fixed task suite; rollout repetitions are averaged within each checkpoint.

The validation-selected gain 1.001 leaves survival unchanged (contrast 0, nominal interval $[0, 0]$) and changes the marginal gap by -0.00016 (nominal interval $[-0.00029, -0.00004]$). The prespecified paired comparisons use this selected gain; the exploratory 1.003 control appears in Appendix C.4.4.

Figure 6 summarizes the fitted-grid measurements.

Table 4: Native offline DreamerV3 reference on 4 eligible / 4 registered priority tasks. Values are equal-task means ± training-seed SE. NLL is a common external Gaussian image score, not the native training likelihood. Return uses the planned physics-step budget. Observed-frame exposure is matched at 25.6M frames per model; the native reference is granted the larger capacity and context (8.85M vs. 3.26–3.29M parameters, 64-frame vs. 16-frame training windows) at fewer updates (25,000 vs. 100,000); wall time is not equated. Bold marks the lowest measured NLL and MSE means.

Model	Parameters	Obs. NLL ↓	10-step MSE ↓	Return
Native DreamerV3 offline	8.85M	$\mathbf{-1.2990 \pm 0.0008}$	$\mathbf{0.0024 \pm 0.00004}$	0.0734 ± 0.0013
Occupancy	3.26–3.29M	-0.0861 ± 0.4763	0.1206 ± 0.0244	failed
Gaussian	3.26–3.29M	1.3022 ± 2.5705	0.0542 ± 0.0514	failed

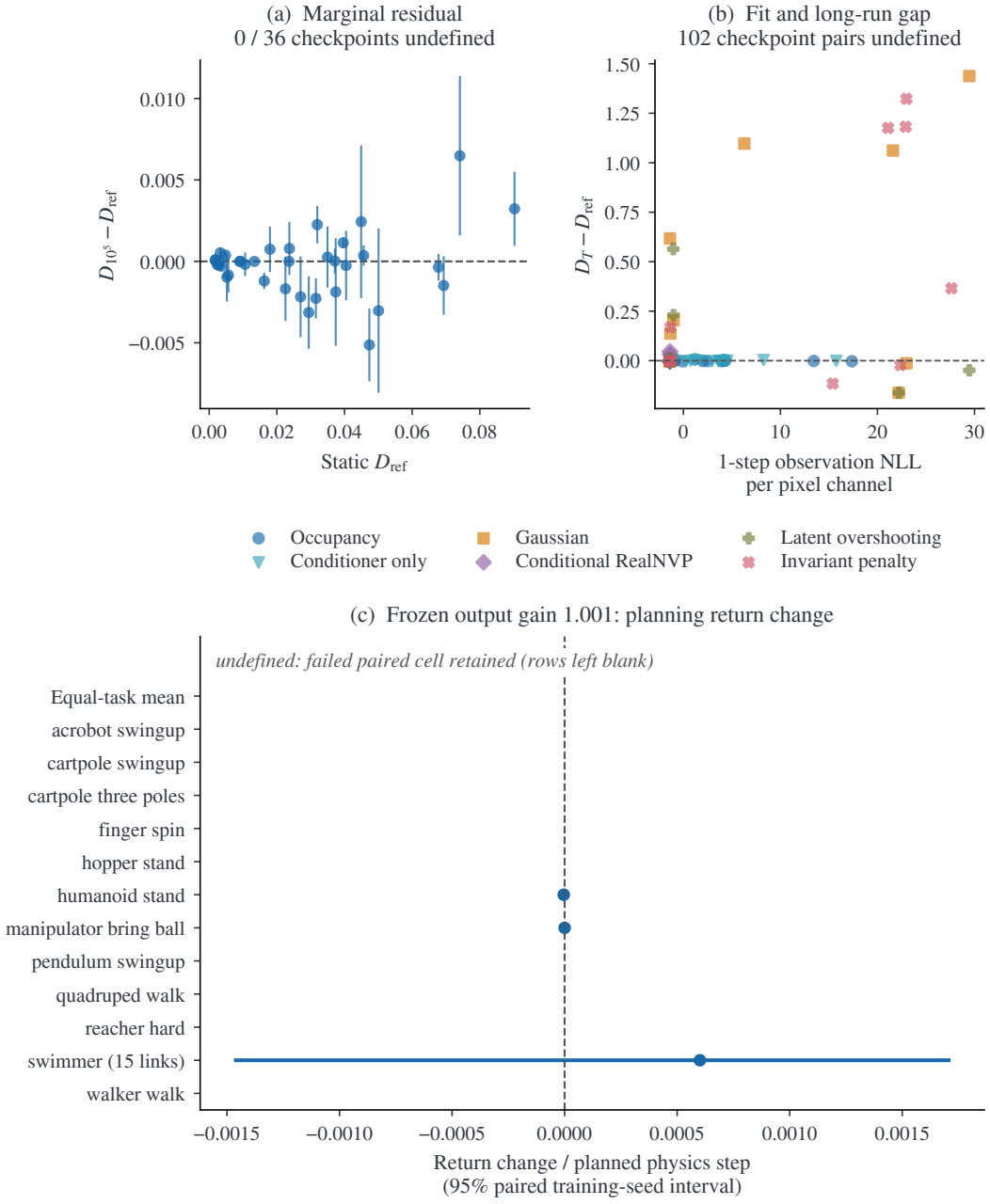


Figure 6: **Marginal agreement, conditional fit, and planning are distinct outcomes.** **(a)** The 10^5-step rollout-minus-static energy gap versus the static score for occupancy checkpoints; vertical bars show ± 1 paired evaluation SE of that gap; horizontal uncertainty is not shown. **(b)** Observation NLL versus marginal gap for cells with uncensored estimates; 102 censored cells are excluded from this scatter and retained in survival summaries. **(c)** Validation-selected gain return contrasts with nominal paired training-seed percentile-bootstrap intervals; failed comparisons remain unestimated. D_{ref} is a limiting level for preserving models under their convergence assumptions.

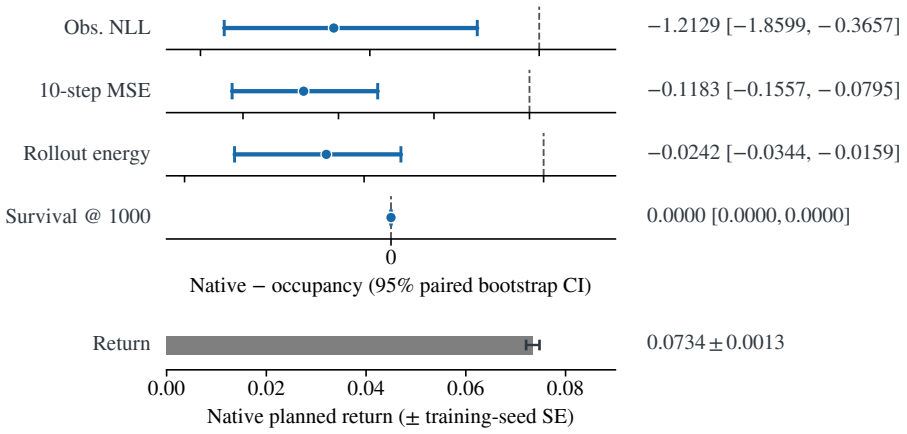
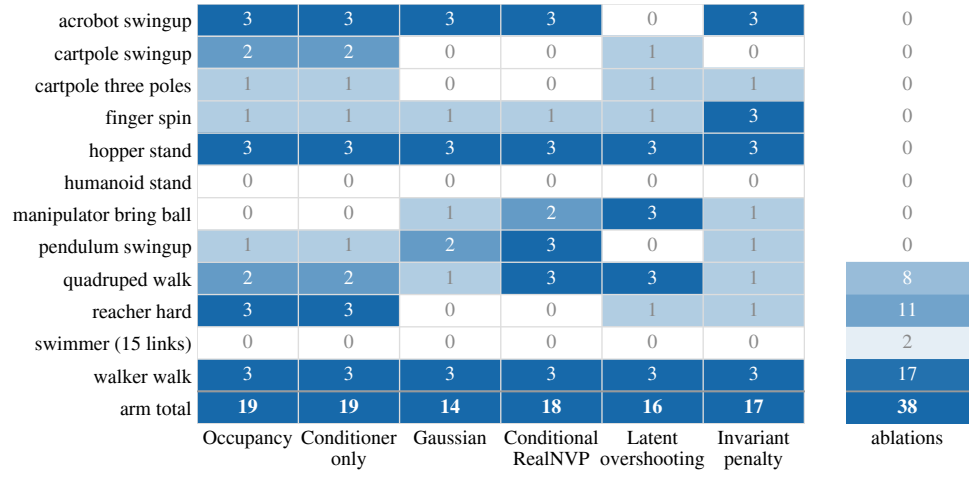


Figure 7: **The native DreamerV3 reference is the stronger short-horizon predictor; survival is tied at the shared evaluation horizon.** Native − occupancy contrasts with nominal 95% paired training-seed percentile-bootstrap intervals (4 priority tasks × 3 seeds, horizon 1000): observation NLL, 10-step MSE, and rollout energy all favor the native model with intervals excluding zero, while survival is identical (1.0 vs. 1.0). Bar: native planned return 0.0734 ± 0.0013 with all 12 plan cells completed (a full-grid occupancy return is unavailable because of planner failures, Fig. 8). The native 10^5-step marginal is unmeasured, and no Gaussian-reference identity exists for its categorical state.



Primary cells: 3 seeds each. Gain 1.001: 19 failures, in the same cells as occupancy.
Horizon sweep: 12 failures (occupancy 8, Gaussian 4). Total: 172 failures.

Figure 8: **Planner completion varies across tasks and transition families.** Each cell counts failures among three training seeds: the CEM planner has fewer finite candidates than required elites. The margin strip counts 38 ablation failures. The selected gain shares the same 19 failed task–seed cells as occupancy, and 12 planning-sweep jobs fail, giving 172 failures among 348 planning evaluations. Primary failure counts range from $14/36$ for Gaussian to $19/36$ for occupancy and conditioner only.

Per-task results. Tables 5–7 summarize primary tasks, retrained variants, and planning horizons. The supplementary `recurrent_checkpoints.csv` retains all 324 primary, selected-gain, and ablation records; `recurrent_planning_sweep.csv` retains 48 horizon records from 24 sweep jobs. The accompanying field definitions specify units and uncertainty. A *failed* return denotes a planner failure; a *censored* gap denotes a rollout with escaped chains. Completed-cell summaries use only measured returns.

Table 5: **Task-level recurrent outcomes.** Each cell lists mean survival at 10^5 steps, mean marginal gap $D_{10^5} - D_{\mathrm{ref}}$, and completed planners out of three seeds (top to bottom). Gap means require all three seeds; cens. denotes chain censoring. The gap uses RMS pixel distance. Exact checkpoint scores and paired evaluation SEs: `recurrent_checkpoints.csv`.

Task	Occupancy	Conditioner only	Gaussian	Conditional RealNVP	Latent overshooting	Invariant penalty
	1.00000	1.00000	1.00000	0.66667	1.00000	1.00000
	−0.0015	−0.0018	−0.0011	*cens.*	+0.0060	−0.0017
Acrobot swingup	0/3	0/3	0/3	0/3	3/3	0/3
	1.00000	1.00000	0.00000	0.00045	0.00000	0.00590
	−0.0023	−0.0028	*cens.*	*cens.*	*cens.*	*cens.*
Cartpole swingup	1/3	1/3	3/3	3/3	2/3	3/3
	1.00000	1.00000	0.04183	0.00000	0.04228	0.33830
	−0.0006	−0.0004	*cens.*	*cens.*	*cens.*	*cens.*
Cartpole three poles	2/3	2/3	3/3	3/3	2/3	2/3
	1.00000	1.00000	0.33333	0.62048	0.00000	0.61182
	−0.0007	−0.0006	*cens.*	*cens.*	*cens.*	*cens.*
Finger spin	2/3	2/3	2/3	2/3	2/3	0/3
	1.00000	1.00000	1.00000	0.66659	0.33333	0.66667
	+0.0005	−0.0003	+0.6653	*cens.*	*cens.*	*cens.*
Hopper stand	0/3	0/3	0/3	0/3	0/3	0/3
	1.00000	1.00000	0.66463	0.00000	0.00000	0.33931
	+0.0001	+0.0001	*cens.*	*cens.*	*cens.*	*cens.*
Humanoid stand	3/3	3/3	3/3	3/3	3/3	3/3
	1.00000	1.00000	0.33333	0.09753	0.00000	0.33333
	+0.0000	+0.0001	*cens.*	*cens.*	*cens.*	*cens.*
Manipulator bring ball	3/3	3/3	2/3	1/3	0/3	2/3
	1.00000	1.00000	1.00000	0.66667	1.00000	1.00000
	−0.0010	−0.0001	−0.0027	*cens.*	−0.0027	−0.0027
Pendulum swingup	2/3	2/3	1/3	0/3	3/3	2/3
	1.00000	1.00000	0.55558	0.02747	0.33341	0.00073
	+0.0018	+0.0011	*cens.*	*cens.*	*cens.*	*cens.*
Quadruped walk	1/3	1/3	2/3	0/3	0/3	2/3
	1.00000	1.00000	0.00000	0.04008	0.00000	0.33333
	+0.0010	+0.0015	*cens.*	*cens.*	*cens.*	*cens.*
Reacher hard	0/3	0/3	3/3	3/3	2/3	2/3
	1.00000	0.99915	0.98295	0.00000	0.99996	0.42175
	+0.0005	*cens.*	*cens.*	*cens.*	*cens.*	*cens.*
Swimmer (15 links)	3/3	3/3	3/3	3/3	3/3	3/3
	1.00000	1.00000	0.26685	0.08728	0.34717	0.22815
	+0.0001	+0.0006	*cens.*	*cens.*	*cens.*	*cens.*
Walker walk	0/3	0/3	0/3	0/3	0/3	0/3

Table 6: **Retrained architectural variants.** Four tasks (Walker, Reacher, Swimmer, Quadruped), three seeds each. All variants retain every chain; gaps are equal-task means ± training-seed SE. Full NLL, MSE, gap, and return records: `recurrent_checkpoints.csv`; aggregate SEs: `recurrent_method_summary.csv`.

Variant	Surviving cells	Marginal gap	Completed planners
$c_\star = 0.99$	12/12	$+0.00044 \pm 0.00065$	4/12
$c_\star = 0.9999$	12/12	$+0.00047 \pm 0.00071$	4/12
$r = 0$	12/12	-0.00025 ± 0.00035	8/12
$r = 0.01$	12/12	$+0.00042 \pm 0.00075$	3/12
$d = 512$	12/12	-0.00052 ± 0.00055	6/12
$d = 1024$	12/12	$+0.00035 \pm 0.00032$	9/12

Table 7: **Planning-horizon sweep.** Counts are completed seeds out of three; a mean return is shown only when all three complete. Returns use the planned physics-step denominator. The 48 seed–horizon rows in `recurrent_planning_sweep.csv` represent 24 sweep jobs. Twelve jobs fail during $H = 5$; their $H = 50$ evaluations are unattempted.

		Completed		Mean return	
Task	Model	$H = 5$	$H = 50$	$H = 5$	$H = 50$
Quadruped walk	Occupancy	1/3	1/3	—	—
Quadruped walk	Gaussian	2/3	2/3	—	—
Reacher hard	Occupancy	0/3	0/3	—	—
Reacher hard	Gaussian	3/3	3/3	0.0293	0.0065
Swimmer (15 links)	Occupancy	3/3	3/3	0.1227	0.1174
Swimmer (15 links)	Gaussian	3/3	3/3	0.1184	0.1186
Walker walk	Occupancy	0/3	0/3	—	—
Walker walk	Gaussian	0/3	0/3	—	—

C.4.8 Scope of the recurrent comparisons

The comparisons condition on the collected data, eligible tasks, fixed fitted behaviour policy, three paired seeds, and reported budgets. Because the matching gates reject every alternative transition family, their contrasts combine architecture with achieved fit; the selected gain isolates output scaling. Planning contrasts use common completed cells, with completion counts reported alongside returns. The prior certificate applies to the specified initialization bank; for second moment 6.9×10^9, its total-variation cap becomes informative near 10^4 steps.

D Lorenz calibration and competing stabilization objectives

A fitted encoder determines its own decoded-reference distribution. The first comparison asks whether that static score predicts the same model's long-run distributional error. It does not assume that all encoders have the same plateau or that a predictable plateau is small. Throughout the appendix, D_∞ is the static decoded-reference score of Section 4; the recurrent-study tables print the same quantity as D_{ref}.

Table 8: **Lorenz prediction–rollout comparison by training seed.** Static and terminal energy distances use mean ± Monte Carlo SE from eight repetitions, with sample SD. The standardized gap combines these SEs in quadrature; it is nominal because the saved summaries do not retain their covariance. The final column is an energy distance in Gaussian coordinates, not a TV bound.

Seed	Static D_∞	Rollout D_{10^5}	Nominal gap/SE	$D(\hat{\mu}^z_{10^5}, \gamma)$
0	0.1963 ± 0.0098	0.1849 ± 0.0087	-0.87	0.00215
1	0.2228 ± 0.0133	0.2076 ± 0.0112	-0.87	0.00069
2	0.1548 ± 0.0087	0.1593 ± 0.0084	$+0.37$	0.00111

Table 9: **Lorenz rollout fidelity, survival, and scale.** Three training seeds and seven sampled horizons. D_T is energy distance to data; f_T is the fraction of finite chains; $m_T = d^{-1}\mathbb{E}\|z_T\|^2$ uses finite chains. All stationary-model chains are finite. Unconstrained distances are conditional on survival, and extreme values can be numerically unreliable.

		Stationary			Unconstrained		
Seed	T	D_T	f_T	m_T	D_T	f_T	m_T
0	0	−0.0018	1.0000	0.8057	−0.0017	1.0000	1.0110
	1	−0.0007	1.0000	0.8191	−0.0022	1.0000	1.0161
	10	0.0616	1.0000	0.8712	−0.0025	1.0000	1.0048
	100	0.1789	1.0000	0.9926	9.25×10^{17}	0.9984	—
	1000	0.1851	1.0000	1.0035	1.17×10^{10}	0.9801	—
	10000	0.1775	1.0000	1.0027	1.07×10^{29}	0.8254	—
	100000	0.1849	1.0000	0.9965	−0.0035	0.1395	2.04×10^{9}
1	0	0.0034	1.0000	0.8566	0.0032	1.0000	1.0137
	1	0.0071	1.0000	0.8694	0.0025	1.0000	1.0125
	10	0.1740	1.0000	0.9535	−0.0030	1.0000	1.0090
	100	0.2129	1.0000	1.0271	0.0025	0.9999	1.0069
	1000	0.2158	1.0000	1.0228	0.0077	0.9994	1.0031
	10000	0.2113	1.0000	1.0252	0.0047	0.9932	1.0136
	100000	0.2076	1.0000	1.0322	0.0026	0.9394	1.0172
2	0	0.0012	1.0000	0.6588	0.0012	1.0000	1.0096
	1	0.0028	1.0000	0.6685	0.0008	1.0000	1.0061
	10	0.0538	1.0000	0.7216	−0.0007	1.0000	1.0055
	100	0.1538	1.0000	0.9676	3.68×10^{13}	0.9993	—
	1000	0.1634	1.0000	1.0113	1.11×10^{9}	0.9925	—
	10000	0.1588	1.0000	1.0040	3.54×10^{9}	0.9236	—
	100000	0.1593	1.0000	1.0100	−0.0020	0.4602	9.53×10^{15}

Table 10: **Static reference scores of the unconstrained Lorenz models.** Mean ± Monte Carlo SE over eight static draws. These scores compare decoded Gaussian samples with data; the unconstrained rollout need not converge to that reference law.

Seed	$D(E_{\#}^{-1}\gamma, p_{\mathrm{data}})$
0	0.0464 ± 0.0030
1	0.0308 ± 0.0034
2	0.0226 ± 0.0042

The three stationary models have different static distances, and each terminal score is within one nominal combined Monte Carlo SE of its own prediction. The unconstrained models have lower static reference scores but retain 13.95%, 93.94%, and 46.02% of their chains at 10^5 steps. Their static decoded-Gaussian scores are not predictions of their limiting laws. In particular, near-zero survivor-conditioned distances at the last horizon coexist with numerical failure and, in two seeds, extreme moments. These comparisons do not establish that an unconstrained kernel lacks an invariant distribution.

D.1 Reweighting the marginal fit

This comparison tests whether improving the fitted marginal changes the rollout plateau. The four constrained settings share data, schedule, and a 30,000-update budget; the $w = \infty$ setting fits the marginal first and then fits the conditional with the chart frozen. Two unconstrained conditional-flow arms use $w = 1, 10$. Each static score is evaluated from the same fitted checkpoint as its rollout.

Table 11: **Complete marginal-weight sweep.** Finite weights share a 30,000-update budget; $w = \infty$ uses marginal-first training. NLLs are nats per observation dimension. Static D_∞ uses decoded Gaussian samples for every arm, without predicting an unconstrained limit. Distance SEs use eight Monte Carlo repetitions and sample SD. Terminal distances are omitted when chains are lost.

		Held-out NLL					
Class	w	Pair	Marginal	Conditional	Static D_∞	D_{10^5}	Finite
Stationary	1	2.2679	2.6477	-0.3799	0.1126 ± 0.0098	0.1066 ± 0.0070	1.0000
	3	2.3891	2.3382	$+0.0509$	0.0809 ± 0.0056	0.0881 ± 0.0107	1.0000
	10	2.4548	2.2713	$+0.1835$	0.0165 ± 0.0042	0.0162 ± 0.0035	1.0000
	∞	3.5045	2.2582	$+1.2463$	0.0072 ± 0.0026	0.0028 ± 0.0036	1.0000
Unconstrained	1	1.4906	2.3240	-0.8334	0.0532 ± 0.0031	—	0.0599
	10	1.4511	2.2648	-0.8137	0.0057 ± 0.0025	—	0.0641

For constrained $w = 1$ to 10, held-out pair NLL increases by 0.1869 nats per dimension while the static distance decreases. The corresponding unconstrained NLL changes by -0.0395. These are fitted, finite-budget outcomes; an ordering of globally optimized objectives need not order their held-out scores. Both unconstrained arms retain only about 6% of chains at 10^5. The negative distance estimate -2.26 is therefore not evidence of better full-distribution fidelity. The smallest constrained plateau is unresolved at the experiment's sampling resolution.

D.2 Soft stabilization objectives

This comparison measures fidelity and survival under three soft stabilization objectives. The three penalties share an unconstrained transition and use weights $\{0.01, 0.1, 1\}$, 8,000 Adam updates, and matched initializations at seed zero. They add an eight-step Huber rollout loss, a Jacobian-direction penalty, or a four-step differentiable energy-distance penalty to total three-dimensional pair NLL. The pair NLL entries below are consequently nats per pair, not nats per dimension. The stationary arm uses the same encoder specification and update budget with a different transition family.

Huber targets are detached encoder states; coordinates in the training penalty are clamped to $[-10^3, 10^3]$. The Jacobian probe uses two normalized transpose-Jacobian iterations on a random direction. It is neither an exact spectral norm nor an operator-contraction certificate. The measure penalty uses rollout samples and independent real states, following the motivation of invariant-measure regularization (Schiff et al., 2024). Checkpoints minimize recorded validation NLL; the seed table states the unequal evaluation intervals for seed zero and seeds one–two.

Table 12: **Long-horizon regularization on Lorenz-63.** Seed 0, 8,000 updates per configuration. The three penalties share the unconstrained architecture; OURS uses a different transition family. D_∞ is the static decoded-Gaussian reference score, not a predicted limit for the unconstrained models. Rollout distances use mean $\pm$ Monte Carlo SE over eight repetitions and are displayed only when every chain is finite.

Objective	w	Pair NLL	D_∞	Finite at 10^5	D_{10^5}
FREE	—	5.604	0.1108	0.00027	—
Multistep	0.01	5.836	0.1334	0.13644	—
	0.1	6.253	0.0997	1.00000	0.6776 ± 0.0200
	1	5.840	0.1572	0.95813	—
Jacobian probe	0.01	5.391	0.1055	0.00003	—
	0.1	5.599	0.0910	0.00000	—
	1	5.498	0.1551	0.00104	—
Measure penalty	0.01	6.410	0.2089	0.00000	—
	0.1	7.574	0.2170	1.00000	0.1058 ± 0.0142
	1	6.438	0.1594	0.43649	—
OURS	—	8.003	0.2039	1.00000	0.2125 ± 0.0091

Table 13: **Seed-specific outcomes for the selected multistep weight.** All runs use $w = 1$ for multistep and 8,000 updates. Checkpoints minimize validation NLL over the recorded evaluations: every 1,000 updates for seed 0 and every 4,000 for seeds 1–2, plus the final update. Survivor-conditioned distances are included descriptively without Monte Carlo SE; they do not score the complete rollout law.

Arm	Seed	Finite	D_∞	D_{10^5}	D_{10^5}/D_∞
Multistep	0	0.95813	0.1572	0.0224	0.143
	1	0.99991	0.2271	0.0282	0.124
	2	0.31396	0.1342	—	—
OURS	0	1.00000	0.2039	0.2125	1.042
	1	1.00000	0.1380	0.1444	1.046
	2	1.00000	0.0924	0.0836	0.905

Only two of the nine penalized seed-zero configurations retain at least 99% of chains. Both retain every chain: multistep weight 0.1 has distance 0.6776, and measure-penalty weight 0.1 has distance 0.1058, versus 0.2125 for the stationary model. The latter is a direct example of a soft penalty attaining better long-run fidelity. Its static reference score does not become a predicted limit by virtue of that success. At the selected multistep weight one, only one of three seeds retains at least 99% of chains; the stationary model retains every chain in all three. These results identify a survival–fidelity trade-off under the recorded budgets, without ruling out successful stabilization by a different regularizer or tuning procedure.

Time-varying negative control. On the ramped Lorenz system, stationary-model pair NLL is 4.7793 versus 3.5166 for the unconstrained model, in nats per dimension. Its rollout distance reaches 0.841 at 10^3 steps. The unconstrained model also fails: its second moment is 1.6×10^4 by step ten and becomes nonfinite later. The reference two-half distance is only 0.0052, showing that this simple finite-data diagnostic misses the imposed nonstationarity. A model's invariant marginal need not track a changing data marginal.

E FIXED-LATENT PIXEL MODELS: MARGINAL PREDICTABILITY AND CONDITIONAL FAILURE

Each task has a separately trained and frozen convolutional autoencoder. All transition families use its same 256-dimensional latent, Gaussian smoothing scale 0.1, decoder, and data split. Walker's held-out reconstruction PSNR is 30.25 dB. These are transition comparisons in a fixed representation, not end-to-end recurrent pixel-world-model comparisons. Learned invertible state charts, where present, are trained separately for each transition model.

The data contain 64 training and 32 held-out uninterrupted trajectories, yielding 959,936 and 479,968 pairs before burn-in. Burn-in is applied to both training pairs and reference observations. For Walker the latent screen selects 6,000 observation steps, leaving 575,936 training and 287,968 held-out pairs. Its matched 250-frame window baseline is 0.05408 ± 0.01691, with threshold 0.11908, half-to-half distance 0.03223, estimated ESS 1,141, spread ratio 0.986, and no frozen coordinate. Observation-step burn-in differs from the physical-state screen's simulator-step units. These diagnostics support the retained finite window; they do not verify a population invariant law.

Table 14: Walker transition budgets and held-out NLL, in nats per frozen-latent dimension. Parameter counts exclude the common autoencoder. A dash denotes an unevaluated or inapplicable likelihood component. Source: the recorded Walker comparison.

	Parameters (M)		NLL (nats/dim)		
Model	Coordinates	Transition	Marginal	Conditional	Pair
Stationary	4.20	7.20	-0.0307	-0.0732	-0.1039
Free flow	4.20	5.52	-0.0256	-0.1549	-0.1805
Gaussian with chart	4.20	8.37	$+0.1390$	$+0.1197$	$+0.2587$
Gaussian raw	—	8.37	—	$+0.4717$	$+0.4717$
Flow matching	—	8.47	—	—	—

The Walker comparison uses transition heads with 5.52–8.47M parameters; capacity and optimization differ across families, including conditional flow matching (Lipman et al., 2023). Native recurrent-model comparisons appear in Appendix C.4.6. The twelve-task comparison includes the stationary transition, raw Gaussian, and flow matching; the other two chart-based arms are additional Walker controls.

E.1 Complete twelve-task outcomes

The task-wide comparison separates prediction of each model's own limiting distance from accuracy relative to the data.

Table 15: Static and terminal latent energy distances on all twelve pixel tasks. Levels use the same eight paired Monte Carlo repetitions of each frozen checkpoint. The paired difference is terminal minus static, with sample SE. These uncertainties do not measure training-seed or collection variation. Every constrained chain remains finite.

Task	D_∞	D_{10^5}	Paired $\Delta \pm$ SE
Acrobot swingup	0.143	0.144	$+0.0008 \pm 0.0040$
Cartpole swingup	0.212	0.212	-0.0005 ± 0.0024
Cartpole three poles	0.116	0.116	$+0.0002 \pm 0.0016$
Finger spin	0.169	0.175	$+0.0065 \pm 0.0062$
Hopper stand	0.193	0.193	$+0.0007 \pm 0.0038$
Humanoid stand	0.023	0.023	$+0.0008 \pm 0.0016$
Manipulator bring ball	0.224	0.229	$+0.0049 \pm 0.0036$
Pendulum swingup	0.082	0.086	$+0.0037 \pm 0.0050$
Quadruped walk	0.114	0.114	-0.0009 ± 0.0042
Reacher hard	0.305	0.299	-0.0061 ± 0.0038
Swimmer (15 links)	1.100	1.101	$+0.0010 \pm 0.0132$
Walker walk	0.160	0.156	-0.0036 ± 0.0022

Table 16: **Twelve-task survival and pixel fidelity at $T = 10^5$.** Energy estimates use Euclidean distance on flattened, downsampled 16×16 decoder outputs with intensities in $[0, 1]$; this ground distance is not RMS-normalized. The flow-matching/ours ratio is reported only when both methods retain every chain. The paired static-versus-rollout comparison is given separately in Table 15.

	Finite-chain fraction			Pixel distance		
Task	Ours	Gauss. raw	FM	Ours	FM	FM/ours
Acrobot swingup	1.00	0.00	1.00	0.0106	0.0031	0.30
Cartpole swingup	1.00	0.00	1.00	0.0159	0.0010	0.06
Cartpole three poles	1.00	0.00	1.00	0.0022	0.0141	6.5
Finger spin	1.00	0.74	1.00	0.0153	0.0013	0.08
Hopper stand	1.00	0.95	1.00	0.0105	0.0007	0.07
Humanoid stand	1.00	0.00	1.00	0.0040	0.0084	2.1
Manipulator bring ball	1.00	0.00	0.43	0.0070	0.5780	—
Pendulum swingup	1.00	0.00	1.00	0.0028	0.0002	0.09
Quadruped walk	1.00	0.00	1.00	0.0154	0.0831	5.4
Reacher hard	1.00	0.98	0.23	0.0168	0.4449	—
Swimmer (15 links)	1.00	1.00	1.00	0.6031	1.6668	2.8
Walker walk	1.00	0.00	1.00	0.0277	2.2742	82

swimmer_swimmer15 carries the metastability flag from the occupancy screen and is retained with that qualification.

The stationary models retain every chain on all twelve tasks, but their pixel fidelity is not uniformly best. Of ten tasks on which flow matching also retains every chain, five favour flow matching and five favour the stationary model. The other two flow-matching scores condition on partial survival. Swimmer's large stationary distance and metastability qualification remain part of the result. Static-versus-rollout agreement in Table 15 therefore answers a different question from which model has the smallest pixel distance.

Sampling conventions. Rollouts reproduce the external OU driver's coefficients, clipping, and interleaved actions, initializing its state from the stationary driver law. Conditional flow matching uses 16 Euler steps. Its one-step 8-versus-16-step sensitivity reaches 14.6% of observed displacement on Swimmer, exceeding the 10% sensitivity heuristic; the reported scores use 16 Euler steps. The compact Walker comparison below separates stochastic samples from mean steps, which define different rollout laws.

Table 17: Walker decoded-pixel distance and numerical survival at 10^5 steps. Entries summarize the recorded full trajectory grid; distances for partially surviving models are conditional on survival. A dash means no surviving sample. Finite-sample uncertainty for the paired stationary plateau is reported separately in Table 15.

Transition	Rollout convention	Finite fraction	Pixel distance
Stationary	Sampling	1.0000	0.02770
Free flow	Sampling	0.6030	0.03458
Gaussian with chart	Sampling	0	—
Gaussian with chart	Mean step	0	—
Gaussian raw	Sampling	0	—
Gaussian raw	Mean step	0	—
Flow matching	Sampling	1.0000	2.27422
Flow matching	Mean step	1.0000	2.26613

At 10^3 steps, Walker mean-step flow matching and the stationary model have similar distances (0.02738 and 0.02778). By 10^5, both flow-matching conventions reach approximately 2.27. The sampled flow-matching second moment is 1.06, closer to the data's 1.01 than the stationary model's 0.80 despite its worse distributional distance: a matched moment alone does not identify the more faithful distribution.

E.2 TYPICAL CONDITIONAL FIT AND EXTREME LOSSES

The conditional-loss distribution distinguishes a broad fit failure from a rare extreme tail.

Conditional-likelihood distributions. Held-out pair costs distinguish poor fit on typical transitions from a small number of extreme losses. The tables report the empirical mean, median, maximum, and upper-tail fraction directly. Large observed losses do not establish infinite population moments. Pair-level standard errors would additionally require accounting for temporal dependence.

Table 18: **Conditional NLL across held-out pairs.** All costs are nats per latent dimension. The tail fraction is the empirical proportion exceeding ten nats per dimension. Each row compares the same held-out pairs under the constrained model and the raw-latent Gaussian head.

	Ours				Gaussian raw
Task	Median	Mean	Maximum	Fraction > 10	Median
Acrobot swingup	-0.5204	-0.4951	$+4.19 \times 10^{2}$	1×10^{-5}	$+0.6007$
Cartpole swingup	-0.4823	$+33.4751$	$+8.43 \times 10^{6}$	3×10^{-4}	-0.0620
Cartpole three poles	$+0.4501$	$+0.4126$	$+1.26 \times 10^{2}$	6×10^{-6}	$+0.9424$
Finger spin	-0.6457	$+0.5314$	$+3.88 \times 10^{5}$	2×10^{-5}	-0.5393
Hopper stand	-0.2634	-0.2357	$+3.03 \times 10^{2}$	4×10^{-5}	$+0.1258$
Humanoid stand	$+0.4159$	$+0.4170$	$+3.2921$	0	$+0.9886$
Manipulator bring ball	$+0.4458$	$+1.61 \times 10^{3}$	$+2.58 \times 10^{7}$	6×10^{-3}	$+0.5640$
Pendulum swingup	-0.7428	$+2.84 \times 10^{3}$	$+3.00 \times 10^{8}$	2×10^{-5}	-0.3783
Quadruped walk	-0.3659	$+4.07 \times 10^{3}$	$+6.50 \times 10^{7}$	8×10^{-3}	$+0.4521$
Reacher hard	$+0.8509$	$+1.11 \times 10^{4}$	$+1.42 \times 10^{8}$	4×10^{-3}	$+0.4041$
Swimmer (15 links)	$+1.43 \times 10^{7}$	$+2.67 \times 10^{7}$	$+2.94 \times 10^{8}$	0.97	$+0.2616$
Walker walk	-0.0797	-0.0732	$+1.4903$	0	$+0.4798$

Table 19: **Sensitivity to extreme pair losses.** The trimmed mean removes the largest 0.1% of costs and therefore changes the estimand. The 90% column gives the fraction of pairs whose largest costs account for 90% of the total, when defined. The final column is the Gaussian head's median cost on the constrained model's worst 64 pairs.

Task	Trimmed mean	Fraction for 90%	Marginal NLL	Gaussian on worst
Acrobot swingup	-0.4977	—	-0.4885	$+0.8250$
Cartpole swingup	-0.5002	6×10^{-6}	-0.4935	$+1.3204$
Cartpole three poles	$+0.4116$	0.73	$+0.4159$	$+1.0937$
Finger spin	-0.6325	3×10^{-6}	-0.6097	$+0.3829$
Hopper stand	-0.2397	—	-0.2151	$+0.6000$
Humanoid stand	$+0.4165$	0.83	$+0.5093$	$+1.1412$
Manipulator bring ball	$+2.4806$	2×10^{-4}	$+0.6216$	$+0.5382$
Pendulum swingup	-0.7399	1×10^{-5}	-0.7285	$+0.8168$
Quadruped walk	$+29.7949$	3×10^{-4}	-0.2371	-0.1430
Reacher hard	$+2.9868$	3×10^{-4}	$+0.9779$	$+0.0687$
Swimmer (15 links)	$+2.66 \times 10^{7}$	0.43	$+6.6530$	$+0.9427$
Walker walk	-0.0741	—	-0.0307	$+0.9733$

The Swimmer median of 1.43×10^{7} and 97% of pairs above ten nats per dimension identify broad conditional misfit. Reacher instead has a median near 0.85 with an extreme upper tail. The trimmed statistics in Table 19 summarize the retained pairs. The raw Gaussian assigns moderate costs to many pairs in the constrained model's worst tail, so those losses are specific to the fitted conditional model.

E.3 ALTERNATE OBSERVATION-SPACE DIAGNOSTICS

The frozen Walker checkpoints are evaluated in decoded 16×16 pixels, fixed random CNN features, standardized state-probe outputs, and frozen AE latents. All four metrics share trajectories within this evaluation, which uses separate rollout draws from the preceding tables. Random CNN features supply a nonlinear image geometry, not a validated perceptual metric. The state probe is a two-hidden-layer multilayer perceptron (MLP) of width 512 trained on up to 60,000 real validation observations and 18 recorded physical observables. Its $R^2 = 0.616$ is a checkpoint-selection score on the final ordered 20% of that pool, not an independent test score. Probe error, ambiguity, and failures on generated images limit the physical interpretation.

Table 20: **Alternate metrics on Walker walk.** Four energy distances on the same trajectories within this evaluation. Finite and in-box fractions are distinct; distances use the in-box subset. Values are means over up to four Monte Carlo repetitions at one checkpoint per arm, without training-seed uncertainty. The state column compares probe outputs with recorded standardized states, and its $T = 0$ value includes probe error.

Arm	T	Finite	In box	Pixels	Features	State probe	AE latent
OURS	0	1.0000	1.0000	0.0004	0.0001	0.0373	0.0014
	1	1.0000	1.0000	0.0011	0.0003	0.0787	0.0072
	10	1.0000	1.0000	0.0221	0.0038	0.1102	0.1261
	100	1.0000	1.0000	0.0301	0.0072	0.1190	0.1608
	1000	1.0000	1.0000	0.0300	0.0073	0.1221	0.1595
	10000	1.0000	1.0000	0.0285	0.0065	0.1186	0.1544
	100000	1.0000	1.0000	0.0290	0.0068	0.1176	0.1582
GAUSS_E	0	1.0000	1.0000	0.0004	0.0001	0.0373	0.0014
	1	1.0000	1.0000	0.0035	0.0011	0.0875	0.0140
	10	1.0000	0.6594	0.1105	0.0139	0.1499	1.2502
	100	0.0000	0.0000	—	—	—	—
	1000	0.0000	0.0000	—	—	—	—
	10000	0.0000	0.0000	—	—	—	—
	100000	0.0000	0.0000	—	—	—	—
FREE	0	1.0000	1.0000	0.0004	0.0001	0.0373	0.0014
	1	1.0000	1.0000	0.0028	0.0003	0.0404	0.0181
	10	1.0000	0.9998	0.0075	0.0007	0.0642	0.0274
	100	0.9979	0.9977	0.0116	0.0016	0.0721	0.0382
	1000	0.9932	0.9932	0.0129	0.0021	0.0728	0.0430
	10000	0.9492	0.9492	0.0141	0.0023	0.0758	0.0529
	100000	0.6082	0.6079	0.0211	0.0044	0.0750	0.0587

Table 21: **Monte Carlo variation for the stationary model under alternate metrics.** Mean $\pm$ SE from four evaluation repetitions, using sample SD. Every chain is finite and inside the box. These SEs describe evaluation sampling at the fixed checkpoint and probe.

T	Pixels	Features	State probe	AE latent
0	0.0004 ± 0.0004	0.0001 ± 0.0002	0.0373 ± 0.0022	0.0014 ± 0.0010
1	0.0011 ± 0.0007	0.0003 ± 0.0001	0.0787 ± 0.0033	0.0072 ± 0.0021
10	0.0221 ± 0.0006	0.0038 ± 0.0002	0.1102 ± 0.0048	0.1261 ± 0.0017
100	0.0301 ± 0.0007	0.0072 ± 0.0003	0.1190 ± 0.0026	0.1608 ± 0.0021
1000	0.0300 ± 0.0011	0.0073 ± 0.0003	0.1221 ± 0.0022	0.1595 ± 0.0030
10000	0.0285 ± 0.0014	0.0065 ± 0.0002	0.1186 ± 0.0009	0.1544 ± 0.0039
100000	0.0290 ± 0.0010	0.0068 ± 0.0001	0.1176 ± 0.0033	0.1582 ± 0.0044

These distances retain finite AE latents within a box whose center is the data midpoint and whose half-width is five times its coordinate range. The stationary arm keeps all chains; its late-distance ranges vary little under each of the four geometries. The free flow has lower survivor-conditioned distances in every geometry, while only 60.82% remain finite and 60.79% remain inside the box

at the final horizon. The probe distance 0.0373 already present at initialization is a diagnostic reference, not a subtractable physical error floor. All reported retained fractions remain essential to interpreting these alternative metrics.

F ARCHITECTURAL CONTRACTION, FIDELITY COST, AND RECURRENT MEMORY

The cap experiment uses the sequential block scan of Appendix I.2. Its uniform coefficient cap gives the L^2 operator bound of Theorem 47. For finite initial density-ratio L^2 error, the sufficient horizon for a 100-fold relative reduction is $\lceil \log(10^{-2}) / \log c_\star \rceil$. This gives 44, 459, 4,603, and 46,050 steps for caps $0.9, 0.99, 0.999, 0.9999$. The default numerical noise floor $\sin \phi_{\min} = 10^{-4}$ gives approximately 9.2×10^8 steps. These relative guarantees are not absolute pixel-error tolerances and do not apply directly to an atomic initial law. The parallel finite-start argument in Section 4 addresses the latter issue with a different construction.

Table 22: **Angle-capped transitions on twelve pixel tasks.** All models use $c_\star = 0.999$, giving a 100-fold contraction of initial L^2 error within $4{,}603$ steps when that error is finite. D_∞ and D_{10^5} use the same eight paired evaluation repetitions; t is their mean difference divided by its Monte Carlo standard error, computed with sample variance. $\widehat{\sigma}_2$ is a finite-basis estimate, not a certified lower endpoint. ΔNLL compares with the frozen default-cap checkpoint. Dashes indicate unavailable rate estimates.

Task	$\widehat{\sigma}_2$	D_∞	D_{10^5}	Paired t	ΔNLL
`acrobot_swingup`	—	0.146	0.144	-0.51	-0.0034
`cartpole_swingup`	—	0.211	0.207	-1.28	-33.9489
`cartpole_three_poles`	0.9613	0.119	0.117	-0.87	-0.0009
`finger_spin`	0.9773	0.180	0.182	$+0.68$	-1.1688
`hopper_stand`	0.9591	0.189	0.192	$+0.75$	-0.0006
`humanoid_stand`	0.9823	0.049	0.046	-1.03	-0.0029
`manipulator_bring_ball`	0.9847	0.232	0.230	-0.45	-1.6×10^3
`pendulum_swingup`	0.9815	0.204	0.212	$+0.51$	-2.8×10^3
`quadruped_walk`	0.9906	0.084	0.087	$+1.54$	-4.1×10^3
`reacher_hard`	0.9597	0.309	0.301	-5.28	-1.1×10^4
`swimmer_swimmer15`	0.9928	1.105	1.123	$+0.72$	-2.7×10^7
`walker_walk`	0.9816	0.170	0.166	-1.85	$+0.0000$

All twelve capped runs pass the specified flatness check between 10^4 and 10^5: the difference is within three combined Monte Carlo SEs plus $0.01 D_\infty$. Both horizons exceed 4,603, so that observation does not locate convergence or verify completion by the theoretical horizon. Eleven of the twelve static-versus-terminal paired gaps have $|t| < 3$. Reacher's is -0.0079 ± 0.0015 ($t = -5.28$); the static resampling of Appendix G and the independent rollout replication of Appendix G.1.1 are consistent with variation in the original static draw.

Table 23: **Contraction horizon and fitted-model trade-offs.** $T^\star = \lceil \log(10^{-2})/\log c_\star \rceil$ bounds the horizon for a 100-fold reduction of finite initial L^2 error. Default refers to the numerical noise floor $\sin\phi_{\min} = 10^{-4}$, which gives a finite but impractical bound of approximately 9.2×10^8 steps. ΔNLL is relative to the frozen default checkpoint for that task; these comparisons include variation between fitted checkpoints.

Task	$c_\star$	$T^\star$	D_∞	ΔNLL
pendulum_swingup	0.9	44	0.128	−2842.147
pendulum_swingup	0.99	459	0.108	−2842.134
pendulum_swingup	0.999	4603	0.211	−2842.084
pendulum_swingup	0.9999	46050	0.203	−2841.520
pendulum_swingup	Default	9.2×10^8	0.076	+0.000
reacher_hard	0.9	44	0.337	−11114.228
reacher_hard	0.99	459	0.322	−11114.865
reacher_hard	0.999	4603	0.310	−11114.873
reacher_hard	0.9999	46050	0.298	−11114.307
reacher_hard	Default	9.2×10^8	0.306	+0.000
quadruped_walk	0.9	44	0.091	−4069.268
quadruped_walk	0.99	459	0.108	−4068.519
quadruped_walk	0.999	4603	0.084	−4064.337
quadruped_walk	0.9999	46050	0.092	−4009.974
quadruped_walk	Default	9.2×10^8	0.118	+0.000

Tighter caps shorten the relative-contraction bound but do not order fitted fidelity or held-out likelihood. Moving from 0.999 to 0.9 improves conditional NLL on Pendulum and Quadruped while worsening Reacher by about 0.65 nats per dimension. Default-cap retraining still gives very poor pair NLL on Pendulum and Quadruped ($9,123$ and $25,438$, compared with frozen values $2,841$ and $4,069$). Thus the extreme default losses persist under these additional fits. Static distances also change: the best capped level is worse than the matched default on Pendulum and Reacher, and better on Quadruped. The shortest bound and smallest static discrepancy occur at different cap settings.

On matched Lorenz seeds, cap 0.999 increases pair NLL by 0.075–0.162, and cap 0.99 by 0.376–0.503. Enforcing a noise floor can therefore cost accuracy on sharp conditionals while reducing extreme held-out log-loss on other tasks.

F.1 Deterministic memory and partial noise

The synthetic memory comparison has $d_h = d_z = 8$ and trains six configurations for 4,000 updates. Rotation-only memory conserves $\|h\|^2$ pathwise; quantile swirls remove that radius invariant but do not by themselves supply strict one-step contraction. A cap on the noisy z block does not certify the whole (h, z) state when deterministic memory remains. The recurrent pixel model of Appendix C.4 instead includes every persistent coordinate in its stochastic prior.

The initial memory variance is 1.44, with the other coordinates standard Gaussian, giving finite divergence

$$\chi^2(\mu_0\|\gamma) = \{1.44(2 - 1.44)\}^{-d_h/2} - 1 = 1.365.$$

Finite divergence alone does not imply contraction.

Table 24: **Relaxation with finite initial divergence.** All models have $d_h = d_z = 8$ and start with memory variance 1.44 ($\chi^2 = 1.365$). The operator column is a sampled estimate, not a bound. $c_\star$ is the architectural upper bound, and T_{rel} is sufficient for a 100-fold reduction of L^2 error. The default noise floor gives $c_\star = \sqrt{1 - 10^{-8}}$. Distances at $T = 1000$ are normalized by a single independent-Gaussian control distance; this ratio has no calibrated significance threshold.

Memory update	Cap	$\hat{\sigma}$	T_{rel}	$\mathbb{E}\|h_T\|^2$	D_T/D_{ctrl}
Joint update	Default	0.974	9.2×10^8	8.01	1.24
Rotation	1	1.002	—	11.51	11.64
Quantile swirl	1	1.016	—	11.45	11.66
Stochastic memory	Default	0.973	9.2×10^8	8.02	1.09
Rotation	1 (requested .99)	1.002	—	11.51	12.45
Stochastic memory	0.99	0.973	459	8.02	0.99

Rotation memory retains second moment 11.51 instead of eight. The quantile swirl remains near that value at 1,000 steps, a finite-horizon non-relaxation rather than evidence about asymptotic mixing. Both stochastic-memory models reach second moments near eight and distances near the single Gaussian control. With cap 0.99 on every updated block, the relative horizon is 459 and the separate absolute bound in the paper's L^1-TV convention is $\|\mu_0 K^T - \gamma\|_{\mathrm{TV}} \le 0.99^T\sqrt{1.365} \le 0.01$ for $T \ge 474$ (total variation, not the energy statistic of Table 24). The stochastic-memory pair has NLL 0.4130 with the cap and 0.4153 with the default floor — one fitted pair, not a general cost estimate; deterministic-memory scores concern fewer coordinates and are not directly comparable to that pair.

G Static-reference and finite-start diagnostics

G.1 Independent static resampling at a fixed Reacher checkpoint

The static estimate in 256-dimensional frozen-autoencoder coordinates, using unnormalized Euclidean distance, is resampled at the capped Reacher checkpoint independently: eight fresh Gaussian clouds of 2,048 points per reference pool, averaged within pool, giving eight reference-level values for uncertainty estimation. The 64 clouds are not treated as 64 independent replications of the reference data or model; the recorded terminal rollout samples are reused.

Table 25: Independent static resampling for the capped Reacher checkpoint. Eight fresh Gaussian clouds of 2,048 points are averaged within each of eight existing reference pools. Mean ± sample SE uses the eight reference-level values. Recorded rollout samples are reused; Appendix G.1.1 reports the independent rollout replication.

Quantity	Energy distance
Fresh static distance	0.30264 ± 0.00141
Stored static minus fresh static	0.00595 ± 0.00235
Stored data-start rollout minus fresh static	-0.00199 ± 0.00190
Stored reference-start rollout minus fresh static	-0.00041 ± 0.00250

The resampled static average sits below the recorded static draw, and both recorded rollout laws are closer to the resampled average than to the recorded draw, with unresolved residual contrasts (Table 25). Scoring the same Gaussian sample with float32 and float64 decoder arithmetic changes this static distance by about 9×10^{-9}, six orders of magnitude below the sampling contrast. An independent replication with 32 fresh 10^5-step rollout pairs also gives a zero-compatible gap (Appendix G.1.1). Together, the results are consistent with sampling variation in the original static draw.

G.1.1 Independent Reacher rollout replication

The capped Reacher checkpoint is evaluated with 32 fresh paired repetitions of 2,048 points each at $T = 100000$. The data-initialized gap $D_T - D_{\mathrm{ref}}$ is 0.001711 ± 0.002227, with interval

$[-0.002831, 0.006252]$ and cross/within sample covariance -0.007606. The Gaussian-initialized gap $D_T^\gamma - D_{\mathrm{ref}}$ is 0.0008676 ± 0.002131, with interval $[-0.003478, 0.005213]$ and cross/within sample covariance -0.005885. The initialization contrast $D_T - D_T^\gamma$ is 0.000843 ± 0.002063, with interval $[-0.003365, 0.005051]$ and cross/within sample covariance -0.008885. Uncertainties are sample SEs with 95% Student intervals, paired per repetition.

Action, initialization, Gaussian-start, transition, reference, and static streams use explicit seeds. Disjointness from the original evaluation is verified for all except the transition stream, whose original seed is unavailable. The paired uncertainty conditions on this one checkpoint and the fixed held-out data pool. This evaluation uses frozen-autoencoder coordinates and clipped OU actions; it measures rollout-sampling variation without replicating training.

G.2 Reference mixtures bound losses without restoring dynamics

This rescoring measures how much of the conditional density is supplied by the reference component when a mixture controls extreme loss. For frozen checkpoint density p and its decoded-reference density p_E, the post-hoc mixture

$$p_r(x' \mid x, a) = (1 - r)p(x' \mid x, a) + rp_E(x')$$

retains reference preservation when p has it. It bounds per-dimension NLL above by $[-\log p_E(x') - \log r]/d$, and increases the original NLL by at most $-\log(1 - r)/d$ pointwise. For $r = 10^{-3}$ and $d = 256$, the latter is 3.9082×10^{-6}. The rescoring evaluates all validation pairs at this common reset probability, without retraining or new rollouts.

Table 26: Post-hoc reset-mixture rescoring on all held-out pairs: Reacher 479,968; Swimmer 247,968. Costs are nats per latent dimension ($d = 256$). The common reset probability is $r = 10^{-3}$. Ref. dom. is the percentage of pairs for which the reference component has posterior mixture responsibility above one half; it is not the reset sampling frequency. These CPU evaluations include no retraining or new rollout.

Checkpoint	Mean	Median	99.9%	Maximum	% > 10	Ref. dom. %
Reacher, default	1.118×10^4	0.8509	1.211×10^4	1.415×10^8	0.3665	—
Reacher, default $+r$	0.8547	0.8433	2.9334	4.6512	0.0000	15.82
Reacher, cap .999	0.8389	0.8300	2.5030	192.7131	0.0002	—
Reacher, cap .999 $+r$	0.8328	0.8265	2.4403	4.0283	0.0000	8.76
Swimmer, default	2.674×10^7	1.433×10^7	1.887×10^8	2.942×10^8	97.1081	—
Swimmer, default $+r$	6.6779	6.7025	16.3152	22.0964	12.6641	99.15
Swimmer, cap .999	314.0148	294.4076	1175.8230	1987.7783	94.7945	—
Swimmer, cap .999 $+r$	6.7789	6.7911	16.4691	22.0765	14.7499	98.26

Reacher's default extreme losses become bounded by the reference component; the median changes little. Its worst-pair decomposition identifies very small conditional scales and large standardized residuals as the main source of the loss, rather than the state-chart Jacobian. The corresponding capped checkpoint is much better conditioned on those pairs. The CPU and GPU mean scores differ by approximately 0.6% for the uncapped Reacher model; both reveal the same severe conditional tail.

For Swimmer, the reference component has posterior density responsibility above one half on 99.15% of default-checkpoint pairs and 98.26% of capped-checkpoint pairs; its sampling probability remains 0.1%. The guarded default median remains about 6.70 nats per dimension, and 12.66% of pairs remain above ten. The reduction in scored log-loss is supplied mainly by the static reference component.

G.3 Untrained parallel kernels and atomic starts

The parallel-kernel CPU experiment uses random, untrained weights. It checks Gaussian starts, a fixed uniform 16-point bank in eight dimensions, and exact conditional tanh-Gaussian policy shifts. The bank's initial KL to the Gaussian is infinite, while its saved second moment is finite. The implemented bound therefore begins after the first noisy transition. Bounded-observable intervals use

Hoeffding's inequality with a union bound over at most 200 comparisons and family failure probability 0.01; the independent unit is a rollout chain at a fixed horizon. All 114 reported comparisons are compatible with their bounds. These finite-sample checks test implementation consistency.

Table 27: Finite-start bounds in a two-dimensional OU specialization. The initial law is uniform on $\{(-3,0),(3,0)\}$, so its initial KL is infinite. The update coefficient is 0.75 and the architectural cap used in the bound is 0.8; no reset is used. KL and L^1-TV are computed by 220×220 Gauss–Legendre quadrature on $[-12,12]^2$. These are numerical integrals of an analytic mixture law, not high-dimensional empirical TV estimates.

Horizon	KL integral	KL upper bound	L^1-TV integral	L^1-TV upper bound
1	2.1033	3.2617	1.5094	2.0000
2	0.8497	2.0875	0.9991	2.0000
5	0.0418	0.5472	0.2119	1.0461
10	0.0002	0.0588	0.0123	0.3428

The two-dimensional analytic specialization supplies a complementary direct numerical integral. Its one-step implemented density agrees with the analytic conditional Gaussian density at the checked inputs, and the reported mixture KL and L^1-TV integrals lie below their respective finite-start bounds. Quadrature and finite-sample checks do not replace the proof. In a conditioning negative control, both correct $\alpha \sim \gamma_1$ independent of $z \sim \gamma_2$ and incorrect $\alpha = -z_1$ have the correct marginal action law. After the same $\pi/4$ joint rotation and coefficient-0.8 noise update, the first-coordinate variances are 0.998 and 1.650, against theoretical values 1 and 1.64. Thus marginal Gaussian action fit alone does not supply the conditional chart.

Table 28: CPU timing of the parallel prior, with four CPU threads, float32, batch size 32, six action coordinates, hidden width 128, four joint layers and two post layers. Times are medians of five repetitions. The measurement excludes the optimizer, data loading, CNN encoder/decoder, and simulator. Parameter memory is not peak process memory; these measurements do not predict GPU throughput or trained-model accuracy.

State width	Parameters	Parameter MiB	Forward/backward ms	Sampling ms
256	302,856	1.16	7.63	2.50
512	484,360	1.85	8.84	3.24
1024	847,368	3.23	13.40	4.57

The timing experiment is a four-thread CPU microbenchmark of the prior alone. At latent widths 256, 512, and 1024, the kernel uses 0.30, 0.48, and 0.85 million parameters (Table 28; timing medians are plotted in Figure 3c). The benchmark excludes observation networks, optimization, and data loading; each fitted recurrent checkpoint additionally records GPU memory and full-model throughput. Numerical GPU results for the recurrent model are reported under Appendix C.4.

H Controlled occupancy and downstream planning

H.1 A system with an exact conditional action chart

This system separates the cost of an unconditional action chart from the cost of requiring preservation at every action. The controlled Gaussian data follow

$$a \mid z \sim \mathcal{N}(\tau z, 1-\tau^2), \qquad z' = \varrho a + \sqrt{1-\varrho^2}\,\xi, \qquad \xi \sim \gamma_1. \tag{35}$$

Starting with $z \sim \gamma_1$ gives standard Gaussian state and action marginals for $|\tau|, |\varrho| < 1$. Their dependence changes with τ. The exact chart $\alpha = (a - \tau z)/\sqrt{1-\tau^2}$ removes that dependence; a marginal action transform cannot, since a is already marginally Gaussian. The excess action-chart NLL from modeling $H(a)$ instead of $H(a \mid z)$ is $-\frac{1}{2}\log(1-\tau^2)$, or 0.2231 and 0.8304 at $\tau = 0.6, 0.9$. This chart cost is distinct from transition NLL.

For independent actions, the optimal excess transition NLL over all per-action-preserving kernels is $-\frac{1}{2}\log(1-\varrho^2)$ (Proposition 29). For dependent actions we optimize the narrower implemented family $z' = c(a)z + \sqrt{1-c(a)^2}\varepsilon$ by minimizing its Gaussian conditional cross entropy for each action and integrating over $a \sim \gamma_1$. At $\varrho = 0.8$, its excess costs are $0.5108, 0.3681, 0.1408$ for $\tau = 0, 0.6, 0.9$. Only the first is an unrestricted per-action-class floor. A 250×250 copula-grid optimization followed by truncated 24-node action quadrature gives approximately 0.32 and 0.04 for the latter two settings. These approximate, uncertified class optima show that the scalar OU restriction contributes materially to the measured advantage.

Table 29: Controlled Gaussian conditional fit. Excess transition NLL is relative to the true conditional entropy. The scalar OU optimum equals the unrestricted per-action floor only at $\tau = 0$. Each configuration has one training seed.

			Transition excess NLL		Per-action OU family	
ϱ	τ	Chart	Occupancy	Per-action	Optimum	Loss/optimum
0.3	0.0	Conditional	0.0011	0.0470	0.0472	1.00
0.5	0.0	Conditional	0.0011	0.1428	0.1438	0.99
0.8	0.0	Conditional	0.0011	0.5089	0.5108	1.00
0.95	0.0	Conditional	0.0011	1.1623	1.1640	1.00
0.8	0.6	Conditional	0.0011	0.3668	0.3681	1.00
0.8	0.6	Marginal	0.0011	0.3667	0.3681	1.00
0.8	0.9	Conditional	0.0011	0.1399	0.1408	0.99
0.8	0.9	Marginal	0.0011	0.1399	0.1408	0.99

Each configuration uses 400,000 training and 200,000 held-out transitions, 20,000 updates, and 32,768 Gaussian-start chains through 10^5 steps. The occupancy model's approximately 0.0011-nat excess is a finite-sample fit statistic, not a certified optimization optimum.

Table 30: **Action-chart and closed-loop diagnostics.** A dash indicates an unrecorded chart diagnostic; correlation refers to the chart output and state in the occupancy model. Second moments are measured at $T = 10^5$ under the behaviour policy, and all listed chains remain finite. The final column is a separate 200-step Gaussian-start KS diagnostic for the per-action model under that policy. It is not a test of preservation for a fixed action.

				$\mathbb{E}\lvert z_{10^5}\rvert^2$		
ϱ	τ	Chart	$\mathrm{Corr}(\alpha, z)$	Occupancy	Per-action	Per-action KS
0.3	0.0	Conditional	—	1.0071	0.9936	0.0014
0.5	0.0	Conditional	—	1.0089	0.9936	0.0015
0.8	0.0	Conditional	—	1.0107	0.9936	0.0014
0.95	0.0	Conditional	—	1.0102	0.9936	0.0015
0.8	0.6	Conditional	0.0011	1.0108	1.0666	0.0089
0.8	0.6	Marginal	0.6003	1.0107	1.0666	0.0088
0.8	0.9	Conditional	0.0039	1.0132	1.1256	0.0179
0.8	0.9	Marginal	0.9004	1.0130	1.1256	0.0179

Table 31: **Held-out chart likelihood.** Costs are nats per action coordinate. The entropy is analytic for the Gaussian behaviour policy. The difference from this entropy is an empirical excess cost and can be slightly negative from sampling.

ϱ	τ	Chart	NLL	$H(a \mid z)$	Excess
0.8	0.6	Conditional	1.1953	1.1958	-0.0005
0.8	0.6	Marginal	1.4186	1.1958	$+0.2228$
0.8	0.9	Conditional	0.5880	0.5886	-0.0005
0.8	0.9	Marginal	1.4204	0.5886	$+0.8318$

At $\tau = 0.9$, the fitted conditional and marginal charts have state correlations 0.0039 and 0.9004; their chart NLL gap is 0.8324 nats. These check the learned transform against the analytic 0.8304-nat reference. The ideal transform, rather than this finite correlation test, gives exact independence.

A fixed-model dependence intervention. A raw-action-conditioned OU kernel preserves the reference for each external action while potentially failing under state-dependent actions. At $\tau = 0.9$, the fitted model has 200-step KS discrepancy 0.0179 under the behaviour policy versus a Gaussian control of 0.0020. A separate fixed-model intervention uses 200,000 training transitions, 8,000 updates, and 2,000 rollout steps. State-dependent driving gives second moment 1.115; independent driving of the same model gives 0.9985, with KS 0.0014 versus control 0.0015. The action marginals agree under the reference state law, while their state dependence differs. This intervention isolates that dependence in the tested system.

H.2 Planning on the stochastic Duffing oscillator

This comparison measures planning error under learned transitions with different preservation constraints. The system is $\mathrm{d}x = v\mathrm{d}t$ and $\mathrm{d}v = (-\delta v - \alpha_D x - \beta x^3 + \mathrm{drive}\, a)\mathrm{d}t + \sigma \mathrm{d}W$, with $\mathrm{d}t = 0.05$ and $(\alpha_D, \beta, \delta, \mathrm{drive}, \sigma) = (-1, 1, 0.2, 2.5, 0.25)$. Collection uses $a_t = \mathrm{clip}(0.6 \tanh x_t + u_t, -1, 1)$ with an OU driver u_t. The learned action chart is therefore approximate: clipping creates boundary mass and the driver retains memory. This planning dataset does not have the known nonsingular innovation chart used in the recurrent protocol of Appendix C.4.

Each model uses the same 400,000 transitions and 20,000 updates. The state chart and action chart have 534,576 and 78,944 parameters; the occupancy transition has 503,004 and the Gaussian transition 792,070. The ablated mixer has 269,364 parameters. The per-action arm feeds the raw action to its transition, so its name here differs from the conditioner-only ablation of the recurrent study. The objective adds total state-marginal, action-chart, and conditional NLL; the implementation does not divide the two-dimensional conditional NLL by two.

CEM uses 512 candidates, four iterations, and 64 elites to minimize a swing-up-and-hold cost. The real simulator executes each selected action. Three seeds share evaluation episodes and planner randomness across models at horizons $5, 15, 50, 150, 500$. The independent replication unit is the seed, with correlated repeated horizons within a seed.

Table 32: **Paired planning costs on the controlled Duffing oscillator.** Oracle cost and each model's excess cost Δ over the oracle are mean $\pm$ sample standard deviation across three seeds. Models share evaluation episodes and planner seeds within each replication. Horizons are repeated measurements; they are not treated as independent samples.

H	Oracle cost	Ours: Δ	Per-action: Δ	Gaussian: Δ
5	0.204 ± 0.083	-0.004 ± 0.004	$+2.047 \pm 0.988$	$+0.003 \pm 0.002$
15	0.212 ± 0.091	-0.000 ± 0.007	$+1.746 \pm 0.802$	$+0.002 \pm 0.005$
50	0.253 ± 0.117	$+0.005 \pm 0.022$	$+1.630 \pm 0.789$	$+0.003 \pm 0.004$
150	0.433 ± 0.250	$+0.016 \pm 0.099$	$+1.441 \pm 0.630$	-0.013 ± 0.023
500	0.796 ± 0.550	$+0.242 \pm 0.181$	$+1.060 \pm 0.745$	$+0.001 \pm 0.015$

Table 33: **Oracle-normalized calibration gaps.** The gap is the absolute difference between predicted and achieved cost, divided by the oracle's gap at the same horizon and seed. Entries are mean $\pm$ sample standard deviation across three seeds; the oracle is 1 by definition.

H	Ours	Per-action	Gaussian
5	1.44 ± 0.07	2.27 ± 0.37	1.04 ± 0.03
15	1.48 ± 0.14	3.06 ± 0.30	0.97 ± 0.08
50	1.43 ± 0.34	4.87 ± 0.48	0.99 ± 0.14
150	1.18 ± 0.20	2.28 ± 0.38	1.07 ± 0.07
500	0.73 ± 0.15	0.96 ± 0.10	1.01 ± 0.06

The Gaussian baseline is close to the oracle at every reported horizon. The occupancy model is also close at short horizons but has mean excess 0.242 at horizon 500. The per-action arm's excess cost is positive in all 15 seed–horizon cells. Averaging horizons within each seed gives excess costs $1.325, 0.976, 2.453$ for per-action, $-0.009, 0.089, 0.076$ for occupancy, and $-0.001, 0.002, -0.003$ for Gaussian. A two-sided sign test with only three independent seeds gives $p = 0.25$ for per-action and $p = 1$ for each other arm.

The fixed-budget oracle itself becomes less effective at long horizons: its cost increases from 0.204 to 0.796. The normalized calibration ratio therefore depends strongly on the oracle denominator and is not monotone in horizon. The per-action deficit also coincides with worse conditional fit (-8.70 versus -9.34 nats per transition). The comparison does not isolate occupancy drift as its cause. One Gaussian seed has a nonfinite dispersion diagnostic by 10^3 autonomous steps despite 98.86% survival, while its planner remains competitive, showing that autonomous-rollout and planning evaluations measure different properties.

I Numerical exactness and structural diagnostics

I.1 Independent density and preservation checks

For a diffeomorphism T, change of variables gives

$$D_{\mathrm{KL}}(T_{\#}\gamma\|\gamma) = \mathbb{E}_{z\sim\gamma}\left[\tfrac{1}{2}(\|T(z)\|^2 - \|z\|^2) - \log|\det J_T(z)|\right]. \qquad (\ddagger)$$

A preserving map makes the integrand zero pointwise. A small sampled mean alone is insufficient because large signed residuals can cancel. Independent automatic differentiation and general matrix factorization check the analytic Jacobian against this identity.

Table 34: **Pointwise Gaussian identity.** Six perturbed transition families, 8,192 inputs, $d = 32$, CPU double precision. The analytic log-determinant is independently checked against 16 autograd Jacobians per family. Ranges show the two sides of equation †; exactness is an architectural property.

Layer	Max identity defect	$\log\lvert\det J\rvert$	$\frac{1}{2}\Delta\lVert z\rVert^2$	Exact?
Rotation	3.1×10^{-13}	$[0,0]$	$[0,0]$	yes
Quantile swirl	4.1×10^{-13}	$[-0.015,+0.014]$	$[-0.015,+0.014]$	yes
Grid swirl	4.2×10^{-13}	$[-0.074,+0.064]$	$[-0.074,+0.064]$	yes
Rotation + quantile	4.6×10^{-13}	$[-0.012,+0.012]$	$[-0.012,+0.012]$	yes
Affine	7.8×10^{-1}	$[-0.309,+0.121]$	$[-0.856,+0.567]$	no
Additive	2.2×10^{-1}	$[0,0]$	$[-0.194,+0.217]$	no

The additive control's zero log-determinant does not preserve the Gaussian: its identity residual is 0.217. The largest analytic-versus-autograd log-determinant discrepancy among the tested map families is 8.8×10^{-14}. A separate bijective-noise sampler check compares $\log p(z' \mid z, a)$ with $\log\gamma(\varepsilon) - \log|\det(\partial z'/\partial\varepsilon)|$. Across seven cells with dimensions 4–16, one to four noise blocks, and parameter perturbations through scale three, its maximum discrepancy is 4.4×10^{-12}. These are finite numerical checks, not proofs for arbitrary weights or conditioning.

Orthogonal-map tests on 8,192 inputs show residual scales near 10^{-14} in float64, 10^{-6} in float32, and 10^{-2} with TF32 enabled. Exactness checks disable TF32. Checkpoint geometry checks also find small, nonzero floating-point orthogonality defects. Neither these local measurements nor agreement of a few marginal moments supplies a uniform long-run roundoff bound.

Table 35: Sequential-scan rotation–OU arithmetic check: 16,384 chains, dimension 32, two noise blocks, float32 with TF32 disabled, and independent actions. Selected horizons from the complete recorded sweep. The independent Gaussian row has the same sample size. This checks numerical consistency at fixed weights, not a uniform error theorem.

Horizon	RMS	Maximum mean	Maximum covariance defect	Kurtosis
1	0.997796	0.0156	0.0266	2.9885
10^2	1.000537	0.0172	0.0326	2.9993
10^4	0.999767	0.0181	0.0291	2.9894
10^6	1.001119	0.0158	0.0325	2.9985
Independent Gaussian	1.000560	0.0204	0.0247	3.0110

The conditioning graph is equally important: additive layers with angle-tied noise give RMS 1.0014 but KS 0.0361, and choosing identity for $|x| < 1$ with an independent Gaussian refresh otherwise gives standard deviation 0.7186 and KS 0.1087, versus 1.0003 and 0.0023 when the branches are selected by an independent draw instead. Each component preserves the reference; selection based on the coordinate being updated need not.

I.2 Architecture scope: parallel prior and sequential scan

The main construction rotates the joint state–innovation input, reads only the transformed innovation coordinates to determine its diagonal noise coefficients, and updates all state coordinates in parallel. Its postmaps are Gaussian rotations with unit Jacobian determinant. A full-reference reset gives an exact two-component density mixture. The cap applies to every persistent stochastic state coordinate. The proof of its finite-start entropy bound concerns this conditioning graph and must not be inferred from a finite-basis operator estimate.

The sequential-scan experiments use a scan whose block-i conditioner reads $(y_{<i}, x_{>i}, a)$ and excludes the updated x_i. Rotation couplings and CDF-conjugated quantile swirls surround the scan. A quantile swirl uses smooth, compactly supported planar rotations on the open unit cube; conjugating by the Gaussian CDF balances its radius change against its Jacobian. These maps can preserve the Gaussian without preserving Euclidean volume in Gaussian coordinates. The sequential-scan L^2 bound (Theorem 47) and the parallel bound (Section 4, Appendix B) are proved under their respective conditioning graphs.

Finite copulas and noise-only maps. The copula family used with the sequential scan mixes Gaussian copulas and finite cosine expansions with positive normalized weights and zero-integral cosine basis functions. A sufficient coefficient bound enforces positivity while retaining uniform margins. Its conditioner excludes the updated coordinate; sampling inverts the monotone conditional CDF using 40 double-precision bisection steps. This is a finite preserving subfamily of the scalar class in Theorem 41. Gaussian-preserving transformations of fresh noise alone leave its conditional law, hence the transition kernel, unchanged (Lemma 40). This observation does not restrict useful state–noise interactions or complementary-coordinate conditioners.

I.3 A scalar illustration of error sensitivity

At $\lambda = 0.999$, the coefficient perturbation $\delta = 10^{-4}\lambda$ is a relative error of 0.01%. It increases the stationary variance by about 11% while costing 2.5×10^{-6} nats of conditional KL averaged under the true unit-variance state law (Proposition 36 states and proves the calculation). Tying the fitted coefficient and scale through $m^2 + s^2 = 1$ removes this particular variance error.

J Capacity and operator diagnostics

J.1 Finite transition capacity

This experiment compares the finite transition families on a fixed preserving target. Combining rotations and swirls improves on either primitive alone. Broader stochastic dependence also helps,

but the unconstrained models remain better at this training budget. Their substantial late-training improvement makes these outcomes fitted-model comparisons, rather than converged approximation limits.

Table 36: Conditional capacity at a fixed Gaussian reference with no learned encoder. All models use 6,000 updates on the same preserving target kernel. NLL is per latent dimension. Deterministic depth and noise expressivity are distinct changes.

Transition	Parameters	NLL/dim
Quantile swirls, two OU blocks	$670,406$	0.6981
Rotations, two OU blocks	$672,366$	0.6609
Both primitives, two OU blocks	$671,386$	0.6265
Both, twice the deterministic depth	$1,209,134$	0.5943
Both, four OU blocks	$738,202$	0.5824
Both, two copula blocks	$705,442$	0.5394
Unconstrained affine stack	$607,089$	-0.6276
Unconstrained conditional flow	$544,872$	-0.7622
Unconstrained flow, twice the depth	$1,089,744$	-0.9843

The kernel-only copula improvement does not automatically transfer to joint chart–transition learning: at 60,000 updates with three noise blocks, the OU pair NLL is 2.1777 nats per dimension against 2.3221, 2.1895, and 2.3104 for the three broader copula variants despite their 3–10% more parameters. The representation theorem describes an ideal class of stationary pair laws; it does not certify the capacity or optimization of these finite families.

J.2 A NON-GAUSSIAN COPULA EXAMPLE

Let z, α be independent standard Gaussians and define

$$s = \mathrm{frac}(\Phi(z) + \Phi(\alpha)), \qquad c(s, q') = 1 + 0.9\cos(2\pi s)\cos(\pi q'), \qquad z' = \Phi^{-1}(q').$$

The density is at least 0.1 and has uniform margins. For each fixed α, s is uniform when z is Gaussian, so the kernel also preserves the Gaussian action by action. It separates a scalar Gaussian-copula restriction from the occupancy-versus-per-action distinction. Numerical quadrature gives information $\int c \log c = 0.111716$ nats. For $x = \Phi^{-1}(s)$ and $y = \Phi^{-1}(q')$, symmetry gives $\mathbb{E}[xy] = 0$. A global Gaussian-copula correlation ρ therefore has expected log density ratio $-\frac{1}{2}\log(1-\rho^2) - \rho^2/(1-\rho^2) \le 0$, with equality at independence. That scalar family loses the entire information in this example.

Table 37: **A non-Gaussian copula witness.** Scores are held-out log density ratios relative to independent Gaussian output, in nats, with mean $\pm$ sample SD across three training/data seeds. The truth has population score 0.111716 by numerical quadrature. The adaptive Gaussian arm is a conditional-density baseline without a guaranteed Gaussian output margin.

Family	Parameters	Recovered score	Truth minus score
Global Gaussian copula (analytic optimum)	1	0	0.1117
Input-dependent Gaussian conditional	$33,537$	0.0636 ± 0.0005	$+0.0481$
Cosine copula, order 3	9	0.1123 ± 0.0005	-0.0005

The order-three learned cosine basis constrains $\sum_{kl} |\beta_{kl}| \le 0.499$ and contains the truth at $\beta_{2,1} = 0.45$. Its three-seed score 0.1123 ± 0.0005 uses sample SD; the slight excess over population information is held-out sampling variation. The adaptive Gaussian arm recovers 0.0636 ± 0.0005 by letting its coefficient depend on the scalar being updated. It is a normalized conditional density, but violates the exact noise stage's conditioning rule and has no guaranteed Gaussian output marginal. Its remaining gap is a fitted-family result, not an ablation of an exact preserving component. The cosine basis contains this specific truth, without parameterizing every positive copula.

J.3 FINITE-BASIS OPERATOR ESTIMATES

Population compressions onto a finite basis cannot certify the full Markov operator norm. Sampling error can further inflate their estimates, while truncation can omit slow modes. Constant-OU controls recover known norms within 0.0047 and orthogonal deterministic maps return norm one. A non-normal Gaussian example has singular norm 0.900 but spectral radius 0.474, showing why those quantities cannot be interchanged. An OU step with coefficient 0.9 followed by a fair global sign flip removes odd correlations: a linear-only estimate is 0.004, while quadratic estimates 0.806 and 0.812 detect the true quadratic mode 0.810.