STRUCTURE AND LIMITS OF COMPOSABLE COARSE-TO-FINE GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study coarse-to-fine generative models that combine deterministic coarse evolution with stochastic reconstruction. We characterize the temporal laws of Markov rollouts with transition kernels $Q_t(u) = G(\phi_t(Ru))$ that compose exactly at all positive times. Under Lipschitz regularity of the restriction, quadratic-Wasserstein Lipschitz regularity of the decoder, and controlled coarse-flow regularity, every positive-time joint law is a mixture of independent emissions along a deterministic flow of persistent classes. This representation does not require exact consistency between reconstructed samples and their prescribed coarse states. It implies that mean-square continuity forces the visited class laws to be point masses. Within the same exactly composable setting, we bound the path-law distance to independent reconstruction under approximate fiber consistency, with dependence on decoder sensitivity and fiber error. Explicit constructions show that weaker decoder regularity permits stochastic class evolution and that increasing decoder sensitivity can preserve temporal dependence as the fiber error vanishes. Controlled temporal experiments separate identical marginals from path statistics, and a small learned reconstruction study shows why improved temporal correlation alone is insufficient.

1 INTRODUCTION

A coarse description of a physical system generally does not uniquely determine its fine-scale state. Conditional generative models represent this ambiguity by sampling fine-scale states given the coarse description, with applications to atmospheric downscaling (Mardani et al., 2025), turbulence super-resolution (Fan et al., 2025), molecular reconstruction (Wang et al., 2022), and cosmological super-resolution (Li et al., 2021). Using reconstruction in a dynamical model requires understanding how the generated states depend on one another over time.

Let X and C be the fine and coarse state spaces, and let $G(c)$ denote the reconstruction distribution for coarse input $c \in C$. We study an architecture that restricts a fine state $u \in X$ to a coarse state $Ru \in C$, evolves it under a deterministic coarse semigroup $(\phi_t)_{t\geq 0}$, and samples from $G(\phi_t(Ru))$. Its transition kernel (Kallenberg, 2021, p. 233) is

$$Q_t(u, \cdot) = G(\phi_t(Ru)), \qquad t > 0. \tag{1}$$

For a measurable set $A \subseteq X$, $Q_t(u, A)$ is the probability of reaching A after a transition of duration t from u. The dot in $Q_t(u, \cdot)$ leaves the set argument unspecified, denoting the entire probability measure, also abbreviated $Q_t(u)$. The coarse semigroup property is $\phi_{s+t} = \phi_t \circ \phi_s$. The restriction R and decoder G are fixed across physical time, and each successive state is sampled conditionally on the current state. These transitions define a Markov rollout (Kallenberg, 2021, p. 234).

For consistency across time partitions, we require transitions of lengths s then t to have the same endpoint distribution as one transition of length $s+t$. This requirement is expressed by the Chapman–Kolmogorov relation (Kallenberg, 2021, p. 233), which we impose for every initial state $u \in X$, measurable set $A \subseteq X$, and pair of positive times:

$$(Q_s Q_t)(u, A) := \int_X Q_t(v, A)\, Q_s(u, dv) = Q_{s+t}(u, A), \qquad s, t > 0. \tag{2}$$

Here $Q_s(u, dv)$ is the distribution of the intermediate state v after time s. The integral averages $Q_t(v, A)$ over this distribution, giving the probability that the two-step transition from u ends in A.

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.

Equation (2) averages over the intermediate state to compare endpoint distributions. To study how intermediate and later states depend on one another, we consider their joint law, which describes the distribution of those states together. Starting from a fixed fine state $U_0 = x$, let U_{t_j} denote the generated fine state at observation time t_j, where $0 < t_1 < \cdots < t_n$. For measurable sets $A_j \subseteq X$,

$$\mathrm{Law}_x(U_{t_1}, \ldots, U_{t_n})(A_1 \times \cdots \times A_n) = \Pr_x(U_{t_1} \in A_1, \ldots, U_{t_n} \in A_n)\,.$$

Here Law_x denotes the distribution and $\Pr_x$ the probability for a rollout started at x. The right-hand side is the probability that, for every j, the generated state U_{t_j} lies in A_j.

The architecture restricts how this temporal dependence can arise. At each observation time t_i, the generated state U_{t_i} enters the next transition only through the coarse state RU_{t_i}, obtained using the same restriction R. That coarse state is evolved and used to condition the next reconstruction. Random differences between generated fine states can therefore affect later samples only through information retained by R.

We prove that, for models satisfying Eqs. (1) and (2) and the regularity assumptions in Section 2, these joint laws are mixtures of product distributions. The initial choice is a random variable Λ taking values in a space of class labels. Each label is a reconstruction distribution on X shared by the states in that class (Section 3). Given Λ, samples at any finite collection of distinct positive times are independent. In particular, observing an intermediate sample does not change the predicted distribution of a future state once Λ is fixed:

$$\Pr_x(U_t \in A \mid \Lambda, U_s) = \Pr_x(U_t \in A \mid \Lambda), \qquad 0 < s < t.$$

Here $A \subseteq X$ is measurable. Equation (2) alone does not imply this conditional independence, as the Brownian motion example in Section 3 illustrates. Theorem 1 derives it from the architecture and regularity assumptions together.

This theorem gives practical guidance for selecting reconstruction models in dynamical applications. An architecture of this form can be considered when dependence across time is intended to arise from the initial choice Λ, with later reconstructions sampled independently conditional on that choice. If random fluctuations generated during reconstruction must influence subsequent evolution even after Λ is fixed, the architecture cannot meet that requirement under the theorem's assumptions.

Some applications require the expected squared difference between states at nearby times to vanish as the time interval shrinks (e.g., Wang et al., 2021, Section 3). This property is called mean-square continuity (Rasmussen and Williams, 2006, Section 4.1.1, p. 81). Under our assumptions, Corollary 3 shows that it requires each positive-time reconstruction to be deterministic once Λ is fixed. Thus, taking smaller time steps cannot satisfy this requirement while the conditional reconstructions retain nonzero sampling variance. Section 4 develops this consequence and gives an example where relaxing the assumed control on how $G(c)$ varies with the coarse input c allows intermediate randomness to affect future predictions.

An important special case requires reconstructed fields to reproduce the coarse input exactly when restricted, a constraint used to preserve prescribed coarse-scale quantities in scientific super-resolution and downscaling (Geiss and Hardin, 2023; Harder et al., 2023). For $U \sim G(c)$, this requirement is $RU = c$ almost surely. We refer to this property as exact fiber consistency. Equivalently, $R_{\#}G(c) = \delta_c$, where the pushforward $R_{\#}G(c)$ is the distribution of RU (Peyré and Cuturi, 2019, Definition 2.1) and δ_c is the point mass at c. Together with the coarse semigroup property, this constraint implies Eq. (2). For a fixed initial coarse state, it also makes the positive-time joint law a product of independent reconstructions (proofs in Appendix A). This explicit joint law provides a baseline for evaluating correlations between reconstructed states and uncertainty in quantities averaged over time.

We then quantify how imperfect coarse recovery changes this baseline, while retaining exact composition. Theorem 5 bounds the distance between the joint distribution of a generated sequence and that of independent reconstructions. The bound depends on the coarse recovery error, how strongly the reconstruction distribution responds to changes in its coarse input, and how much coarse evolution enlarges differences between inputs. With these controls and the total time interval fixed, the bound does not grow with the number of time steps.

When this bound tends to zero, the model approaches independent reconstruction. Corollary 8 quantifies the remaining error in the joint distribution when the desired evolution is mean-square

continuous and remains random after the initial coarse state is fixed. This error stays positive as time steps shrink, even if the distribution at each time is matched. We next examine whether temporal dependence can persist as coarse recovery errors vanish. Proposition 9 constructs such a model using rotation, which carries an initial random direction through time without changing its magnitude. The model encodes this direction in small deviations from exact coarse recovery. As these deviations shrink, the generated sequences approach the prescribed rotational evolution, while the decoder responds increasingly strongly to small input changes.

The analysis also covers Euclidean state spaces that are not bounded, under explicit bounds on the expected squared size of states and on responses to input changes.

Controlled experiments and learned PDE reconstruction studies examine these consequences. In periodic advection, processes with identical single-time distributions have different temporal correlations and uncertainty in time averages. Rotation experiments illustrate how temporal dependence persists as coarse recovery errors decrease, while a separate perturbation experiment shows how coarse errors accumulate when composition is only approximate. In a learned Burgers study, using a previous generated state improves temporal correlation but worsens single-time distribution scores and overestimates time-average variability. Evaluations on Allen–Cahn, Burgers, and Navier–Stokes further examine coarse recovery, endpoint accuracy, and behavior over repeated reconstruction.

2 Setting and relation to existing theory

Let (X, d_X) and (C, d_C) be fine and coarse metric spaces, $R : X \to C$ a restriction, and $G : C \to \mathcal{P}_2(X)$ a probability kernel. Here d_X, d_C are the respective distances, $\mathcal{P}_2(X)$ denotes probability laws with finite second moment, and W_2 is their quadratic Wasserstein distance (Villani, 2009, Definitions 6.1 and 6.4). The deterministic coarse maps form a semigroup: $\phi_0 = \mathrm{Id}$ and $\phi_{s+t} = \phi_t \circ \phi_s$. We assume joint continuity: $(t, c) \mapsto \phi_t(c)$ is continuous in time and initial coarse state together. For a bounded measurable function f, the transition kernel Q_t acts by $Q_t f(x) = \int_X f(y) Q_t(x, dy)$, the expected value of f after a transition from x. For an initial distribution ν, write $(\nu Q_t)(A) = \int_X Q_t(x, A)\nu(dx)$ for the resulting distribution. Unless specified, integrals are over the full space on which the integrating measure is defined. The classification assumes Eq. (2) for all initial states and all $s, t > 0$, without requiring exact fiber consistency. Agreement at finitely many numerical step sizes is insufficient.

We first work with compact X, C to obtain uniform moment bounds and convergent subsequences in the proofs. This includes models whose fine and coarse states stay in closed bounded subsets of finite-dimensional spaces. Appendix B treats unbounded Euclidean spaces. Assume R and G are Lipschitz, meaning

$$d_C(Rx, Ry) \le L_R d_X(x, y), \qquad W_2(G(c), G(c')) \le L_G d_C(c, c')$$

for all $x, y \in X$ and $c, c' \in C$. On each finite horizon T, assume

$$L_T := \sup_{0 \le t \le T} \mathrm{Lip}(\phi_t) < \infty, \qquad \sup_{c \in C} d_C(\phi_h(c), c) \le B_T h \quad (0 < h \le T). \tag{3}$$

The first bound limits how much ϕ_t amplifies distances between coarse inputs. The second requires each coarse state to move by at most $B_T h$ over a step of duration h. One example is the solution map ϕ_t of an autonomous Lipschitz ODE whose solutions starting in C remain in C for all $t \ge 0$ (Appendix A.1) (Teschl, 2012, Section 2.4). For the transition $Q_t(x) = G(\phi_t(Rx))$ in Eq. (1), the Lipschitz assumptions and Eq. (3) imply

$$\sup_{0 \le t \le T} \mathrm{Lip}_{W_2}(Q_t) \le K_T := L_G L_R L_T, \qquad \sup_x W_2(Q_h(x), P(x)) \le L_G B_T h, \tag{4}$$

where $P(x) = G(Rx)$ and Q_0 in the first bound denotes P. This convention concerns the kernel bound. The rollout itself starts from $U_0 = x$, so $\mathrm{Law}_x(U_0) = \delta_x$, even when $P(x) \ne \delta_x$. In each step from time s to $t > s$, it samples U_t from $Q_{t-s}(U_s, \cdot)$ using only the current state U_s. This sampling rule and the initial law determine the rollout's finite-dimensional distributions: the joint laws at finitely many observation times used in the introduction.

Related work. Our question is what temporal dependence deterministic coarse evolution and stochastic reconstruction can retain under exact composition and the stated regularity assumptions.

A starting point is the structure of kernels satisfying $P^2 = P$, called idempotent: applying P twice has the same distributional effect as applying it once. Classical results group their output states into classes on which P returns the same probability law, while allowing an initial state to select a class randomly (Blackwell, 1942). Modern results extend this decomposition to standard Borel kernels and continuous kernels on compact Hausdorff spaces (Fritz et al., 2026, arXiv v5, Corollaries 4.5.5 and 4.5.10). This decomposition alone does not determine how the classes evolve over time. For the architecture in Eq. (1), we derive deterministic class evolution from Eq. (4) and prove that all positive-time reconstructions are independent once the initial class is fixed.

A related question is whether coarse observations of a fine Markov process themselves form a Markov process. Markov-function criteria address when this holds (Rogers and Pitman, 1981). Eliminating variables can also make the retained dynamics depend on past states (Lapolla and Godec, 2019). To predict a dynamical system when only some variables are retained, Chorin et al. (2002) derive equations with memory terms that account for this dependence. These approaches start from a fine process and examine its coarse description. In the other direction, finite-state memoryless embeddings lift a coarse Markov chain by drawing each fine state from a distribution determined by its coarse state (Wolfer and Watanabe, 2024, Section 4.4.2). These embeddings provide an analogue of reconstruction with exact coarse recovery. Our main theorem starts from the prescribed fine transition in Eq. (1) and determines its joint laws without assuming exact coarse recovery.

Matching distributions at individual times is distinct from matching dependence across times. Under their stated assumptions, Brunick and Shreve (2013, Corollaries 3.7, 3.9, and 3.10) construct stochastic differential equations whose solutions match the distribution of a given process at each time. Their construction also covers time integrals and the maximum reached up to each time. These results allow the dynamics to be chosen to match the desired statistics. Here the transition architecture is fixed. Theorem 1 characterizes its positive-time joint laws under exact composition and regularity. Under its additional conditions, Corollary 8 quantifies the remaining joint-distribution error as the single-time distributions are matched, with a positive limit for targets that remain random after the initial coarse state is fixed.

3 A CLASSIFICATION OF THE TEMPORAL LAW

We retain the compact spaces of Section 2. To group states by their reconstruction distributions, define

$$S = \overline{\bigcup_{x \in X} \operatorname{supp} P(x)} \subseteq X, \qquad \ell(u) = P(u), \qquad \mathcal{M} = P(S) \subseteq \mathcal{P}_2(X). \tag{5}$$

For a Borel set $A \subseteq X$, write $P(x, A) := P(x)(A)$ for the probability that a draw from $P(x)$ lies in A. The closure is taken in X, so S is compact. Every $P(x)$ is supported on S, hence $P(x, S) = 1$. Define $u \sim v$ when $P(u) = P(v)$ for $u, v \in S$. The equivalence class $[u] = \{v \in S : P(v) = P(u)\}$ groups states with the same reconstruction law. We label this class by that common law $P(u)$. Thus $\mathcal{M} = P(S)$ represents the quotient space $S/\!\sim$ through $[u] \mapsto P(u)$, and contains only these class laws. The label map is $\ell : S \to \mathcal{M}$.

Theorem 1. *Let X, C be compact metric spaces, let $R : X \to C$ be L_R-Lipschitz, and let $G : C \to \mathcal{P}_2(X)$ be L_G-Lipschitz in W_2. Under Eqs.* (1)–(3)*:*

1. *$P^2 = P$ and $PQ_t = Q_t P = Q_t$. For every $x \in X$ and $t > 0$, $Q_t(x, S) = 1$. For $u \in S$, a draw $V \sim P(u)$ stays in the same class with probability one: $P(V) = P(u)$ almost surely.*

2. *With $\mathcal{M}$ equipped with W_2, there is a deterministic, jointly continuous forward evolution $F_t : \mathcal{M} \to \mathcal{M}$, with $F_0 = \mathrm{Id}$ and the semigroup property, such that*

$$F_t(\mu) = \mu Q_t, \qquad Q_t(u) = F_t(P(u)) \quad (u \in S,\ t > 0). \tag{6}$$

   In particular, $P(U_t) = F_t(P(u))$ almost surely when $U_t \sim Q_t(u)$ and $u \in S$.

3. *Fix any $x \in X$. For $V \sim P(x)$, define $\rho_x(D) := \Pr(P(V) \in D)$ for every Borel set $D \subseteq \mathcal{M}$. This is a probability law on the label space $\mathcal{M}$. If $x \in S$, part 1 gives $\rho_x = \delta_{P(x)}$. The joint distributions of the rollout from $U_0 = x$ can be generated as follows. Draw one random label Λ in $\mathcal{M}$ with distribution ρ_x. Given its value μ, sample the states independently at distinct positive*

observation times, with

$$\mathrm{Law}_x(U_t \mid \Lambda = \mu) = F_t(\mu), \qquad t > 0,$$

for ρ_x-almost every μ. The initial state remains $U_0 = x$. Thus, for $0 < t_1 < \cdots < t_n$,

$$\boxed{\mathrm{Law}_x(U_{t_1}, \ldots, U_{t_n}) = \int_{\mathcal{M}} \bigotimes_{j=1}^{n} F_{t_j}(\mu)\, \rho_x(d\mu).} \tag{7}$$

Here $\bigotimes$ denotes the product probability measure, assigning each measurable rectangle $A_1 \times \cdots \times A_n \subseteq X^n$ probability $\prod_j F_{t_j}(\mu)(A_j)$. Averaging these products over the possible labels according to ρ_x gives the joint law without conditioning on Λ.

For ease of reference, Table 1 summarizes the main operators and probability laws introduced so far. A label $\mu \in \mathcal{M}$ is a probability law on X, whereas ρ_x is a law on $\mathcal{M}$. For P and Q_t, functions are written to the right of the kernel, and distributions to its left.

Table 1: Main operators and probability laws used in the classification.

Expression	Input or parameter	Value or type
$R(x)$	State $x \in X$	Coarse state in C
$\phi_t(c)$	Coarse state $c \in C$	Evolved coarse state in C
$G(c)$	Coarse state $c \in C$	Reconstruction law on X
$Q_t(x),\ P(x)$	State $x \in X$	Probability law on X
$Q_t f,\ Pf$	Bounded measurable $f : X \to \mathbb{R}$	Expectation of f as a function of x
$\nu Q_t,\ \nu P$	Probability law ν on X	Resulting probability law on X
$\ell(u) = P(u)$	State $u \in S$	Class label in $\mathcal{M}$
$F_t(\mu) = \mu Q_t$	Class label $\mu \in \mathcal{M}$	Evolved class label in $\mathcal{M}$
ρ_x	Fixed initial state $x \in X$	Distribution of Λ on $\mathcal{M}$

Remark. For $x \in S$, $\rho_x = \delta_{P(x)}$, so every generated sequence uses the same initial label $P(x)$, although its sampled states may still be random. For a fixed $x \notin S$, different sequences may select different initial labels according to ρ_x. Each sequence selects its initial label only once. Appendix A gives an example.

The sampling procedure in part 3 of Theorem 1 gives the same joint distributions as the original rollout from $U_0 = x$, as proved in Appendix A. To explain the conditional independence, fix $0 < s < t$. Part 3 states that, given $\Lambda = \mu$, U_s has distribution $F_s(\mu) \in \mathcal{M}$. By the definition $\mathcal{M} = P(S)$, this law equals $P(v)$ for some $v \in S$. Part 1 of Theorem 1 therefore gives $P(U_s) = P(v) = F_s(\mu)$ almost surely. Since $U_s \in S$ almost surely, part 2 of Theorem 1 and the semigroup property give

$$Q_{t-s}(U_s) = F_{t-s}(P(U_s)) = F_{t-s}(F_s(\mu)) = F_t(\mu).$$

This distribution does not depend on the sampled value of U_s. In Appendix A, the label is chosen initially, and later transitions follow the Markov sampling rule of Section 2: given $U_s = u$ and $\Lambda = \mu$, the next distribution is $Q_{t-s}(u, \cdot)$. Conditioning first on U_s, the conditional law of total probability gives, for Borel $A, B \subseteq X$ and ρ_x-almost every μ,

$$\begin{aligned}
\Pr_x(U_s \in B, U_t \in A \mid \Lambda = \mu) &= \int_B \Pr_x(U_t \in A \mid U_s = u, \Lambda = \mu)\, F_s(\mu)(du) \\
&= \int_B Q_{t-s}(u, A)\, F_s(\mu)(du) \\
&= F_s(\mu)(B)\, F_t(\mu)(A).
\end{aligned}$$

The integration is over B because the event requires $U_s \in B$. The last line assigns $B \times A$ its probability under the product measure $F_s(\mu) \otimes F_t(\mu)$, as defined in Eq. (7). Thus observing U_s does not change the distribution given Λ.

To obtain the probability given only $\Lambda = \mu$, integrate over the conditional law $F_s(\mu)$ of U_s from part 3 of Theorem 1:

$$\begin{aligned}\Pr_x(U_t \in A \mid \Lambda = \mu) &= \int_X \Pr_x(U_t \in A \mid U_s = u, \Lambda = \mu)\, F_s(\mu)(du) \\ &= \int_X F_t(\mu)(A)\, F_s(\mu)(du) \\ &= F_t(\mu)(A).\end{aligned}$$

The first line uses the conditional law of total probability. The integral is over all of X because no restriction is placed on U_s. The second line substitutes the conditional transition probability obtained above. The last line follows because $F_t(\mu)(A)$ does not depend on u and $F_s(\mu)$ has total mass one. Combining the two conditional probabilities gives the identity in the introduction:

$$\Pr_x(U_t \in A \mid \Lambda, U_s) = F_t(\Lambda)(A) = \Pr_x(U_t \in A \mid \Lambda), \qquad 0 < s < t, \tag{8}$$

almost surely. The same argument applies at any finite collection of distinct positive times. Figure 1 illustrates the sampling procedure.

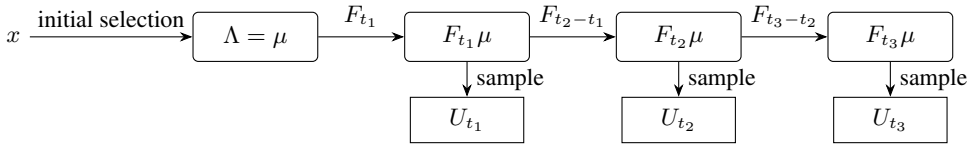

Figure 1: The representation with a shared label in Theorem 1. A single label $\Lambda \sim \rho_x$ selects the initial class. Its value μ determines the distributions along the top row. Given this label, the states below are independent samples. At time t, the current class label is $P(U_t) = F_t(\mu)$ almost surely.

Exact composition versus conditional independence. Brownian motion provides a counterexample to deriving Eq. (8) from the composition relation in Eq. (2) alone. Standard Brownian motion started at $B_0 = 0$ satisfies Eq. (2). Its zero-time kernel is $P(b) = \delta_b$, so the initial label is fixed at δ_0. Nevertheless, observing $B_s = b$ changes the predicted mean of B_t from zero to b for $0 < s < t$ (Pitman and Yor, 2018, Section 4.1). Appendix A gives the calculation. Theorem 1 obtains conditional independence using the additional bounds in Eq. (4), which follow from the architecture and regularity assumptions.

A criterion stated directly in terms of the transitions. In Theorem 1, the architecture and the regularity of its components give the bounds on Q_t in Eq. (4). The proof of deterministic class evolution then uses these bounds together with exact composition, without further use of the decomposition $Q_t = G \circ \phi_t \circ R$. This suggests imposing the needed properties directly on the transition kernels. The following proposition gives sufficient conditions for the same classification for general Markov semigroups, including the continuity and convergence assumptions. Its proof is in Appendix A.

Proposition 2. *Let Q_t, $t > 0$, be a Markov semigroup on compact X, with continuous kernels and continuous $t \mapsto Q_t(x)$ for $t > 0$. Suppose $Q_h \to P$ uniformly in W_2 as $h \downarrow 0$, and define S as in Eq.* (5)*. If, for every $T < \infty$,*

$$\sup_{0 \le t \le T} \mathrm{Lip}_{W_2}(Q_t|_S) < \infty, \qquad \sup_{u \in S} W_2(Q_h(u), P(u)) \le \omega_T(h) = o(\sqrt{h}), \tag{9}$$

where $o(\sqrt{h})$ means $\omega_T(h)/\sqrt{h} \to 0$ and $Q_0 = P$ denotes only the limit, then all conclusions of Theorem 1 hold. The spatial bound is required only on S, whereas zero-time convergence is uniform on X.

Section 4 gives an example showing why the $o(\sqrt{h})$ condition cannot be replaced by $O(\sqrt{h})$. Proposition 6 in Appendix B separately extends the classification to unbounded Euclidean spaces, with global Lipschitz bounds and second-moment control.

4 CONTINUOUS TRAJECTORIES AND A REGULARITY BOUNDARY

Independent draws need not become close when their observation times become close. We work on compact subsets of Euclidean space under the assumptions of Theorem 1, or on Euclidean space under those of Proposition 6. For a distribution μ, let $m(\mu) = \int u\,\mu(du)$ be its mean and $v(\mu) = \int \|u - m(\mu)\|^2\,\mu(du)$ its total variance. Write $\mathbb{E}_x$ for expectation with $U_0 = x$ fixed, without conditioning on Λ.

Corollary 3. *For any fixed initial state $U_0 = x$ and any $t > 0$,*

$$\lim_{h\downarrow 0} \mathbb{E}_x \|U_{t+h} - U_t\|^2 = 2\int_{\mathcal{M}} v(F_t(\mu))\rho_x(d\mu). \tag{10}$$

If the initial label is fixed at μ_0, then $\rho_x = \delta_{\mu_0}$ and the right-hand side reduces to $2v(F_t(\mu_0))$. In particular, $x \in S$ gives $\mu_0 = P(x)$. Thus, for each $t > 0$, mean-square continuity requires $F_t(\Lambda)$ to be a point mass (a Dirac distribution) with probability one when the initial label Λ is drawn from ρ_x.

Appendix C proves the limit and shows that convergence to every starting state also constrains the dimension of S: if $Q_h(u) \Rightarrow \delta_u$ as $h \downarrow 0$ for every $u \in S$, then $\dim_H S \le d_c$ when $C \subseteq \mathbb{R}^{d_c}$.

A counterexample under a weaker temporal bound. Proposition 2 requires $W_2(Q_h(u), P(u)) = o(\sqrt{h})$ uniformly on S. The following counterexample shows that Proposition 2 no longer holds if this condition is weakened to $O(\sqrt{h})$, even when all other assumptions remain satisfied. The example retains the architecture and exact composition. Its decoder is $1/2$-Hölder continuous in W_2, meaning that $W_2(G(c), G(c')) \le H d_C(c, c')^{1/2}$ for all $c, c' \in C$ and some finite constant H. This is the standard Hölder condition (Evans, 2010, Section 5.1), with output distances measured by W_2.

Proposition 4. *There are compact spaces X, C, a linear restriction, a smooth deterministic coarse flow, and a globally $1/2$-Hölder decoder satisfying Eqs.* (1)–(2)*. On the two-point support $S = \{u_-, u_+\}$, with $\|u_+ - u_-\| = 2$, $P(u) = \delta_u$ and Q_t is a continuous-time Markov chain with flip rate $1/2$: the probability of changing state in a short interval h is $h/2 + o(h)$. The decoder is not Lipschitz, and*

$$W_2^2(Q_h(u_+), \delta_{u_+}) = 2(1 - e^{-h}). \tag{11}$$

Appendix D gives the construction and verifies all the remaining assumptions of Proposition 2.

5 COARSE RECOVERY AND DECODER SENSITIVITY

Exact fiber consistency, the application-motivated special case from the introduction, requires $RU = c$ almost surely for $U \sim G(c)$. For fixed $c_0 = RU_0$, $RU_t = \phi_t(c_0)$ almost surely, so reconstruction at each time uses a deterministic coarse input. The positive-time samples are independent: newly sampled fine information cannot affect future predictions. Exact composition also follows (Appendix A). We ask how far the joint law can depart from this independent baseline when coarse recovery is approximate.

Retain exact composition as a separate assumption. In Euclidean fine and coarse spaces, assume $W_2(G(c), G(c')) \le L_G \|c - c'\|$ and

$$\sup_c \left(\int \|Ru - c\|^2\, G(c, du) \right)^{1/2} \le \varepsilon. \tag{12}$$

For an input c and a draw $U \sim G(c)$, the recovered coarse state RU may now differ from c. Thus ε bounds the root-mean-square (RMS) recovery error uniformly over inputs (Chai and Draxler, 2014, cf. Eq. (2)). The constant L_G measures how much the reconstruction distribution can change when its input changes.

For $0 = t_0 < t_1 < \cdots < t_n = T$, set $U_k = U_{t_k}$, $a_k = \phi_{t_k}(c_0)$, and $\Delta_k = t_k - t_{k-1}$. Write $\mathsf{Q}_{\mathrm{roll}}$ for the joint law of $(U_1, \ldots, U_n)$ generated from fixed U_0, and $\mathsf{Q}_{\mathrm{fresh}} = \bigotimes_{k=1}^n G(a_k)$ for independent draws at the prescribed coarse states a_k. Choose any deterministic weights $w_k > 0$ with $\sum_k w_k = T$, fixed before sampling. For example, $w_k = \Delta_k$. The squared sequence distance is $d_w^2(u, v) = \sum_k w_k \|u_k - v_k\|^2$. Its Wasserstein distance $W_{2,w}$ is the square root of the smallest expected weighted squared error over joint pairings of the two sequences.

Theorem 5. *If $\ell_k = \mathrm{Lip}(\phi_{\Delta_k}) < \infty$, then*

$$W_{2,w}(\mathsf{Q}_{\mathrm{roll}}, \mathsf{Q}_{\mathrm{fresh}}) \le L_G \varepsilon \left(\sum_{k=2}^{n} w_k \ell_k^2 \right)^{1/2}. \tag{13}$$

In particular the bound is at most $L_G \varepsilon \ell \sqrt{T}$ when $\ell_k \le \ell$. With T, L_G, ε, and this coarse-flow bound fixed, increasing the number of steps does not enlarge this upper bound.

With bounded coarse-flow Lipschitz constants, $L_G \varepsilon \to 0$ therefore forces the generated sequence to approach independent reconstruction in $W_{2,w}$ (proof in Appendix E).

Approximating a target process. Let Z_t denote the fine-scale evolution that an application asks the model to reproduce. For example, it could consist of fine-grid PDE solutions from random fine initial states sharing the same coarse initial state c_0. This target is supplied by the application (e.g., Fan et al., 2025; Mardani et al., 2025; Schillinger et al., 2026). It is not defined by R, ϕ_t, and G. Suppose it is mean-square continuous and has nonzero variance at some time after c_0 is fixed.

There are two different comparisons: comparing the distributions at each time and summing their squared Wasserstein distances with weights Δ_k, or comparing the joint distributions of the entire sampled sequences. Corollary 8, stated and proved in Appendix E.2, makes the difference precise. Under the assumptions of Theorem 5, if the first error and $L_G \varepsilon$ tend to zero as the grid is refined, and the coarse-flow Lipschitz constants remain bounded, the squared joint-distribution error tends to twice the variance of Z_t integrated over time, a positive limit. Thus accurate distributions in this time-weighted sense and small recovery error do not suffice to reproduce such a target: the model must also depart from independent reconstruction.

Allowing greater decoder sensitivity. The preceding restriction applies when $L_G \varepsilon \to 0$. Rather than accept larger recovery errors, we ask whether growing decoder sensitivity can permit joint approximation as $\varepsilon \to 0$. Proposition 9 in Appendix F gives one example: an exactly composing model stores a random direction in a shrinking coarse offset and transports it by rotation. Its joint laws approach a mean-square continuous rotational target with temporal dependence, while its decoder remains Lipschitz with a constant growing inversely with the offset. Which mean-square continuous targets admit such joint approximation by models of Eq. (1), with exact composition, vanishing coarse recovery error, and decoder Lipschitz constants allowed to grow? To our knowledge, a general characterization is not available.

6 Temporal diagnostics and PDE illustrations

The preceding results predict how independent reconstruction affects temporal statistics and relate departures from it to coarse recovery error and decoder sensitivity. We first compare controlled experiments with these predictions, then examine their implications for evaluating learned PDE reconstructions. Appendices G and H give all protocols and detailed results.

Identical marginals, different temporal observables. Theorem 1 gives conditional independence, and Corollary 3 predicts nonvanishing mean-square increments when the conditional variance is positive. To observe this consequence at fixed coarse context, we compare independent reconstruction with transport of one initial random phase in periodic advection. The reconstruction satisfies exact fiber consistency and composition. Both models have the same coarse trajectory and identical full-field distributions at every positive time. As the time spacing decreases, the measurements follow the analytic predictions: independent reconstruction retains nonzero increments and its time-average variance tends to zero, while transport has vanishing increments and positive limiting time-average variance at a fixed location (Figure 2, Appendix G.1).

The role of exact composition. Under exact composition, Theorem 5 bounds the joint-distribution error relative to independent reconstruction without growth in the number of steps on a fixed horizon. To examine this assumption, we use a Gaussian decoder with small coarse recovery error and a small discrepancy between one transition and two composed transitions. Repeated reconstruction nevertheless accumulates coarse error and moves farther from independent reconstruction as the number of steps increases (Appendix G.2).

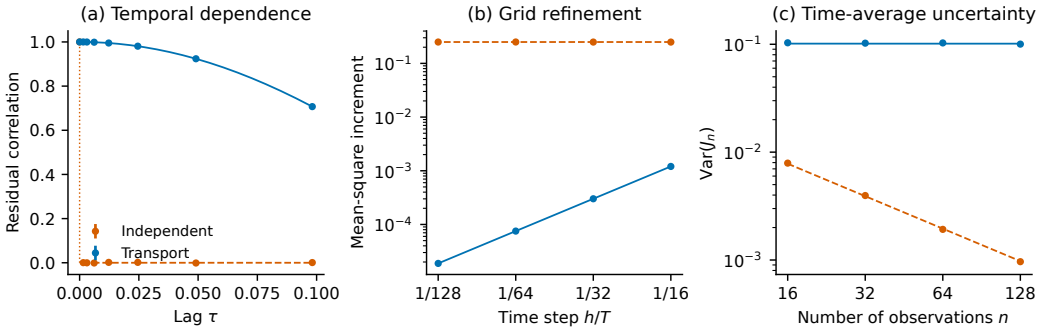

Figure 2: Advection at fixed coarse context. Dots average three seeds, bars show twice the standard error using trajectories as independent units, and lines are analytic predictions. The residual correlation, spatial mean-square increment, and variance of the fixed-location time average distinguish laws with identical single-time marginals. $T = \pi/16$ and $n = 16, 32, 64, 128$. Some bars are smaller than the markers.

Vanishing recovery error with growing sensitivity. Section 5 asks whether increasing decoder sensitivity permits target approximation despite vanishing recovery error. We implement the rotational construction of Proposition 9 for its specified target, shrinking the coarse offset while allowing sensitivity to grow at the rate established by the proposition. The numerical results agree with the predicted vanishing joint-distribution error while temporal dependence persists (Appendix G.3).

Evaluating a learned reconstruction. The distinction between single-time and joint accuracy in Section 5 motivates checking temporal statistics alongside single-time scores in learned models. We compare two Gaussian decoders for one-dimensional Burgers dynamics. Both receive the same reference coarse sequence and preserve its block averages, but only one also uses the previously generated residual. Using this history brings correlation and mean-square increments closer to the reference while worsening the single-time distribution score and overestimating time-average variance (Appendix G.4, Table 2). The supplied coarse sequence isolates reconstruction, so this experiment does not test exact composition.

Coarse recovery in larger PDE reconstructions. Section 5 separates coarse recovery from accuracy of the generated fine evolution. To assess this distinction in larger PDE reconstructions, we compare five methods with different ways of enforcing the coarse constraint on Allen–Cahn, Burgers, and Navier–Stokes problems. Accurate coarse recovery does not give a consistent endpoint accuracy advantage, and some constrained reconstructions become numerically unstable over repeated steps (Appendix H).

7 Conclusion and Limitations

Under exact composition and the stated regularity, the prescribed architecture generates samples at distinct positive times that are independent given the initial class. Thus any required temporal dependence must be carried by that class. The theory and experiments support evaluating temporal correlations and time-average uncertainty alongside single-time accuracy and coarse recovery.

However, our theory assumes that the reconstruction distribution for a given coarse input does not change with physical time. Extending the results to approximate composition requires understanding how composition errors propagate over repeated steps. Models that retain earlier fine information or additional hidden states, or evolve coarse states randomly, also require a separate analysis of their transition rules. We recognize that checking finitely many states and time steps cannot establish that the assumptions hold everywhere, particularly the global bounds required on unbounded spaces. Our experiments illustrate the predicted phenomena and their implications for evaluation, but do not establish that these reconstruction models outperform other PDE solvers.

Reproducibility statement

The appendix contains complete proofs and identifies their classical ingredients. Appendix G specifies the controlled and learned temporal experiments, including analytic predictions, data splits, all seeds, and runtime. Their JAX implementation records configurations, software versions, code hashes, checks, and model checkpoints. Appendix H documents the two-dimensional PDE pipeline and numerical caveats. The horizon analysis reads its saved metrics without retraining. Reproduction commands accompany the experiment source.

Ethics statement

The experiments use synthetic PDE data. The mathematical results do not certify the physical reliability of a deployed simulator. We report numerical failures alongside accuracy metrics.

AI use statement

Large language models assisted with the theoretical framework, statements and counterexamples, proof construction and checking, literature searches and comparisons, interpretation of experimental artifacts, and preparation of the manuscript and reproducibility code. The authors are responsible for the claims, proofs, references, code, and final submission.

References

Luigi Ambrosio and Daniele Puglisi. Linear extension operators between spaces of Lipschitz maps and optimal transport. *Journal für die reine und angewandte Mathematik*, 764:1–21, 2020. https://doi.org/10.1515/crelle-2018-0037.

David Blackwell. Idempotent Markoff chains. *Annals of Mathematics*, 43(3):560–567, 1942. https://doi.org/10.2307/1968811.

Adam Bobrowski, Bogdan Kaźmierczak, and Markus Kunze. An averaging principle for fast diffusions in domains separated by semi-permeable membranes. *Mathematical Models and Methods in Applied Sciences*, 27(4):663–706, 2017. https://doi.org/10.1142/S0218202517500130.

Gerard Brunick and Steven Shreve. Mimicking an Itô process by a solution of a stochastic differential equation. *The Annals of Applied Probability*, 23(4):1584–1628, 2013. https://doi.org/10.1214/12-AAP881.

T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014. https://doi.org/10.5194/gmd-7-1247-2014.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712, 2023. https://proceedings.mlr.press/v202/chen23o.html.

Alexandre J. Chorin, Ole H. Hald, and Raz Kupferman. Optimal prediction with memory. *Physica D: Nonlinear Phenomena*, 166(3–4):239–257, 2002. https://doi.org/10.1016/S0167-2789(02)00446-3.

Lawrence C. Evans. *Partial Differential Equations*. Volume 19 of *Graduate Studies in Mathematics*, second edition. American Mathematical Society, 2010. https://doi.org/10.1090/gsm/019.

Xiantao Fan, Deepak Akhare, and Jian-Xun Wang. Neural differentiable modeling with diffusion-based super-resolution for two-dimensional spatiotemporal turbulence. *Computer Methods in Applied Mechanics and Engineering*, 433:117478, 2025. https://doi.org/10.1016/j.cma.2024.117478.

Bálint Farkas and Henrik Kreidler. Towards a Koopman theory for dynamical systems on completely regular spaces. *Philosophical Transactions of the Royal Society A*, 378(2185):20190617, 2020. https://doi.org/10.1098/rsta.2019.0617.

Tobias Fritz, Tomáš Gonda, Antonio Lorenzin, Paolo Perrone, and Dario Stein. Absolute continuity, supports and idempotent splitting in categorical probability. *Theory and Applications of Categories*, 45(39):1583–1682, 2026. https://www.tac.mta.ca/tac/volumes/45/39/45-39abs.html. Author version: https://arxiv.org/abs/2308.00651v5.

Andrew Geiss and Joseph C. Hardin. Strictly enforcing invertibility and conservation in CNN-based super resolution for scientific datasets. *Artificial Intelligence for the Earth Systems*, 2(1):e210012, 2023. https://doi.org/10.1175/AIES-D-21-0012.1.

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

Paula Harder, Alex Hernandez-Garcia, Venkatesh Ramesh, Qidong Yang, Prasanna Sattegeri, Daniela Szwarcman, Campbell D. Watson, and David Rolnick. Hard-constrained deep learning for climate downscaling. *Journal of Machine Learning Research*, 24(365):1–40, 2023. https://jmlr.org/papers/v24/23-0158.html.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. https://papers.nips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887, 2022. https://proceedings.mlr.press/v162/hoogeboom22a.html.

Olav Kallenberg. *Foundations of Modern Probability*. Volume 99 of Probability Theory and Stochastic Modelling. Springer, Cham, third edition, 2021. https://doi.org/10.1007/978-3-030-61871-1.

Alessio Lapolla and Aljaž Godec. Manifestations of projection-induced memory: General theory and the tilted single file. *Frontiers in Physics*, 7:182, 2019. https://doi.org/10.3389/fphy.2019.00182.

Yin Li, Yueying Ni, Rupert A. C. Croft, Tiziana Di Matteo, Simeon Bird, and Yu Feng. AI-assisted superresolution cosmological simulations. *Proceedings of the National Academy of Sciences*, 118(19):e2022038118, 2021. https://doi.org/10.1073/pnas.2022038118.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6:124, 2025. https://doi.org/10.1038/s43247-025-02042-5.

Emanuele Naldi and Giuseppe Savaré. Weak topology and Opial property in Wasserstein spaces, with applications to gradient flows and proximal point algorithms of geodesically convex functionals. *Atti della Accademia Nazionale dei Lincei. Rendiconti Lincei. Matematica e Applicazioni*, 32(4):725–750, 2021. https://doi.org/10.4171/RLM/955. Authors' preprint: https://arxiv.org/abs/2104.06121.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. https://doi.org/10.1561/2200000073.

Jim Pitman and Marc Yor. A guide to Brownian motion and related stochastic processes. arXiv preprint arXiv:1802.09679, 2018. https://arxiv.org/abs/1802.09679.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. https://gaussianprocess.org/gpml/.

L. C. G. Rogers and J. W. Pitman. Markov functions. *The Annals of Probability*, 9(4):573–582, 1981. https://doi.org/10.1214/aop/1176994363.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

Maybritt Schillinger, Maxim Samarin, Xinwei Shen, Reto Knutti, and Nicolai Meinshausen. EnScale: Temporally-consistent multivariate generative downscaling via proper scoring rules. *Journal of Advances in Modeling Earth Systems*, 18:e2025MS005593, 2026. https://doi.org/10.1029/2025MS005593.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. https://arxiv.org/abs/2010.02502.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*. Volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, 2012. https://doi.org/10.1090/gsm/140.

Cédric Villani. *Optimal Transport: Old and New*. Volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009. https://doi.org/10.1007/978-3-540-71050-9.

Junyang Wang, Jon Cockayne, Oksana Chkrebtii, T. J. Sullivan, and Chris. J. Oates. Bayesian numerical methods for nonlinear partial differential equations. *Statistics and Computing*, 31(5):55, 2021. https://doi.org/10.1007/s11222-021-10030-w.

Wujie Wang, Minkai Xu, Chen Cai, Benjamin K. Miller, Tess Smidt, Yusu Wang, Jian Tang, and Rafael Gomez-Bombarelli. Generative coarse-graining of molecular conformations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23213–23236, 2022. https://proceedings.mlr.press/v162/wang22ag.html.

Geoffrey Wolfer and Shun Watanabe. Geometric aspects of data-processing of Markov chains. *Transactions of Mathematics and Its Applications*, 8(1):tnae001, 2024. https://doi.org/10.1093/imatrm/tnae001.

A PROOF OF THE STRUCTURAL CLASSIFICATION

READING GUIDE AND PROBABILITY NOTATION

The appendices follow the logical order of the main text. Appendix A proves the classification on a compact space, and Appendix B explains the extra estimates needed on an unbounded space. Appendix C derives the continuity consequences. Appendix D checks the example at the regularity boundary. Appendices E and F prove the approximation bound and the rotation construction. Appendices G and H explain the numerical settings, calculations, and results.

We use probability measures to describe random states. If a random variable Y takes values in X, its law $\mathrm{Law}(Y) = \nu$ assigns each measurable set $A \subseteq X$ the probability $\nu(A) = \Pr(Y \in A)$. Here measurable sets are Borel sets, obtained from open sets by complements and countable unions. An integral $\int f(y)\nu(dy)$ is the expectation of $f(Y)$. For a finite set this is just the sum $\sum_y f(y)\Pr(Y = y)$. The notation also covers continuous distributions and mixtures, without requiring a density.

The point mass δ_x gives probability one to the single state x. A statement holds ν-almost surely, or ν-almost everywhere, if its failures form a set of ν-probability zero. The support $\operatorname{supp}\nu$ consists of points whose every open neighborhood has positive probability. A point in the support need not itself have positive probability; a uniform distribution on an interval is a familiar example.

For a transition kernel K, $K(x, A)$ is the probability that the next state lies in A when the input state is x. Its three common uses are

$$x \longmapsto K(x,\cdot), \qquad f \longmapsto Kf, \quad Kf(x) = \int f(y)K(x,dy), \qquad \nu \longmapsto \nu K.$$

The first produces a distribution from a state. The second produces a function of the input state: the expected value of a future observable f. The third averages transition distributions over an uncertain input with law ν. For finite X, these operations are the rows of a stochastic matrix, multiplication of that matrix by a column of function values, and multiplication of a row of probabilities by that matrix. In particular,

$$(KL)(x,A) = \int L(y,A)K(x,dy)$$

means first apply K, then apply L. Testing this measure identity against f gives $K(Lf)$.

A coupling of two laws ν, ν' is any joint choice (Y, Y') having those laws separately. The variables in a coupling need not be independent. The quadratic Wasserstein distance is

$$W_2^2(\nu,\nu') = \inf \mathbb{E}\, d_X(Y,Y')^2,$$

where the infimum is over all couplings. It is the smallest mean squared distance achievable by pairing the two distributions, followed by a square root. The space $\mathcal{P}_2(X)$ consists of probability measures with finite second moment. On compact X, distances are bounded, so every probability measure has this property. We will repeatedly use the elementary consequence

$$\left| \int f\, d\nu - \int f\, d\nu' \right| \le \mathrm{Lip}(f)\, W_2(\nu,\nu').$$

Indeed, for any coupling, the left side is at most $\mathbb{E}|f(Y) - f(Y')|$, then at most $\mathrm{Lip}(f)\mathbb{E}d_X(Y,Y')$, and finally at most $\mathrm{Lip}(f)(\mathbb{E}d_X(Y,Y')^2)^{1/2}$. Minimize over couplings. A function is Lipschitz if $|f(x) - f(y)| \le L d_X(x,y)$ for some finite L; $\mathrm{Lip}(f)$ is the smallest such constant.

These probability conventions and the standard conditional-probability constructions below are treated in Kallenberg (2021). The Wasserstein coupling framework is developed in Villani (2009). We give the calculations needed here explicitly.

Two elementary variance facts explain much of the proof. For a real random variable H with a finite second moment,

$$\mathrm{Var}(H) = \mathbb{E}(H - \mathbb{E}H)^2 = \min_{a\in\mathbb{R}} \mathbb{E}(H-a)^2.$$

The identity follows by expanding $H - a = (H - \mathbb{E}H) + (\mathbb{E}H - a)$. In particular, zero variance means that H equals its mean with probability one. More generally, if $Z \ge 0$ and $\mathbb{E}Z = 0$, then $Z = 0$ with probability one: otherwise some event $\{Z \ge 1/m\}$ would have positive probability and force a positive expectation.

WHAT THE CLASSIFICATION PROOF MUST ESTABLISH

The main distinction is between a fine state and its class label. A label $\mu \in \mathcal{M}$ is itself a probability measure on X. The label distribution ρ_x is a probability measure on the space $\mathcal{M}$ of these labels. Choosing a label and then choosing a state from a distribution associated with that label are two different sampling operations. We first show that reconstruction preserves labels on S, then show that time advances a label deterministically. Only after these steps do we derive the conditional independence of states.

We prove Theorem 1, including initial states whose reconstruction can select different classes. We first derive the decomposition of P into class laws. We then use a backward martingale, a sequence of conditional predictions of a terminal observable, to prove that the class label evolves deterministically. The estimates require neither a differentiable decoder nor assumptions about the domain of an infinitesimal generator.

The deterministic evolution established here needs an argument beyond the static decomposition of P. For example, fast-diffusion limits can average within regions while retaining a stochastic Markov chain between those regions (Bobrowski et al., 2017). Deterministic evolution also has classical operator characterizations: under their stated continuity assumptions, Farkas and Kreidler (2020, Theorems 14 and 17) characterize semigroups induced by continuous semiflows through preservation of constants and products, and their generators through annihilation of constants and the corresponding product rule. The proof below derives determinism directly from the Wasserstein estimates in Eq. (4), without assuming a generator domain on which to verify that rule.

Conventions and assumptions. Let X be a compact metric fine-state space and C a metric coarse-state space. All probability measures on X belong to $\mathcal{P}_2(X)$. Let $R : X \to C$ be L_R-Lipschitz and let $G : C \to \mathcal{P}_2(X)$ be L_G-Lipschitz for W_2. Let $(\phi_t)_{t\geq 0}$ be a deterministic semigroup on C, with $\phi_0 = \mathrm{Id}$. For every $T < \infty$, suppose that there are finite constants L_T, B_T such that

$$\sup_{0\leq t\leq T} \mathrm{Lip}(\phi_t) \leq L_T, \qquad \sup_{x\in X} d_C(\phi_h(Rx), Rx) \leq B_T h \quad (0 < h \leq T). \tag{14}$$

The second bound is needed only for sufficiently small h. For a kernel K on X, write $K(x) = K(x, \cdot)$, $Kf(x) = \int_X f(y)K(x, dy)$ and $(\nu K)(A) = \int_X K(x, A)\nu(dx)$ for measurable $A \subseteq X$. Thus Q_sQ_tf means that the state first makes a transition of length s, then a transition of length t. Define

$$Q_t(x) = G(\phi_t(Rx)) \quad (t > 0), \qquad P(x) = G(Rx), \qquad Q_sQ_t = Q_{s+t} \quad (s, t > 0). \tag{15}$$

The last identity is an assumption for all fine states and all positive times. When used below as notation, Q_0 denotes P, not the identity kernel. Set

$$S = \overline{\bigcup_{x\in X} \operatorname{supp} P(x)}, \qquad \ell(x) = P(x), \qquad \mathcal{M} = P(S). \tag{16}$$

Here S is the closure in X of all reconstruction supports. States $u, v \in S$ have the same class label if $P(u) = P(v)$. The definition of S concerns possible reconstruction values and their limits. It makes no assertion that the process returns to an individual state repeatedly. The equivalence relation $u \sim v$ if $P(u) = P(v)$ is defined on S. The map $[u] \mapsto P(u)$ identifies $S/\!\sim$ with $\mathcal{M} = P(S)$. This is also a homeomorphism: the continuous surjection $P|_S$ maps a compact space onto the Hausdorff space $\mathcal{M}$ and is therefore a quotient map. The metric on $\mathcal{M}$ is the restriction of W_2.

To interpret the topological statement, $[u]$ denotes all states in S with the same law $P(u)$. The quotient $S/\!\sim$ collects these classes as single objects. Giving it the quotient topology means that a collection of classes is open exactly when its inverse image in S is open. The space $\mathcal{M}$ represents each such object by its law, so it has the concrete distance W_2. No additional distribution over $\mathcal{M}$ is introduced in this identification. Compactness makes the identification continuous in both directions: a continuous map from a compact space to a metric space sends closed sets to compact, hence closed, sets.

Step 1: the projection identities. The assumptions imply

$$\sup_{x\in X} W_2(Q_h(x), P(x)) \leq A_T h, \qquad A_T = L_G B_T, \tag{17}$$

$$W_2(Q_t(x), Q_t(y)) \leq K_T d_X(x, y), \qquad K_T = L_G L_R L_T, \quad 0 \leq t \leq T. \tag{18}$$

Here the first estimate follows by applying the decoder bound to the two coarse inputs $\phi_h(Rx)$ and Rx. For the second, apply the bounds for G, ϕ_t, and R in order:

$$W_2(G(\phi_t(Rx)), G(\phi_t(Ry))) \le L_G d_C(\phi_t(Rx), \phi_t(Ry)) \le L_G L_T L_R d_X(x, y).$$

For Lipschitz f, the expectation bound above therefore gives $\|Q_h f - Pf\|_\infty \le \mathrm{Lip}(f) A_T h$, where $\|f\|_\infty = \sup_x |f(x)|$. For general continuous f on compact X, approximate it uniformly by a Lipschitz function f_ϵ. The remaining error is at most

$$\|Q_h f - Pf\|_\infty \le 2\|f - f_\epsilon\|_\infty + \|Q_h f_\epsilon - Pf_\epsilon\|_\infty.$$

First let h tend to zero with the approximating function fixed, and then improve the approximation. Kernels do not increase the supremum norm because they take averages with nonnegative weights summing to one. In particular these kernels are Feller: they map continuous functions to continuous functions. The approximation argument above gives uniform convergence for every continuous test. The coarse semigroup law and (14) also imply continuity in time: for instance, $d_C(\phi_{t+h}(Rx), \phi_t(Rx)) \le L_T B_T h$ on a fixed finite horizon. The analogous left-continuity estimate uses $\phi_t = \phi_{t-h}\phi_h$. Therefore $Q_{t+h} f \to Q_t f$ uniformly when $t > 0$. Passing to the limit $h \downarrow 0$ in $Q_h Q_t = Q_{h+t}$ and $Q_t Q_h = Q_{t+h}$ gives

$$PQ_t = Q_t P = Q_t. \tag{19}$$

For clarity, the two limits in Eq. (19) act on continuous test functions. The first uses $Q_h(Q_t f) \to P(Q_t f)$ because $Q_t f$ is continuous. The second uses $Q_t(Q_h f) \to Q_t(Pf)$ because $Q_h f \to Pf$ uniformly and Q_t takes an average. Both right sides tend to $Q_t f$. Continuous tests determine probability measures, so these equalities of expectations imply the kernel identities.

Then $t \downarrow 0$ gives $P^2 = P$. In particular, every $P(x)$ is a P-invariant probability measure.

Step 2: reconstruction preserves the class label on S. Suppose $g \in C(X)$ is P-harmonic, meaning $Pg = g$. Jensen's inequality and idempotence give

$$V_g = P(g^2) - g^2 \ge 0, \qquad PV_g = 0. \tag{20}$$

The word harmonic here means only that reconstruction preserves the expected value of g: $Pg(u) = g(u)$. It does not refer to a differential equation. For every input u, $V_g(u) = \mathrm{Var}_{P(u)}\, g$ is nonnegative. Moreover,

$$PV_g = P^2(g^2) - P(g^2) = 0.$$

If V_g were positive at a support point, continuity would make it bounded below by a positive number on a neighborhood of that point. That neighborhood has positive probability, contradicting its zero integral. This is why the conclusion holds at every point of S, rather than only outside an exceptional set for one particular initial law. The function V_g is continuous. Its integral vanishes under every $P(x)$, so it vanishes on the support of every $P(x)$, and therefore on S. Consequently

$$\mathrm{Var}_{P(u)}\, g = 0 \quad (u \in S). \tag{21}$$

Choose a countable family $(f_j)_{j \ge 1}$ of Lipschitz functions uniformly dense in $C(X)$, the space of continuous real-valued functions on X. Such a family exists on a compact metric space. Here is a construction explaining this fact. Choose a countable dense set of points, and take all functions that are finite minima of $a_i + L d_X(x, x_i)$, with rational constants a_i and positive rational L. There are countably many such choices, and each is Lipschitz. Continuous functions on a compact space are uniformly continuous; the functions $\inf_y \{f(y) + L d_X(x, y)\}$ approximate f uniformly as $L \to \infty$. A sufficiently fine finite set of y values from the chosen dense set, together with rational approximations of their function values, then approximates each of these infima uniformly. Thus the stated countable family is dense. Each Pf_j is harmonic. For fixed $u \in S$, (21) and a countable intersection give

$$Pf_j(v) = Pf_j(u) \quad \text{for all } j, \qquad P(u)\text{-almost every } v.$$

The countability matters: each test has an exceptional set of probability zero, and the union of countably many such sets still has probability zero. All test identities therefore hold at the same sampled state v. If two measures agree on these tests, uniform approximation makes them agree on every continuous function. Such functions determine the measure: for a closed set E, the continuous functions $\max(0, 1 - m d_X(\cdot, E))$ decrease to its indicator as integers m increase, so bounded

convergence determines its probability. Agreement on closed sets determines Borel probability measures. Hence

$$P(u)\big(\{v \in S : P(v) = P(u)\}\big) = 1, \qquad u \in S. \tag{22}$$

Here $P(x, S) = 1$ for every x follows directly from the definition of S. Also $Q_t P = Q_t$ implies

$$Q_t(x, S) = \int P(y, S) Q_t(x, dy) = 1, \qquad x \in X,\ t > 0. \tag{23}$$

Thus a draw from $P(u)$ preserves the label $P(u)$ when $u \in S$. For an initial point outside S, the label of a draw from $P(x)$ can still be random.

For completeness, the measures in $\mathcal{M}$ are exactly the extreme P-invariant probability measures. Invariance means $\nu P = \nu$. An invariant measure is extreme if it cannot be written as a nontrivial mixture of two distinct invariant probability measures. If $\mu = P(u)$ with $u \in S$ and $\mu = a\nu_1 + (1 - a)\nu_2$ for invariant measures ν_i and $0 < a < 1$, then $\nu_i \ll \mu$. By (22), $P(v) = \mu$ for ν_i-almost every v, and invariance gives $\nu_i = \nu_i P = \mu$. Conversely, every invariant ν is supported on S and satisfies $\nu = \int P(v)\nu(dv)$. If the random measure $P(v)$ is not almost surely constant under ν, a bounded continuous test function and a threshold divide this mixture into two different invariant measures, each obtained by averaging the class laws on its part, with positive mixture weights. Thus an extreme ν has $P(v) = \nu$ almost surely and belongs to $\mathcal{M}$. This is the classical static idempotent-kernel decomposition, here obtained directly under the Feller assumptions.

Here $\nu_i \ll \mu$ means that a μ-probability-zero set also has ν_i-probability zero. In the displayed convex mixture, this follows from nonnegativity and positive weights. For the converse, a nonconstant random label must give a nonconstant value of $Pf_j(V)$ for at least one determining test. Choose a threshold a such that both $Pf_j(V) \le a$ and $Pf_j(V) > a$ have positive probability. Averaging the invariant laws $P(V)$ separately over these two events gives two invariant probability measures with different f_j expectations. Their mixture is ν. This contradicts extremality and completes the converse argument.

Step 3: vanishing dynamical variance. If g is Lipschitz and harmonic and $u \in S$, couple $Y \sim Q_h(u)$ and $Z \sim P(u)$ optimally. By (21), $g(Z) = g(u)$ almost surely, and hence

$$\mathrm{Var}_{Q_h(u)}\, g \le \mathbb{E}|g(Y) - g(u)|^2 \le \mathrm{Lip}(g)^2 A_T^2 h^2. \tag{24}$$

The first inequality uses the variance-minimization identity with the particular constant $g(u)$. For the second, class preservation gives $g(Z) = g(u)$, and Lipschitz continuity gives $|g(Y) - g(Z)| \le \mathrm{Lip}(g) d_X(Y, Z)$. Thus a small Wasserstein displacement from $P(u)$ controls the variance of every such observable. The order h^2 will make the total uncertainty vanish when we subdivide a fixed time interval. Fix $0 < t \le T$, $u \in S$, and a Lipschitz function f. For any integer $n \ge 1$, put $h = t/n$ and let $U_0 = u, U_1, \dots, U_n$ be a Markov chain with transition Q_h. Its terminal law is $Q_t(u)$ by exact composition, independent of n. Define

$$g_k = Q_{t-kh} f \quad (0 \le k < n), \qquad g_n = Pf. \tag{25}$$

Every g_k is harmonic, by (19), and (18) gives $\mathrm{Lip}(g_k) \le K_T \mathrm{Lip}(f)$ uniformly in n, k. The endpoint g_n obeys the same bound because $L_T \ge 1$. The value $g_k(U_k)$ is the conditional prediction of the terminal quantity $Pf(U_n)$. These predictions form a martingale: $Q_h g_k = g_{k-1}$, including the last step, where $Q_h P = Q_h$ is used.

Here is the variance calculation in detail. Let $\mathcal{F}_k$ denote the information in $(U_0, \dots, U_k)$ and put $M_k = g_k(U_k)$. The Markov rule and $Q_h g_k = g_{k-1}$ imply

$$\mathbb{E}[M_k \mid \mathcal{F}_{k-1}] = M_{k-1}.$$

This identity defines a martingale: the current prediction is the expected value of the next prediction. The functions g_k are constructed backward from the terminal observable, while the chain and its information run forward in k. At the terminal step we use Pf, so that the observable measures the class label and is harmonic.

Set $D_k = M_k - M_{k-1}$. For $j < k$, D_j is already known at step $k - 1$, and hence

$$\mathbb{E}[D_j D_k] = \mathbb{E}\{D_j \mathbb{E}[D_k \mid \mathcal{F}_{k-1}]\} = 0.$$

Since M_0 is deterministic and $M_n - M_0 = \sum_k D_k$, expansion of the square gives

$$\mathrm{Var}(M_n) = \sum_k \mathbb{E}D_k^2 = \sum_k \mathbb{E}\{\mathrm{Var}(M_k \mid \mathcal{F}_{k-1})\}.$$

Given the past, U_k has law $Q_h(U_{k-1})$, so the last conditional variance depends only on U_{k-1}. Each g_k has the same Lipschitz bound, which permits the same one-step estimate at all n steps.

All chain states lie in S. Different martingale increments have zero cross expectation, so their variances add. This fact and (24) give

$$\mathrm{Var}_{Q_t(u)}(Pf) = \sum_{k=1}^{n} \mathbb{E}\big[\mathrm{Var}\big(g_k(U_k) \mid U_{k-1}\big)\big] \tag{26}$$

$$\leq nK_T^2 \,\mathrm{Lip}(f)^2 A_T^2 h^2 = K_T^2 \,\mathrm{Lip}(f)^2 A_T^2 \frac{t^2}{n}. \tag{27}$$

The left side does not depend on n, so it is zero. Its mean is $Q_t Pf(u) = Q_t f(u)$. Applying this result to the countable measure-determining tests from Step 2 gives

$$P(v) = Q_t(u) \quad Q_t(u)\text{-almost surely}, \qquad u \in S. \tag{28}$$

This identity describes the label of a sampled state, not the state itself. For each test f_j, zero variance gives $Pf_j(v) = Q_t f_j(u)$ with probability one under $Q_t(u)$. These are the expectations of f_j under the two measures $P(v)$ and $Q_t(u)$. The common countable set of tests makes those measures equal. Individual draws v may still differ while sharing this law as their label. Since $Q_t(u, S) = 1$, this also proves $Q_t(u) \in \mathcal{M}$. The total variance vanishes because the sum of n terms of order $(t/n)^2$ tends to zero. Both the one-step variance estimate and the common Lipschitz bound on the functions g_k are essential.

Step 4: the deterministic flow on classes. If $P(u) = P(v)$, the identity $PQ_t = Q_t$ implies $Q_t(u) = Q_t(v)$. For $\mu = P(u) \in \mathcal{M}$, we may therefore define

$$F_t(\mu) = Q_t(u) = \mu Q_t \quad (t > 0), \qquad F_0(\mu) = \mu. \tag{29}$$

The definition is independent of the representative.

In fact, for any measurable A, $Q_t(u, A) = \int Q_t(v, A) P(u, dv)$ by $PQ_t = Q_t$. This expression depends on u only through $P(u) = \mu$. Furthermore,

$$F_t(F_s(\mu)) = (\mu Q_s)Q_t = \mu Q_{s+t} = F_{s+t}(\mu).$$

The middle equality is exact composition, and Step 3 ensures that every measure appearing as an input of F_t belongs to $\mathcal{M}$. The set $\mathcal{M} = P(S)$ is compact in W_2, since S is compact and P is continuous. First optimally pair the inputs with laws μ and ν, then optimally pair their transition outputs. Integrating these conditional pairings and using (18) yields

$$W_2(F_t\mu, F_t\nu) \leq K_T W_2(\mu, \nu), \qquad 0 < t \leq T. \tag{30}$$

More explicitly, if $\pi(du, dv)$ couples μ and ν, pairing transition outputs conditionally on (u, v) gives a coupling of μQ_t and νQ_t with squared cost at most $K_T^2 \int d_X(u, v)^2 \pi(du, dv)$. Minimize over π. The needed measurable choices of conditional couplings are available on these compact metric spaces; arbitrarily near-optimal choices and a limit give the same estimate. These coupled variables are a proof device for comparing distributions, not an additional sampling rule imposed on the model. At $t = 0$ the Lipschitz constant is one. Moreover, representing $\mu = P(u)$ and using (17) gives

$$W_2(F_h\mu, \mu) \leq A_T h. \tag{31}$$

For example, for $0 \leq s \leq t \leq T$ and $h = t - s$, the semigroup law and the spatial estimate give

$$W_2(F_t\mu, F_s\mu) = W_2(F_s(F_h\mu), F_s\mu) \leq \max(1, K_T) A_T h.$$

Combining this with the bound for changing μ shows that small changes in both time and initial label produce small changes in the output law. This is the meaning of joint continuity.

These inequalities and the semigroup law prove joint continuity of the class flow. In particular it is a jointly continuous deterministic forward evolution with identity at zero, even when the fine kernels do not have that property.

Step 5: all positive-time joint distributions. For a fixed initial state $x \in X$, a draw $V \sim P(x)$ lies in S almost surely. Since $\ell = P|_S : S \to \mathcal{M}$ is continuous, its label has a well-defined distribution ρ_x on $\mathcal{M}$. For every Borel set $D \subseteq \mathcal{M}$,

$$\rho_x(D) = \Pr(P(V) \in D), \qquad V \sim P(x). \tag{32}$$

Specifying this distribution alone does not specify how the label and the generated states are related. We construct them jointly. For $0 < t_1 < \cdots < t_n$, define

$$\widetilde{\mathbb{P}}_x(dv, du_1, \ldots, du_n) = P(x, dv) Q_{t_1}(v, du_1) \prod_{j=2}^{n} Q_{t_j - t_{j-1}}(u_{j-1}, du_j),$$

and set $\Lambda = P(V)$. Integrating out v and using $PQ_{t_1} = Q_{t_1}$ gives

$$Q_{t_1}(x, du_1) \prod_{j=2}^{n} Q_{t_j - t_{j-1}}(u_{j-1}, du_j),$$

which is exactly the joint law of the original rollout from $U_0 = x$. The auxiliary draw V does not replace U_0. These joint constructions are consistent when observation times are removed, by the semigroup law, and therefore give a single shared label for the finite-dimensional distributions.

The displayed joint measure is just a sequential sampling recipe: draw V, then draw U_{t_1} using V, and then use the usual transitions between observation times. Integrating out a variable means taking its expectation without retaining its value. For example, on sets $A_1, \ldots, A_n$, the integral over v replaces $PQ_{t_1}(x, du_1)$ by $Q_{t_1}(x, du_1)$. Every remaining factor is unchanged. This proves equality of the full joint distributions, beyond equality of their separate marginals.

When an observation time is removed, the two adjacent kernels combine by the semigroup law. Removing the earliest positive observation time uses the same law after V. These consistency identities allow a common probability model for all finite lists of times, using the extension theorem for consistent finite-dimensional laws on standard Borel spaces. Compact metric spaces have this measurability property. No continuity of individual fine-state sample paths is needed for this construction.

Because Λ is a function of the auxiliary initial draw V, conditioning on it does not change the later transition rule given the current state. In particular, for $0 < s < t$,

$$\Pr_x(U_t \in A \mid U_s, \Lambda) = Q_{t-s}(U_s, A) \quad \text{almost surely.}$$

This follows from the factors $Q_{t_j - t_{j-1}}$ in the joint construction above, rather than from the marginal distribution of Λ alone.

Given $\Lambda = \mu$, the first state has law $F_{t_1}(\mu)$, because $Q_{t_1}(v) = F_{t_1}(P(v))$ for $v \in S$. Every $F_t(\mu)$ belongs to $\mathcal{M}$, so Eq. (22) gives $P(U_{t_1}) = F_{t_1}(\mu)$ almost surely. The next transition is consequently

$$Q_{t_2 - t_1}(U_{t_1}) = F_{t_2 - t_1}(P(U_{t_1})) = F_{t_2 - t_1}(F_{t_1}(\mu)) = F_{t_2}(\mu).$$

It does not depend on the particular first sample. Repeating this argument proves that the conditional joint law is the product of the $F_{t_j}(\mu)$. Averaging over ρ_x gives

$$\mathrm{Law}_x(U_{t_1}, \ldots, U_{t_n}) = \int_{\mathcal{M}} \bigotimes_{j=1}^{n} F_{t_j}(\mu)\, \rho_x(d\mu). \tag{33}$$

For a concrete interpretation, its value on a rectangle $A_1 \times \cdots \times A_n$ is

$$\Pr_x(U_{t_1} \in A_1, \ldots, U_{t_n} \in A_n) = \int_{\mathcal{M}} \prod_{j=1}^{n} F_{t_j}(\mu)(A_j)\, \rho_x(d\mu).$$

The product inside the integral expresses independence after a label is fixed. The integral then averages those probabilities over the possible initial labels. In a finite label space it is an ordinary weighted sum. The notation $\bigotimes$ in Eq. (33) denotes the product probability measure whose rectangle probabilities are these products. Conditional distributions given $\Lambda = \mu$ are understood as regular conditional

probabilities, valid for ρ_x-almost every μ, including when individual labels have probability zero. The product kernels above provide such a version. Thus no division by $\Pr(\Lambda = \mu)$ is needed.

Concretely, a conditional kernel $K(\mu, A)$ is a version of $\Pr(U_t \in A \mid \Lambda = \mu)$ if $\Pr(U_t \in A, \Lambda \in D) = \int_D K(\mu, A)\rho_x(d\mu)$ for all measurable D. Here we have explicitly constructed $K(\mu, A) = F_t(\mu)(A)$. This integral identity continues to make sense when each individual label has probability zero.

The conditional product law gives the first equality in Eq. (8). To obtain the conditional distribution given only Λ, the tower property of conditional expectation gives

$$\begin{aligned}\Pr_x(U_t \in A \mid \Lambda) &= \mathbb{E}_x\Big[\Pr_x(U_t \in A \mid \Lambda, U_s) \mid \Lambda\Big] \\ &= \mathbb{E}_x[F_t(\Lambda)(A) \mid \Lambda] = F_t(\Lambda)(A).\end{aligned}$$

The last equality holds because $F_t(\Lambda)(A)$ depends only on the given label.

For $x \in S$, Eq. (22) gives $\rho_x = \delta_{P(x)}$. This point mass is on the label space: $\rho_x(D) = 1$ if $P(x) \in D$ and zero otherwise. The state law $P(x)$ is a measure on X, giving its state class $[x]$ probability one, and need not itself be a point mass. For a fixed $x \notin S$, ρ_x may mix several labels. Conditional independence holds for either type of label law. After averaging over a mixture, the states can become dependent, as the following example shows. The classification does not establish that arbitrary initial-label distributions ρ_x and evolutions F_t can be implemented by the architecture.

A fixed initial state with a random class label. Let $X = C = \{a, b, c\}$ carry the discrete metric, take $R = \mathrm{Id}$ and $\phi_t = \mathrm{Id}$, and define $G = P$ by

$$P(a) = \delta_a, \qquad P(b) = \delta_b, \qquad P(c) = \tfrac{1}{2}\delta_a + \tfrac{1}{2}\delta_b.$$

Then $Q_t = P$ for every $t > 0$, and direct substitution gives $P^2 = P$, so composition is exact. The finite spaces are compact, every map is Lipschitz, and the coarse flow is stationary. Thus the assumptions of Theorem 1 hold. Here $S = \{a, b\}$ and $\mathcal{M} = \{\delta_a, \delta_b\}$. Starting from the fixed state $x = c \notin S$, a draw from $P(c)$ selects either class with probability $1/2$:

$$\rho_c = \tfrac{1}{2}\delta_{\delta_a} + \tfrac{1}{2}\delta_{\delta_b}.$$

Each selected class remains unchanged under F_t. Given $\Lambda = \delta_a$, every positive-time state equals a. Given $\Lambda = \delta_b$, every such state equals b. Consequently, for $0 < s < t$,

$$\Pr_c(U_s = a) = \Pr_c(U_t = a) = \tfrac{1}{2}, \qquad \Pr_c(U_s = a, U_t = a) = \tfrac{1}{2} \neq \tfrac{1}{4}.$$

The dependence comes from the shared random label, although $U_0 = c$ is fixed. In this example, observing $U_s = a$ determines the label and gives $\Pr_c(U_t = a \mid U_s = a) = 1$. More generally, an observation may only change the probabilities of the possible labels. This does not contradict independence conditional on the label.

Exact composition without conditional independence. Standard Brownian motion $(B_t)_{t \geq 0}$ started at $B_0 = 0$ has transitions satisfying Eq. (2). Its zero-time kernel is $P(b) = \delta_b$, so its initial class label is fixed. For $0 < s < t$,

$$\mathrm{Law}(B_t \mid B_s = b) = \mathcal{N}(b, t - s),$$

where $\mathcal{N}(b, t - s)$ denotes the normal law with mean b and variance $t - s$ (Pitman and Yor, 2018, Section 4.1). An observed position b therefore changes the distribution predicted at time t. Averaging over the possible positions at time s still gives the same endpoint distribution as a single transition from zero to t. Exact composition alone thus does not imply the conditional independence in Eq. (8).

For this example, the composition identity can be checked directly. Write $B_t = B_s + (B_t - B_s)$, where the increment is independent of B_s and has law $\mathcal{N}(0, t - s)$. The sum has law $\mathcal{N}(0, t)$, so averaging the conditional normal law gives the correct endpoint law. Nevertheless $\mathbb{E}[B_t \mid B_s] = B_s$, whereas $\mathbb{E}B_t = 0$. The small-time Wasserstein displacement from δ_b is $\sqrt{h}$, because its square equals the increment variance h. Brownian motion therefore does not have the vanishing $W_2/\sqrt{h}$ ratio used in the classification argument. The compact example in Appendix D isolates that same temporal boundary within a compact state space.

Exact fibers as a special case. The notation $R_{\#}\nu$ means the law of RY when Y has law ν: $(R_{\#}\nu)(B) = \nu(\{y : Ry \in B\})$. Thus $R_{\#}G(c) = \delta_c$ says exactly that a draw $U \sim G(c)$ satisfies $RU = c$ with probability one. This is a condition on every reconstruction under consideration. Suppose in addition that $R_{\#}G(c) = \delta_c$ for the coarse states under consideration. For $s, t > 0$, exact fiber consistency gives $Rv = \phi_s(Rx)$ for $G(\phi_s(Rx))$-almost every v. Thus, for any measurable $A \subseteq X$,

$$\begin{aligned}(Q_sQ_t)(x, A) &= \int_X G(\phi_t(Rv), A)\, G(\phi_s(Rx), dv)\\ &= G(\phi_t(\phi_s(Rx)), A)\\ &= G(\phi_{s+t}(Rx), A) = Q_{s+t}(x, A).\end{aligned}$$

The last line uses the coarse semigroup. For the independent-reconstruction baseline in Section 5, fix $U_0 = x$, put $c_0 = Rx$, and take $0 = t_0 < t_1 < \cdots < t_n$. Induction using exact recovery gives $RU_{t_j} = \phi_{t_j}(c_0)$ almost surely. The Markov sampling rule therefore gives, for $j \geq 1$,

$$\mathrm{Law}(U_{t_j} \mid U_{t_0}, \ldots, U_{t_{j-1}}) = G\big(\phi_{t_j - t_{j-1}}(RU_{t_{j-1}})\big) = G(\phi_{t_j}(c_0)).$$

This conditional law does not depend on the sampled past, so the joint law is $\bigotimes_j G(\phi_{t_j}(c_0))$. Both this conclusion and exact composition follow without a decoder Lipschitz assumption. In the classification's label representation, $P(y) = G(c_0)$ for $G(c_0)$-almost every y, so $\rho_x = \delta_{G(c_0)}$. Exact recovery is an additional condition for this special case, not an assumption of Theorem 1. Without exact recovery, composition remains a separate requirement. Reusing one noise realization across times specifies a different sampling rule, whose state and transition must be analyzed separately.

Proof of Proposition 2. We explain why only the common support needs spatial Lipschitz control in the abstract criterion. Let X be compact and let Q_t be the Feller positive-time semigroup in that proposition. Uniform convergence $Q_h(x) \to P(x)$ in W_2, for $x \in X$, implies uniform convergence $Q_h f \to Pf$ for continuous f. Continuity of the kernels and their positive-time dependence therefore justify Step 1 without any global spatial Lipschitz bound. In particular, P is continuous and $P^2 = P$, $PQ_t = Q_tP = Q_t$. Step 2 uses only continuous harmonic functions and compactness, so it also remains valid unchanged. It yields both class purity on S and the fact that $Q_t(x)$ and $P(x)$ are supported on S for every ambient x.

Write $K_T^S = \sup_{0 \leq s \leq T} \mathrm{Lip}_{W_2}(Q_s|_S) < \infty$, with $Q_0 = P$. For a Lipschitz test f on X, the backward functions $g_k = Q_{t-kh}f$ and $g_n = Pf$ are continuous on X and have Lipschitz constant at most $K_T^S \,\mathrm{Lip}(f)$ on S. Their regularity outside S is not needed: the chain starts in S when proving deterministic class evolution, and both measures in the one-step coupling are supported there. If

$$\sup_{u \in S} W_2(Q_h(u), P(u)) \leq \omega_T(h), \qquad \omega_T(h)^2/h \longrightarrow 0,$$

the same argument as (24) consequently gives

$$\mathrm{Var}_{Q_t(u)}(Pf) \leq (K_T^S)^2\, \mathrm{Lip}(f)^2 n \omega_T(t/n)^2 \longrightarrow 0, \qquad u \in S.$$

The countable tests, and then Steps 4 and 5, prove all the asserted structural conclusions, including the mixture for an arbitrary initial $x \in X$. For the flow, lifting input couplings only uses measures supported on S, so its spatial Lipschitz bound is K_T^S, and its small-time displacement is at most $\omega_T(h)$. Compactness of $P(S)$ and these bounds give the same joint continuity. Thus neither a rate for $Q_h(x) \to P(x)$ outside S nor a spatial Lipschitz bound outside S is needed.

The decisive arithmetic is $n\omega_T(t/n)^2 = t\,\omega_T(h)^2/h$. The notation $\omega_T(h) = o(\sqrt{h})$ means $\omega_T(h)/\sqrt{h} \to 0$, which makes this expression vanish. A bound $O(\sqrt{h})$ only says that this ratio stays bounded; it need not tend to zero. This explains exactly where the strict small-time condition enters the proof.

A.1 An ODE example for the coarse-flow assumptions

Let $C \subset \mathbb{R}^m$ be compact with the Euclidean metric. Let b be a locally Lipschitz vector field on an open neighborhood of C, with

$$\|b(c) - b(c')\| \leq L\|c - c'\| \quad (c, c' \in C), \qquad M := \sup_{c \in C} \|b(c)\| < \infty.$$

For each initial coarse state $c \in C$, consider

$$\dot{z}(t) = b(z(t)), \qquad z(0) = c.$$

The ODE is autonomous because b has no explicit time dependence. Suppose C is forward invariant: every such solution stays in C for as long as it exists. Compactness then allows continuation for every $t \geq 0$ (Teschl, 2012, Section 2.6). Denote the solution by $z(t; c)$ and define its solution map by

$$\phi_t(c) := z(t; c) \in C.$$

Uniqueness and autonomy give $\phi_0 = \mathrm{Id}$ and $\phi_{s+t} = \phi_t \circ \phi_s$. This map supplies the coarse evolution. Given R and G, the fine transition is then $Q_t(x, \cdot) = G(\phi_t(Rx))$ for $t > 0$, as in Eq. (1).

For two initial states, the integral equation gives

$$\|\phi_t(c) - \phi_t(c')\| \leq \|c - c'\| + L \int_0^t \|\phi_r(c) - \phi_r(c')\| \, dr.$$

Grönwall's inequality therefore yields $\|\phi_t(c) - \phi_t(c')\| \leq e^{Lt}\|c - c'\|$ (Teschl, 2012, Theorem 2.8). Integrating the speed gives

$$\|\phi_h(c) - c\| \leq \int_0^h \|b(\phi_r(c))\| \, dr \leq Mh.$$

Thus $L_T \leq e^{LT}$ and Eq. (3) holds with $B_T = M$. For $s, t \in [0, T]$, the same estimates give

$$\|\phi_t(c) - \phi_s(c')\| \leq e^{LT}\|c - c'\| + M|t - s|,$$

which proves joint continuity. These estimates verify the coarse-flow assumptions. Exact composition of the reconstructed kernels in Eq. (2) remains a separate requirement.

B AN EXTENSION TO UNBOUNDED EUCLIDEAN STATE SPACES

On a compact space all states have bounded norm, so estimates involving second moments are automatic. On $\mathbb{R}^{d_f}$ a rare sample can be arbitrarily large. We need a bound on its contribution to expectations at every intermediate step. A second issue is that a continuous image of a closed unbounded set need not be closed. The proof below addresses these two issues separately; the variance argument itself is the one from Appendix A.

Proposition 6. *Let $X = \mathbb{R}^{d_f}$ and $C = \mathbb{R}^{d_c}$. Assume R is linear, $G : C \to \mathcal{P}_2(X)$ is globally W_2-Lipschitz with constant L_G, and $G(0)$ has finite second moment. Let ϕ be a jointly continuous deterministic semigroup such that, for every finite T,*

$$\sup_{0 \leq t \leq T} \mathrm{Lip}(\phi_t) \leq L_T < \infty, \qquad \|\phi_h(c) - c\| \leq B_T h(1 + \|c\|), \quad 0 < h \leq T. \tag{34}$$

Assume exact positive-time composition of $Q_t = G\phi_t R$. Then the projection identities, class purity on S, deterministic class flow, and full representation (33) *all remain valid. The class space $\mathcal{M} = P(S)$ is closed in $\mathcal{P}_2(X)$ and is Polish: its topology admits a complete separable metric. It need not be compact. The flow is jointly continuous and uniformly spatially Lipschitz on every finite time horizon.*

Proof. We give the additional arguments that replace compactness.

Moment control. The second-moment norm of a law ν is $W_2(\nu, \delta_0) = (\int \|y\|^2 \nu(dy))^{1/2}$, because pairing with the point mass leaves no choice for the second variable. Apply the triangle inequality through $G(0)$ to bound this quantity for $G(c)$. Put $a_0 = (\int \|y\|^2 G(0, dy))^{1/2}$. The Wasserstein triangle inequality gives

$$\left(\int \|y\|^2 G(c, dy) \right)^{1/2} \leq a_0 + L_G\|c\|.$$

Since $\|\phi_t(c)\| \leq L_T\|c\| + B_T T$, there is a finite constant M_T such that

$$\sup_{0 \leq t \leq T} \int \|y\|^2 Q_t(u, dy) \leq M_T(1 + \|u\|^2), \qquad Q_0 = P. \tag{35}$$

The spatial estimate (18) remains valid with $L_R = \|R\|$, whereas the small-time estimate becomes

$$W_2(Q_h(u), P(u)) \le L_G B_T h(1 + \|Ru\|). \tag{36}$$

The estimate for ϕ_t follows from

$$\|\phi_t(c)\| \le \|\phi_t(c) - \phi_t(0)\| + \|\phi_t(0)\| \le L_T\|c\| + B_T t.$$

The bound at zero uses Eq. (34) with $c = 0$ and $\phi_0(0) = 0$. Squaring the resulting affine growth estimate and using $(a + b)^2 \le 2a^2 + 2b^2$ gives Eq. (35). Linearity of R gives $\|Ru\| \le \|R\|\|u\|$.

Limits and class preservation. All kernels map bounded continuous functions to bounded continuous functions. For such f, pointwise weak convergence $Q_h(u) \to P(u)$, bounded convergence inside a fixed kernel, and continuity of $Q_t f$ justify the limits in both compositions. Thus $P^2 = P$ and $PQ_t = Q_t P = Q_t$ without uniform convergence on the entire state space. In place of a uniformly dense subset of $C_b(X)$, use a countable family of bounded Lipschitz tests whose integrals determine a probability measure. For example, compactly supported Lipschitz functions dense in $C_0(X)$ suffice. Here $C_b(X)$ denotes bounded continuous functions and $C_0(X)$ those that tend to zero at infinity. For a test f in this family, $g = Pf$ is bounded and Lipschitz. The function $P(g^2) - g^2$ is bounded, continuous, nonnegative, and has zero P-average. The support argument from Step 2 therefore proves the same class-purity identity on S and the common-support property. The characterization of extreme invariant laws is unchanged because all mixture identities used to prove it involve bounded tests.

For example, $Q_t(Q_h f)(u) \to Q_t(Pf)(u)$ follows by bounded convergence under the fixed measure $Q_t(u)$. In the other order, $Q_h(Q_t f)(u) \to P(Q_t f)(u)$ follows from weak convergence against the bounded continuous function $Q_t f$. A weak limit means convergence of such integrals. This explains why uniform convergence over all $u \in \mathbb{R}^{d_f}$ is unnecessary.

Summing the weighted variances. For bounded Lipschitz harmonic g and $u \in S$, optimal coupling now gives

$$\mathrm{Var}_{Q_h(u)}\, g \le \mathrm{Lip}(g)^2 L_G^2 B_T^2 h^2 (1 + \|Ru\|)^2. \tag{37}$$

Fix $t \le T$ and construct the backward martingale of (25) from $u \in S$. For $k \ge 2$, exact composition gives $\mathcal{L}(U_{k-1}) = Q_{(k-1)h}(u)$, so (35) controls the moments in (37), uniformly in n and k. For the first increment $U_0 = u$ is deterministic: its law is δ_u, not $Q_0(u) = P(u)$, and its contribution is bounded directly. It follows that

$$\sup_{n\ge 1} \max_{0\le k\le n} \mathbb{E}(1 + \|RU_k\|)^2 \le C_T(1 + \|u\|^2).$$

The backward functions have a common Lipschitz bound independent of n. Summing their conditional variances yields

$$\mathrm{Var}_{Q_t(u)}(Pf) \le C_T\, \mathrm{Lip}(f)^2 (1 + \|u\|^2) \frac{t^2}{n} \longrightarrow 0. \tag{38}$$

The countable test family proves deterministic class evolution as before.

The factor depending on the current state causes no accumulation because its expectation is uniformly bounded: $(1 + \|RU_k\|)^2 \le 2 + 2\|R\|^2\|U_k\|^2$. For $k \ge 1$, exact composition identifies the state law with $Q_{kh}(u)$, whose second moment is controlled by Eq. (35). The deterministic step $k = 0$ has the same growth bound directly. Summing n contributions of order h^2 again yields $nh^2 = t^2/n$.

A closed, measurable label space. It remains to justify the topology of the class space: a continuous image of a noncompact closed set need not be closed or even Borel. Set $L_P = L_G\|R\|$ and, for $\mu \in \mathcal{P}_2(X)$, define

$$D(\mu) = \left(\int W_2^2(P(u), \mu)\, \mu(du) \right)^{1/2}. \tag{39}$$

To understand D, draw $U \sim \mu$ and compare its label $P(U)$ with the proposed label μ. The number $D(\mu)$ is the root-mean-square Wasserstein discrepancy. It is zero exactly when those labels agree with probability one. A zero set of a continuous real function is closed, so showing that D is continuous

will give a useful description of $\mathcal{M}$. The moment bound makes this finite. If (U, V) optimally couples μ, ν, the triangle inequality and Minkowski's inequality give

$$D(\mu) \le \big(\mathbb{E}W_2^2(P(U), P(V))\big)^{1/2} + D(\nu) + W_2(\nu, \mu) \le D(\nu) + (L_P + 1)W_2(\mu, \nu).$$

Interchanging μ, ν shows that D is $(L_P + 1)$-Lipschitz. Class purity implies $D(P(u)) = 0$ for $u \in S$. Conversely, $D(\mu) = 0$ implies $P(u) = \mu$ for μ-almost every u. Then $\mu P = \mu$, whence $\mu(S) = 1$, and some such $u \in S$ satisfies $P(u) = \mu$. Therefore

$$\mathcal{M} = \{\mu \in \mathcal{P}_2(X) : D(\mu) = 0\}. \tag{40}$$

This is closed in the Polish space $\mathcal{P}_2(X)$.

A Polish space has a complete metric and a countable dense subset. Completeness controls limits of Cauchy sequences, and separability supplies countable approximations. The Wasserstein space $\mathcal{P}_2(\mathbb{R}^{d_f})$ has these properties, and so does its closed subset in Eq. (40). In particular, its Borel structure supports the conditional-probability construction used for the label.

Continuity and the joint representation. The formula $F_t(\mu) = \mu Q_t$ again gives $W_2(F_t\mu, F_t\nu) \le K_T W_2(\mu, \nu)$ for $0 < t \le T$. To obtain a temporal bound that does not depend on a chosen representative, couple the mixtures μQ_h and $\mu P = \mu$ pointwise:

$$W_2^2(F_h\mu, \mu) \le L_G^2 B_T^2 h^2 \int (1 + \|Ru\|)^2 \mu(du). \tag{41}$$

Together with the moment estimate and the semigroup law, this proves joint continuity. The initial-label law ρ_x defined in Eq. (32) is Borel on the Polish class space, and the conditional construction in Step 5 proves the full finite-dimensional representation without alteration. □

The coarse-flow regularity assumptions in Proposition 6 hold for the flow of a globally Lipschitz vector field. To see this, let the vector field be b with Lipschitz constant L and put $b_0 = \|b(0)\|$. The integral equation and Grönwall's inequality give $\mathrm{Lip}(\phi_t) \le e^{Lt}$ and

$$\|\phi_h(c) - c\| \le \int_0^h (b_0 + L\|\phi_r(c)\|)\, dr \le h e^{LT}(b_0 + L\|c\|), \qquad h \le T.$$

For the second inequality, the growth estimate $\|\phi_r(c)\| \le e^{Lr}\|c\| + b_0(e^{Lr} - 1)/L$ is used when $L > 0$; the case $L = 0$ follows by direct integration. These bounds give the required constants. The decoder's global Wasserstein bound and exact composition remain separate requirements. Applying the result to a trained decoder requires the stated global estimates. Local regularity and finite moments alone do not suffice.

C Continuity and support-dimension consequences

The first question is simple to state. If the model draws independently at two nearby positive times after its initial label is fixed, can the two states nevertheless become close in mean square? The calculation below shows that the only way is for the conditional sampling variance to vanish. The final part asks a different question: what follows if a transition over a very short time must recover every specified starting state?

We prove Corollary 3, explain how to average over the initial labels, and then derive a geometric consequence of convergence to each starting state. Assume the hypotheses of Theorem 1 on a compact fine space $X \subseteq \mathbb{R}^{d_f}$, or those of Proposition 6 on Euclidean space. For a probability measure ν with finite second moment, write

$$m(\nu) = \int y\, \nu(dy), \qquad v(\nu) = \int \|y - m(\nu)\|^2 \nu(dy).$$

Fix $U_0 = x$. First condition on the initial label $\Lambda = \mu$ and put $\mu_t = F_t(\mu)$. Write $\mathbb{E}_\mu$ for this conditional expectation. The product representation in Eq. (7) makes U_t and U_{t+h} independent given the label, for every $t > 0$ and $h > 0$. Hence

$$\mathbb{E}_\mu\langle U_{t+h}, U_t\rangle = \langle m(\mu_{t+h}), m(\mu_t)\rangle.$$

For two independent random vectors Y, Z with finite second moments, expand

$$\|Y - Z\|^2 = \|Y\|^2 + \|Z\|^2 - 2\langle Y, Z\rangle.$$

Independence makes $\mathbb{E}(Y_i Z_i) = \mathbb{E}Y_i \mathbb{E}Z_i$ for each coordinate, so the expected inner product is the inner product of the means. Writing each second moment as variance plus squared mean then gives

$$\mathbb{E}\|Y - Z\|^2 = v(\mathrm{Law}(Y)) + v(\mathrm{Law}(Z)) + \|\mathbb{E}Y - \mathbb{E}Z\|^2.$$

We use this identity conditional on a fixed label; it would generally be false without that conditioning.

Expanding the squared difference and using $\mathbb{E}_\mu\|U_t\|^2 = v(\mu_t) + \|m(\mu_t)\|^2$ gives

$$\mathbb{E}_\mu\|U_{t+h} - U_t\|^2 = v(\mu_{t+h}) + v(\mu_t) + \|m(\mu_{t+h}) - m(\mu_t)\|^2 \longrightarrow 2v(\mu_t). \tag{42}$$

The limit follows from W_2-continuity of the class flow, which gives convergence of both the means and the variances. One can see the continuity of the mean and variance directly from a coupling. If $W_2(\nu_j, \nu) \to 0$, choose $Y_j \sim \nu_j$ and $Y \sim \nu$ with $\mathbb{E}\|Y_j - Y\|^2 \to 0$. Then

$$\|\mathbb{E}Y_j - \mathbb{E}Y\| \le (\mathbb{E}\|Y_j - Y\|^2)^{1/2}, \qquad \left|(\mathbb{E}\|Y_j\|^2)^{1/2} - (\mathbb{E}\|Y\|^2)^{1/2}\right| \le (\mathbb{E}\|Y_j - Y\|^2)^{1/2}.$$

The latter is the triangle inequality for the square-integrable random vectors. Both moments therefore converge, as does their difference defining the variance.

The law of total expectation now yields

$$\mathbb{E}_x\|U_{t+h} - U_t\|^2 = \int_{\mathcal{M}} \mathbb{E}_\mu\|U_{t+h} - U_t\|^2 \, \rho_x(d\mu).$$

On compact X, the squared increments are bounded by $\mathrm{diam}(X)^2$, so dominated convergence permits passing the limit through this integral. In the Euclidean extension, the required domination follows from (35) and

$$\int_{\mathcal{M}} \int \|y\|^2 \mu(dy)\rho_x(d\mu) = \int \|y\|^2 P(x, dy) < \infty, \tag{43}$$

In more detail, for a fixed label μ the moment bound for μQ_r is at most $M_T(1 + \int \|u\|^2\mu(du))$, uniformly for the times in question. The elementary inequality $\|y - z\|^2 \le 2\|y\|^2 + 2\|z\|^2$ gives the same type of bound for the conditional squared increment. Its average over ρ_x is finite by Eq. (43). This is an integrable bound independent of small h, as required for dominated convergence. The moment identity itself follows by first drawing $V \sim P(x)$ and then drawing from $P(V)$: its unconditional law is $P^2(x) = P(x)$. Consequently,

$$\lim_{h\downarrow 0} \mathbb{E}_x\|U_{t+h} - U_t\|^2 = 2\int_{\mathcal{M}} v(F_t(\mu))\rho_x(d\mu),$$

which proves Eq. (10). If $\rho_x = \delta_{\mu_0}$, evaluating the integral at μ_0 recovers the fixed-label limit. Every label $\mu_0 \in \mathcal{M} = P(S)$ can be obtained in this way by choosing $x \in S$ with $P(x) = \mu_0$.

Mean-square continuity at time t requires this limit to be zero. Since the integrand is nonnegative, $v(F_t(\mu)) = 0$ except on a set of labels with probability zero under ρ_x. For each remaining label,

$$\int \|y - m(\mu_t)\|^2 \mu_t(dy) = 0.$$

The squared distance is nonnegative, so it must be zero with probability one. Consequently $\mu_t = \delta_{m(\mu_t)}$. Thus a label Λ drawn from ρ_x gives a point-mass distribution $F_t(\Lambda)$ with probability one, as stated in the corollary. This does not require the label distribution ρ_x itself to be a point mass.

For example, one label could specify the constant path at a and another the constant path at b. A random choice between those two labels produces a process that is random and mean-square continuous on positive times, although each conditional distribution is a point mass. This example makes no assertion about continuity at the prescribed initial state $U_0 = x$. Thus the conclusion permits randomness in the initial choice and removes subsequent sampling uncertainty within a chosen label. If the fixed-label variance at time t were instead $v > 0$, two independent draws from

nearly the same law would have expected squared separation tending to $2v$, however close their times are.

If mean-square continuity is required at every positive time, first apply this statement at positive rational times. The union of their exceptional label sets still has probability zero. Continuity of F_t and closedness of the set of Dirac laws then extend the conclusion to all $t > 0$ for every label outside that union. This assertion concerns labels selected according to ρ_x only. On a compact fine space, stochastic continuity also implies mean-square continuity because the squared increments are uniformly bounded.

Stochastic continuity means that $\Pr(\|U_{t+h} - U_t\| > a) \to 0$ for every $a > 0$. If the diameter is D, split the expectation according to this event to obtain

$$\mathbb{E}\|U_{t+h} - U_t\|^2 \le a^2 + D^2 \Pr(\|U_{t+h} - U_t\| > a).$$

Let $h \to 0$, then $a \to 0$. For the simultaneous statement over all positive times, countability is equally important: first remove the probability-zero exceptional sets for rational times, then use continuity. The set of Dirac laws is closed in W_2, since $W_2(\delta_y, \delta_z) = \|y - z\|$ makes any convergent sequence of such laws correspond to a Cauchy sequence of points.

Recovery of the initial state and support dimension. If additionally $Q_h(u) \Rightarrow \delta_u$ as $h \downarrow 0$ for every $u \in S$, meaning that the probability of leaving any fixed neighborhood of u tends to zero, then

$$G(Ru) = \delta_u, \qquad \|u - v\| \le L_G \, \|Ru - Rv\| \quad (u, v \in S). \tag{44}$$

Consequently $R|_S$ is one-to-one and both it and its inverse on $R(S)$ are Lipschitz. If $C \subseteq \mathbb{R}^{d_c}$, the Hausdorff dimension of S satisfies $\dim_H S \le d_c$. Initial randomness among these states remains possible.

This additional requirement concerns $t = 0$. It was not needed for the positive-time conclusion above. If $Ru = Rv$ but $u \ne v$, the same decoder input would have to reconstruct two different prescribed point masses. The inverse bound makes this obstruction quantitative: states that the coarse map separates only slightly cannot be far apart on a support where exact initial recovery is required.

To prove this, start from the convergence assumption at every $u \in S$:

$$Q_h(u) \Longrightarrow \delta_u \quad (h \downarrow 0). \tag{45}$$

The already established limit $Q_h(u) \to P(u)$ in W_2 then forces $P(u) = \delta_u$. For $u, v \in S$,

$$\|u - v\| = W_2(G(Ru), G(Rv)) \le L_G \|Ru - Rv\|. \tag{46}$$

Consequently $R|_S$ is injective and has a Lipschitz inverse $J : R(S) \to S$. Since R is also Lipschitz, S and $R(S)$ are bi-Lipschitz equivalent. If $C \subseteq \mathbb{R}^{d_c}$,

$$\dim_H S = \dim_H R(S) \le d_c. \tag{47}$$

Hausdorff dimension measures the size of a set through covers by sets of decreasing diameter and applies also to irregular subsets. A bi-Lipschitz map changes all distances by bounded factors, so it preserves this dimension. More explicitly, cover a set E by sets E_i with diameter at most δ, sum $\operatorname{diam}(E_i)^q$, and minimize over such covers. Letting $\delta \downarrow 0$ defines its q-dimensional Hausdorff measure. The Hausdorff dimension is the threshold in q above which that measure vanishes. An L-Lipschitz map sends each cover to a cover with diameters at most $L \operatorname{diam}(E_i)$, so it cannot increase this dimension. Applying this observation to both $R|_S$ and its inverse proves equality of the two dimensions. Finally, a bounded cube in $\mathbb{R}^{d_c}$ can be covered by order δ^{-d_c} cubes of diameter proportional to δ; the sum tends to zero for $q > d_c$. Countably many bounded cubes cover all of $\mathbb{R}^{d_c}$, giving the stated upper bound.

It is an exact-support restriction under the displayed continuity requirement, not an approximation-error lower bound. In particular, graphs with increasingly large Lipschitz constants can approximate laws supported on higher-dimensional sets in Wasserstein distance.

D A STOCHASTIC PROCESS AT THE TEMPORAL BOUNDARY

We prove Proposition 4. Proposition 2 requires the temporal modulus on the common support to satisfy

$$\sup_{u \in S} W_2(Q_h(u), P(u)) \le \omega(h), \qquad \frac{\omega(h)^2}{h} \longrightarrow 0, \tag{48}$$

The construction below satisfies a temporal bound of order $\sqrt{h}$ and all the other assumptions of Proposition 2, but its class evolution is stochastic. It therefore disproves the proposition with $o(\sqrt{h})$ replaced by $O(\sqrt{h})$.

The construction uses

$$X = [-1,1] \times [0,1] \times [-1,1], \qquad C = [-1,1] \times [0,1],$$
$$R(a,r,z) = (a,r), \qquad \phi_t(a,r) = (a, e^{-t}r). \tag{49}$$

The coordinates have distinct jobs. The sign parameter a and the damping coordinate r form the coarse state; z is discarded. The coarse flow leaves a unchanged and decreases r deterministically. The decoder below turns their product into the mean of a random sign and resets the second coordinate to one. Since $|ra| \le 1$, both sign probabilities lie between zero and one and sum to one. Let $G(a,r) = \mathrm{Law}(A,1,0)$, where $\Pr(A=1) = (1+ra)/2$ and $\Pr(A=-1) = (1-ra)/2$. Then $Q_t(a,r,z) = G(a,e^{-t}r)$ has sign mean $e^{-t}ra$. After a transition of length s, the state is $(A,1,0)$ with $\mathbb{E}A = e^{-s}ra$. After another transition of length t, the conditional sign mean is $e^{-t}A$, and its unconditional mean is $e^{-(s+t)}ra$. Since a probability law on two fixed points is determined by its sign mean, this proves $Q_sQ_t = Q_{s+t}$ for every fine starting state. Taking either time to zero also gives $P^2 = P$.

The common support is $S = \{u_-, u_+\}$, where $u_\pm = (\pm 1, 1, 0)$, and $P(u_\pm) = \delta_{u_\pm}$. In the order (u_-, u_+), the restricted transition matrix is

$$Q_t|_S = \frac{1}{2}\begin{pmatrix} 1+e^{-t} & 1-e^{-t} \\ 1-e^{-t} & 1+e^{-t} \end{pmatrix}. \tag{50}$$

This is the symmetric continuous-time Markov chain with flip rate $1/2$. Its class dynamics are stochastic.

Starting at either support point, the probability of switching to the other is $(1-e^{-t})/2$. For small t this is $t/2 + o(t)$, which explains the flip rate. Both labels are Dirac laws, so a random switch between the two points is also a random switch between classes. At every $t > 0$ both transition probabilities are positive, which rules out deterministic class evolution.

The squared distance between the two output atoms is 4. Transporting the excess mass between them therefore gives exactly

$$W_2^2(G(a,r), G(a',r')) = 2|ra - r'a'| \le 2\sqrt{2}\|(a,r) - (a',r')\|. \tag{51}$$

To check the transport calculation, let p, p' be the probabilities of u_+. A coupling can leave the common masses at both points in place, but must move the excess $|p-p'|$ a distance two. No other transport is possible on two points. Thus $W_2^2 = 4|p-p'| = 2|ra - r'a'|$. Also

$$|ra - r'a'| \le r|a-a'| + |a'|\,|r-r'| \le |a-a'| + |r-r'| \le \sqrt{2}\|(a,r) - (a',r')\|.$$

Taking square roots verifies the global $1/2$-Hölder condition stated in Section 4, with $H = 2^{3/4}$. It is not Lipschitz: the inputs $(1,1)$ and $(1, 1-h)$ have distance h and squared output distance $2h$. At u_+,

$$W_2^2(Q_h(u_+), \delta_{u_+}) = 2(1-e^{-h}) \sim 2h. \tag{52}$$

On the invariant support S, the spatial W_2-Lipschitz constant of Q_t is $e^{-t/2} \le 1$: its two input states have distance 2, while their two transition laws have distance $2e^{-t/2}$. Consequently this example retains uniform spatial regularity on the common support while attaining the critical temporal scale.

The mechanism is a jump of fixed size with probability proportional to h. Its mean squared size is proportional to h, hence its Wasserstein size is proportional to $\sqrt{h}$. In the variance argument of Appendix A, n such contributions over steps $h = t/n$ need not vanish: $nh = t$. This matches the failure of deterministic class evolution in the example. The zero-time convergence is uniform even on the entire ambient cube, since $W_2^2(Q_h(a,r,z), P(a,r,z)) = 2|ra|(1-e^{-h}) \le 2(1-e^{-h})$. Thus all the hypotheses of Proposition 2 other than the little-$o(\sqrt{h})$ modulus hold. Equivalently, restricting the ambient fine space itself to the two-point set S makes the same boundary example globally spatially Lipschitz. On the cube the kernels are only Hölder outside S, which the proposition permits. The example does not constitute a complete classification for every decoder Hölder exponent.

E APPROXIMATE COARSE RECOVERY AND DISTANCES BETWEEN JOINT LAWS

This section proves Theorem 5 and states Corollary 8 in full. We first compare the generated sequence with independent reconstruction. We then ask whether either can approximate an externally specified continuous target process. A coupling is a joint pairing with the two prescribed marginal laws.

E.1 A BOUND INDEPENDENT OF THE NUMBER OF TIME STEPS

There are two sequences to compare. The first uses its own sampled fine state to form the next coarse input. The second follows the deterministic coarse path from c_0 and draws a fresh reconstruction at each time. The theorem bounds the best possible pairing of these two entire sequences. It does not require that two independently simulated sequences happen to be close sample by sample. Let $X = \mathbb{R}^{d_f}$, $C = \mathbb{R}^{d_c}$, let R be Borel, and let $G : C \to \mathcal{P}_2(X)$ be W_2-Lipschitz with constant L_G. For this subsection, no compactness or coarse-speed assumption is needed. Assume exact positive-time composition, and the uniform root-mean-square fiber bound

$$\sup_{c\in C}\left(\int \|Ru-c\|^2 G(c,du)\right)^{1/2} \le \varepsilon. \tag{53}$$

Fix u_0, put $c_0 = Ru_0$, and take a grid $0 = t_0 < t_1 < \cdots < t_n = T$ with $\Delta_k = t_k - t_{k-1}$. Write $U_k = U_{t_k}$, $a_k = \phi_{t_k}(c_0)$, and $\ell_k = \mathrm{Lip}(\phi_{\Delta_k}) < \infty$. Let $\mathsf{Q}_{\mathrm{roll}}$ be the law of the Markov rollout

$$U_k \mid U_0, \dots, U_{k-1} \sim G\big(\phi_{\Delta_k}(RU_{k-1})\big), \qquad U_0 = u_0,$$

and let $\mathsf{Q}_{\mathrm{fresh}} = \bigotimes_{k=1}^{n} G(a_k)$. For any deterministic weights $w_k > 0$ with $\sum_k w_k = T$, chosen before sampling, use the product metric $d_w(u,v)^2 = \sum_k w_k \|u_k - v_k\|^2$ and its Wasserstein distance $W_{2,w}$.

The weights tell us how much each observation contributes to the sequence error. For example, $w_k = \Delta_k$ makes the cost a time integral for piecewise constant paths. Requiring their sum to be T keeps the total weight fixed when a time grid is refined. A joint coupling must pair whole sequences with their prescribed joint laws; it cannot choose unrelated optimal pairings at different times if those pairings destroy a sequence's temporal dependence.

Proposition 7. *Under these assumptions,*

$$W_{2,w}(\mathsf{Q}_{\mathrm{roll}}, \mathsf{Q}_{\mathrm{fresh}}) \le L_G\varepsilon \left(\sum_{k=2}^{n} w_k \ell_k^2\right)^{1/2}. \tag{54}$$

In particular the bound is at most $L_G\varepsilon \max_k \ell_k \sqrt{T}$, independently of the number of time steps.

Proof. Exact composition gives $\mathcal{L}(U_k) = Q_{t_k}(u_0) = G(a_k)$. Thus (53) implies $\mathbb{E}\|RU_k - a_k\|^2 \le \varepsilon^2$ at each time directly, without a recursive accumulation estimate. Put $C_k = \phi_{\Delta_k}(RU_{k-1})$. For $k \ge 2$, the coarse semigroup law gives

$$\mathbb{E}\|C_k - a_k\|^2 \le \ell_k^2 \mathbb{E}\|RU_{k-1} - a_{k-1}\|^2 \le \ell_k^2 \varepsilon^2.$$

At the first step, $C_1 = a_1$ deterministically. Recursively, condition on the entire coupling history and optimally couple the next U_k with a variable V_k whose law is the fixed measure $G(a_k)$. Measurable optimal transport couplings are available for these Euclidean probability kernels. Measurable couplings within an arbitrarily small error would give the same bound by a limit. The first conditional marginal is $G(C_k)$, as required, and the second conditional marginal is always $G(a_k)$, independent of the history. Consequently the variables V_k are mutually independent.

For example, condition on all previously generated U and V values. Regardless of that history, $\Pr(V_k \in A \mid \text{history}) = G(a_k)(A)$. Successively taking conditional expectations therefore gives

$$\Pr(V_1 \in A_1, \dots, V_n \in A_n) = \prod_{k=1}^{n} G(a_k)(A_k).$$

Although V_k can be correlated with U_k in the comparison coupling, its law conditional on earlier pairs is always the same. This is enough for the V sequence to have the independent product law

required by the theorem. At $k = 1$ the two conditional laws coincide, so choose $U_1 = V_1$ with zero cost. Decoder regularity gives

$$\mathbb{E}\|U_k - V_k\|^2 \le L_G^2 \mathbb{E}\|C_k - a_k\|^2.$$

Multiplication by w_k and summation proves the result.

Explicitly, the constructed coupling has

$$\mathbb{E}d_w(U,V)^2 = \sum_{k=2}^{n} w_k \mathbb{E}\|U_k - V_k\|^2 \le L_G^2 \varepsilon^2 \sum_{k=2}^{n} w_k \ell_k^2.$$

The optimal transport cost cannot exceed the cost of this particular coupling. Taking square roots gives Eq. (54). For a common bound ℓ on the coarse Lipschitz constants, the sum is at most $\ell^2 T$. Exact composition is used earlier to keep every marginal recovery error at most ε; the absence of an explicit factor n in the final bound is a consequence of that identity and the fixed total weight. □

Without exact composition, the same coupling only proves

$$W_{2,w}(\mathsf{Q}_{\mathrm{roll}}, \mathsf{Q}_{\mathrm{fresh}}) \le L_G \left(\sum_k w_k \mathbb{E}\|C_k - a_k\|^2 \right)^{1/2}. \tag{55}$$

Replacing the RMS errors of these actual coarse inputs by $\ell_k \varepsilon$ requires the marginal identity supplied by exact composition. A small local composition defect does not by itself justify that replacement.

E.2 Comparison with a continuous target process

Fix a coarse initial state c_0. All target distributions and expectations in this subsection refer to this fixed coarse condition. Let $(Z_t)_{0 \le t \le T}$ be a mean-square continuous $\mathbb{R}^{d_f}$-valued process with finite second moments. It represents the evolution to be approximated, supplied independently of the reconstruction architecture. For example, an application may obtain such a target by evolving random fine initial states with the same restriction c_0 using a fine-grid PDE solver. The assumption of mean-square continuity must hold for the chosen target. It is not asserted for every PDE model. The target fine initial state may be random under this coarse condition, whereas the generated sequence below starts from one fixed fine state u_0 with $Ru_0 = c_0$.

Define the integrated variance

$$\mathcal{V}_T = \int_0^T \mathbb{E}\|Z_t - \mathbb{E}Z_t\|^2 \, dt.$$

Take a family of grids $0 = t_0 < \cdots < t_n = T$ with mesh $h = \max_k \Delta_k \to 0$ and weights $w_k = \Delta_k$. The subscript h indexes this family. The grid, its number of points, and its time increments may all vary. Allow the decoder G_h and its recovery bound ε_h to vary with the grid. For each grid, assume the hypotheses of Proposition 7, including exact composition, and write

$$\mathsf{P}_h = \mathrm{Law}(Z_{t_1}, \ldots, Z_{t_n}), \qquad \mathsf{Q}_{\mathrm{fresh,h}} = \bigotimes_{k=1}^{n} G_h(a_k), \qquad a_k = \phi_{t_k}(c_0).$$

Here $\mathsf{Q}_{\mathrm{roll,h}}$ denotes the joint law of the sequence generated from $U_0 = u_0$ by

$$\mathrm{Law}(U_k \mid U_0, \ldots, U_{k-1}) = G_h\big(\phi_{\Delta_k}(RU_{k-1})\big).$$

Write $W_{2,h}$ for the Wasserstein distance whose squared sequence cost is $\sum_k \Delta_k \|u_k - v_k\|^2$, and put $\ell_h = \max_k \mathrm{Lip}(\phi_{\Delta_k})$. Define

$$\eta_h^2 = \sum_k \Delta_k W_2^2\big(G_h(a_k), \mathrm{Law}(Z_{t_k})\big). \tag{56}$$

This quantity first compares the distribution at each time separately and then adds those squared distances with time weights. Each comparison can use a different optimal coupling. In contrast, $W_{2,h}(\mathsf{P}_h, \mathsf{Q}_{\mathrm{roll,h}})$ pairs entire sequences and therefore depends on their temporal relationships. The condition $\eta_h \to 0$ is a time-weighted condition. It does not by itself state convergence at each individual observation time.

For orientation, the three relevant laws are P_h (the target sequence), $\mathsf{Q}_{\mathrm{roll,h}}$ (the architecture's generated sequence), and $\mathsf{Q}_{\mathrm{fresh,h}}$ (independent reconstruction on the deterministic coarse path). Theorem 5 compares the last two. To connect them to the target, the proof introduces F_h, independent draws from the target's own marginals. This fourth law shares every one-time distribution with P_h while discarding all its time dependence. Small η_h makes the two independent product laws close; it does not make the target process independent.

Corollary 8. *Under the assumptions above, if $\eta_h \to 0$, $L_{G_h}\varepsilon_h \to 0$, and $\sup_h \ell_h < \infty$, then*

$$W_{2,h}^2(\mathsf{P}_h, \mathsf{Q}_{\mathrm{roll,h}}) \longrightarrow 2\mathcal{V}_T. \tag{57}$$

Without requiring these three limits or bounds, the following estimate holds:

$$\liminf_{h\to 0} W_{2,h}(\mathsf{P}_h, \mathsf{Q}_{\mathrm{roll,h}}) \geq \left[\sqrt{2\mathcal{V}_T} - \limsup_{h\to 0}\big(\eta_h + L_{G_h}\varepsilon_h\ell_h\sqrt{T}\big)\right]_+. \tag{58}$$

Here $[a]_+ = \max(a, 0)$.

The limit is positive precisely when $\mathcal{V}_T > 0$. It therefore rules out accurate joint approximation in this regime for a target with nonzero integrated variance, even when its time-weighted marginal error vanishes. A simple example is $Z_t = \xi$ for all t, where ξ has nonzero variance: independent draws from $\mathrm{Law}(\xi)$ match every time marginal exactly, yet their joint laws remain separated from the constant-in-time target. This illustrates the difference between retaining a random state over time and redrawing a state independently at each time.

Proof. Let $\mathsf{F}_h = \bigotimes_k \mathcal{L}(Z_{t_k})$ be the independent product of the true one-time marginals. Independent coordinate couplings give $W_{2,h}(\mathsf{Q}_{\mathrm{fresh,h}}, \mathsf{F}_h) \leq \eta_h$. The triangle inequality and (54) bound the difference between the distance to $\mathsf{Q}_{\mathrm{roll,h}}$ and the distance to F_h by $\eta_h + L_{G_h}\varepsilon_h\ell_h\sqrt{T}$. It remains to prove

$$W_{2,h}^2(\mathsf{P}_h, \mathsf{F}_h) \longrightarrow 2\mathcal{V}_T. \tag{59}$$

Step 1: represent a sampled sequence as a function of time. Let $H = L^2([0,T]; \mathbb{R}^{d_f})$. Its elements are time-dependent vectors with finite integral of their squared norm; two functions that differ only on a set of times of length zero represent the same element. Its inner product and norm are

$$\langle f, g\rangle_H = \int_0^T \langle f(t), g(t)\rangle\, dt, \qquad \|f\|_H^2 = \int_0^T \|f(t)\|^2\, dt.$$

Embed a sequence by using its value at t_k throughout $(t_{k-1}, t_k]$. The squared H-distance is exactly $\sum_k \Delta_k\|u_k - v_k\|^2$. Thus this change of notation preserves the Wasserstein distance on every grid.

Mean-square continuity says that $t \mapsto Z_t$ is continuous in the norm $(\mathbb{E}\|Z_t\|^2)^{1/2}$ on the space of square-integrable random vectors. On the compact interval it is uniformly continuous and has bounded second moments. It admits a jointly measurable version, which we use to regard Z as an H-valued random variable. One can construct it as follows. Choose successively fine deterministic time grids so that the piecewise constant approximations $Z^{[j]}$ satisfy $\sup_t \mathbb{E}\|Z_t^{[j]} - Z_t\|^2 \leq 2^{-4j}$. Each approximation is jointly measurable because it uses finitely many random variables on measurable time intervals. For any fixed t, the probability that $\|Z_t^{[j]} - Z_t\| > 2^{-j}$ is at most 2^{-2j} by Markov's inequality. The sum of these probabilities is finite, so with probability one only finitely many of these events occur. The limit of the approximations, set to zero where it fails to exist, is jointly measurable and equals Z_t almost surely for each fixed t. Its expected squared H-norm is finite by integration of the bounded second moments. If $Z^{(h)}$ is the piecewise constant sampled process, then

$$\mathbb{E}\|Z^{(h)} - Z\|_H^2 \leq T \sup_{|s-t|\leq h} \mathbb{E}\|Z_s - Z_t\|^2 \longrightarrow 0.$$

Using the same target realization on both sides is a coupling, so their laws converge in $W_2(H)$. Values at the isolated time zero do not affect this norm.

Step 2: center the variables and record their total variance. Put $m(t) = \mathbb{E}Z_t$ and $X_*(t) = Z_t - m(t)$. Let Y_h be the centered, piecewise constant independent sequence: on the kth grid interval it equals an independent draw with law $\mathrm{Law}(Z_{t_k} - m(t_k))$. Then

$$\mathbb{E}\|X_*\|_H^2 = \mathcal{V}_T, \qquad S_h := \mathbb{E}\|Y_h\|_H^2 = \sum_k \Delta_k \operatorname{Var}(Z_{t_k}) \longrightarrow \mathcal{V}_T.$$

This is a Riemann sum for a continuous variance function. Write $M = \sup_{0\le t\le T} \operatorname{Var}(Z_t) < \infty$, so $S_h \le MT$. We will prove that $\mathbb{E}\langle X_*, Y_h\rangle_H$ tends to zero uniformly over *all* couplings of these two laws. The coupling may correlate the target with the independent sequence. The coordinates of that sequence must still be independent of one another, because this is part of its prescribed marginal law.

Step 3: use averages over finitely many time intervals. Partition $[0, T]$ into r equal intervals I_j, each of length $\delta = T/r$. Let $\Pi_r f$ replace f on each I_j by its average there. This is an orthogonal projection: $\Pi_r f$ is constant on every block, and $f - \Pi_r f$ has zero average on each block. Consequently its inner product with any blockwise constant function is zero.

If α_{jk} is the length of $I_j \cap (t_{k-1}, t_k]$, the average of Y_h on I_j is $\delta^{-1}\sum_k \alpha_{jk} Y_{h,k}$. The variables $Y_{h,k}$ are centered and independent, so cross terms have zero expectation. Since $0 \le \alpha_{jk} \le h$ and $\sum_k \alpha_{jk} = \delta$,

$$\mathbb{E}\left\|\frac{1}{\delta}\sum_k \alpha_{jk} Y_{h,k}\right\|^2 \le \frac{M}{\delta^2}\sum_k \alpha_{jk}^2 \le \frac{Mh}{\delta}.$$

Multiplying by the block length and summing over the r blocks gives $\mathbb{E}\|\Pi_r Y_h\|_H^2 \le rMh$. Thus an average over any fixed finite time partition loses its randomness as the independent sampling grid becomes finer.

The target has the opposite relevant property: it can be approximated by such block averages. Set $e_r = \mathbb{E}\|X_* - \Pi_r X_*\|_H^2$. Jensen's inequality within each interval yields

$$e_r \le T \sup_{|s-t|\le T/r} \mathbb{E}\|X_*(s) - X_*(t)\|^2 \longrightarrow 0.$$

This step is where mean-square continuity of the target enters the comparison.

Step 4: bound the cross term for every pairing. Orthogonality and Cauchy–Schwarz give, for every coupling,

$$\begin{aligned} |\mathbb{E}\langle X_*, Y_h\rangle_H| &\le |\mathbb{E}\langle \Pi_r X_*, \Pi_r Y_h\rangle_H| + |\mathbb{E}\langle X_* - \Pi_r X_*, Y_h\rangle_H| \\ &\le \sqrt{\mathcal{V}_T\, rMh} + \sqrt{e_r S_h}. \end{aligned} \tag{60}$$

Both terms depend only on the marginal laws, so the bound holds uniformly over all ways of coupling them. First fix r and let $h \to 0$; the first term vanishes. Then let $r \to \infty$; the second vanishes because $S_h \le MT$ and $e_r \to 0$.

Finally, expand the squared norm for any coupling:

$$\mathbb{E}\|X_* - Y_h\|_H^2 = \mathcal{V}_T + S_h - 2\mathbb{E}\langle X_*, Y_h\rangle_H.$$

The uniform cross-term bound shows that the infimum over couplings tends to $2\mathcal{V}_T$. The mean functions $m^{(h)}$ of the sampled target converge to m in H. For any two centered laws and deterministic means, adding the means changes the squared transport distance by the squared distance between those means, since the cross terms with centered variables have zero expectation. Thus the same limit holds for the uncentered target law and F_h. The coupling in Step 1 and the triangle inequality transfer it from the continuous target to its sampled law P_h, proving Eq. (59).

Step 5: return to the generated sequence. Writing $d_h = W_{2,h}(\mathsf{P}_h, \mathsf{F}_h)$ and $b_h = \eta_h + L_{G_h}\varepsilon_h \ell_h \sqrt{T}$, the initial comparison says

$$|W_{2,h}(\mathsf{P}_h, \mathsf{Q}_{\mathrm{roll,h}}) - d_h| \le b_h.$$

Under the stated limits, $b_h \to 0$ and $d_h \to \sqrt{2\mathcal{V}_T}$, proving Eq. (57). Without those limits, the one-sided inequality $W_{2,h}(\mathsf{P}_h, \mathsf{Q}_{\mathrm{roll,h}}) \ge d_h - b_h$, together with nonnegativity, gives Eq. (58). □

The limit (59) also follows from the classical Wasserstein Opial property (Naldi and Savaré, 2021, preprint, Theorem 5.1). We use this transport fact together with the architecture's nonaccumulating estimate (54). For random initial contexts the argument applies conditionally. Identifying that conditional comparison with an unconditional transport distance requires further proof.

F THE ROTATION CONSTRUCTION AND ITS GLOBAL SENSITIVITY BOUND

This section states and proves Proposition 9. The target is a single random direction transported by rotation, namely $Z_t = (0, O_t\Theta)$ with the maps and random variable defined below. The construction shows that an exactly composing model can approach this target as coarse recovery error vanishes, provided the decoder's Lipschitz constant is allowed to grow. It establishes this possibility for the specified target, without asserting an approximation construction for arbitrary target processes.

The probability-valued map used below is related to strong random projections: these assign a probability measure on a subset to each ambient point, agree with the point mass on that subset, and vary Lipschitz-continuously in W_1 (Ambrosio and Puglisi, 2020, Definition 2.7). Here W_1 uses expected distance as transport cost, whereas W_2 uses expected squared distance followed by a square root. This provides a precedent for replacing a deterministic projection by a probability-valued one. Our sphere construction requires a global Lipschitz estimate in W_2, which does not follow from a W_1 Lipschitz estimate alone. We prove that estimate directly, together with rotational equivariance and the inverse-error scale of decoder sensitivity.

Let $d \geq 4$, $X = \mathbb{R}^d \times \mathbb{R}^d$, $C = \mathbb{R}^d$, and $R(c, z) = c$. Fix $b > 0$ and let Θ be uniform on $\mathbb{S}^{d-1}$. For $\|x\| < 1$, define

$$H(x, \Theta) = \frac{x + (1 - \|x\|)\Theta}{\|x + (1 - \|x\|)\Theta\|}. \tag{61}$$

For each fixed x, the denominator vanishes only on a sphere-null set. Assign any unit vector there. Define

$$G_b(c) = \begin{cases} \mathcal{L}\big(bH(c/b, \Theta), H(c/b, \Theta)\big), & \|c\| < b, \\ \delta_{(c, c/\|c\|)}, & \|c\| \geq b. \end{cases} \tag{62}$$

The input c is a coarse vector, while the output has two vectors: a coarse coordinate and a unit-direction coordinate. Inside the radius-b ball, c/b lies in the unit ball. The numerator in H adds that vector to a random direction with weight $1 - \|c/b\|$ and then normalizes its length. At $c = 0$ the output is $(b\Theta, \Theta)$. As the input approaches a boundary point be, the random contribution vanishes and the output tends to (be, e). Outside the ball the decoder returns that deterministic graph point directly. This explains how randomness is confined to an initial reconstruction, while a small nonzero coarse coordinate can retain the selected direction afterward.

The only possible zero denominator occurs when $\|x\| = 1/2$ and $\Theta = -x/\|x\|$. A single direction has probability zero under the uniform sphere law, so its assigned output does not change $G_b(c)$. The later regularity proof will also check a common probability-zero exceptional set along each segment; a pointwise observation alone would not justify integrating derivatives along an entire segment. A matrix A is skew-symmetric when $A^\top = -A$. Its maps $O_t = \exp(tA)$ are rotations and preserve Euclidean norms. Set $\phi_t(c) = O_t c$ and $\Phi_t(c, z) = (O_t c, O_t z)$.

Proposition 9. *The kernels $Q_t = G_b \phi_t R$ compose exactly for all positive times and all fine starting states. The uniform root-mean-square fiber error is at most $2b$, and G_b is globally W_2-Lipschitz with optimal Lipschitz constant L_{G_b} satisfying*

$$\frac{\sqrt{2(1 + b^2)}}{b} \leq L_{G_b} \leq \sqrt{\frac{8(d-2)}{d-3}}\,\frac{\sqrt{1 + b^2}}{b}. \tag{63}$$

From initial coarse context $c_0 = 0$, every positive-time grid law equals the grid law of $(bO_t\Theta, O_t\Theta)$. If $\mathbb{P}$ is the law of the target $Z_t = (0, O_t\Theta)$, obtained by applying Φ_t to the random initial state $(0, \Theta)$, and $\mathbb{Q}_b$ is the law of the generated positive-time process, then in $L^2([0, T]; X)$,

$$W_2^2(\mathbb{P}, \mathbb{Q}_b) = Tb^2. \tag{64}$$

Proof. All outputs lie on the closed graph

$$D_b = \{(c, c/\|c\|) : \|c\| \geq b\}.$$

For interior inputs, the output coarse coordinate has norm b, so its distance from the input is at most $b + \|c\| \le 2b$ pointwise. For exterior inputs, that coordinate is exactly c. This proves the fiber bound.

Rotational equivariance means that rotating an input rotates its output distribution by the same amount. For any orthogonal map O, rotational invariance of Θ and the identity $H(Ox, O\Theta) = OH(x, \Theta)$ away from its null singular set give

$$G_b(Oc) = (O \oplus O)_{\#} G_b(c). \tag{65}$$

The graph D_b is invariant under Φ_t, and, on that graph, $Q_t(c, z) = \delta_{\Phi_t(c,z)}$. Starting from an arbitrary u, a transition of length s enters D_b. Its next transition of length t pushes its law forward by Φ_t. Equation (65) therefore yields

$$Q_s Q_t(u) = (\Phi_t)_{\#} G_b(O_s R u) = G_b(O_t O_s R u) = Q_{s+t}(u).$$

The pushforward $(\Phi_t)_{\#}\nu$ above is the law of $\Phi_t(U)$ for $U \sim \nu$. Once reconstruction has placed a state on D_b, its coarse coordinate already determines the direction, so further decoder calls add no new randomness. This is why exact composition holds even for input states initially outside the graph.

Why the normalization needs a moment estimate. To prove the global regularity bound, we first control the singular normalization in the interior. Couple two interior inputs using the same Θ. For $x \neq 0$, put $r = \|x\|$, $v = x + (1 - r)\Theta$, and $q = \Theta \cdot x/r$. Then

$$\|v\|^2 = 1 - 2r(1 - r)(1 - q) \ge \frac{1 + q}{2}. \tag{66}$$

Where differentiable, $\|D_x v\| \le 2$, and the derivative of normalization has norm $1/\|v\|$. Hence $\|D_x H\, a\| \le 2\|a\|/\|v\|$.

Here D_x denotes a derivative with respect to the input vector; its operator norm is the maximum enlargement of a unit perturbation. The derivative of $v \mapsto v/\|v\|$, applied to a vector a, is

$$\frac{1}{\|v\|}\left(a - \frac{v\langle v, a\rangle}{\|v\|^2}\right).$$

The expression in parentheses is an orthogonal projection, of norm at most $\|a\|$. Also $D_x v$ is the identity minus the map $a \mapsto \Theta\langle x/\|x\|, a\rangle$, so its norm is at most two. The denominator can be small for some random directions. We therefore control the expected square of this derivative, which is the quantity needed for a W_2 estimate. The one-dimensional marginal of a uniform sphere point has density proportional to $(1 - q^2)^{(d-3)/2}$ on $(-1, 1)$. A slice at coordinate q has radius $\sqrt{1 - q^2}$, contributing a factor $(1-q^2)^{(d-2)/2}$ from its $(d-2)$-dimensional spherical area. The sphere's sloping surface contributes the factor $(1 - q^2)^{-1/2}$ when converting slice thickness to dq. Their product gives the stated exponent. Only the density up to normalization is needed, since the normalizing constant cancels in the ratio of integrals. The beta-integral ratio gives

$$\mathbb{E}\frac{1}{1 + q} = \frac{d - 2}{d - 3}, \qquad d \ge 4. \tag{67}$$

Write $B(a, b) = \int_0^1 w^{a-1}(1 - w)^{b-1}\, dw$ for the beta integral. Substituting $w = (1 + q)/2$ and using $B(a - 1, a)/B(a, a) = (2a - 1)/(a - 1)$ with $a = (d - 1)/2$, together with the factor $1/2$ from $(1 + q)^{-1}$ proves the identity.

For completeness, the beta ratio can be obtained without special-function formulas. Integrating the derivative of $w^{a-1}(1 - w)^a$ over $[0, 1]$ gives $(a - 1)B(a - 1, a + 1) = aB(a, a)$; the endpoint terms vanish because $a > 1$. Also $B(a - 1, a) = B(a - 1, a + 1) + B(a, a)$. Combining them gives the stated ratio. Near $q = -1$, the integrand for $\mathbb{E}(1 + q)^{-1}$ behaves as $(1 + q)^{(d-5)/2}$, which is integrable exactly when $d > 3$. This is the reason for the dimension assumption in this construction's estimate. Combining the preceding bounds gives

$$\mathbb{E}\|D_x H\, a\|^2 \le \frac{8(d - 2)}{d - 3}\|a\|^2. \tag{68}$$

For a fixed nondegenerate straight segment in the open unit ball, a vanishing denominator requires $\|x\| = 1/2$ and $\Theta = -x/\|x\|$. The segment meets that sphere in at most two points, so the exceptional directions have sphere measure zero. Outside that null set, the map along the segment is absolutely continuous, and the possible corner of $\|x\|$ at zero is harmless. Integrating its derivative and applying Minkowski's inequality in $L^2(\Theta)$ proves

$$\big(\mathbb{E}\|H(x,\Theta) - H(x',\Theta)\|^2\big)^{1/2} \le \sqrt{\frac{8(d-2)}{d-3}}\,\|x - x'\|.$$

To spell out the integration step, put $x_\theta = x + \theta(x' - x)$ for $0 \le \theta \le 1$. For almost every sphere direction,

$$H(x',\Theta) - H(x,\Theta) = \int_0^1 D_x H(x_\theta,\Theta)(x' - x)\,d\theta.$$

Taking the square-mean norm in Θ and applying the integral triangle inequality bounds it by the integral of those square-mean derivative norms. Equation (68) bounds every integrand by the same constant times $\|x' - x\|$. A possible passage through $x_\theta = 0$ contributes only a single parameter value and does not affect the integral.

Rescaling by b and including both output coordinates yields the upper bound in (63) for interior pairs. The bound extends to the boundary: for x approaching a unit vector e, $H(x,\Theta) \to e$ almost surely, and bounded convergence gives convergence in L^2. The resulting boundary law agrees with the exterior definition. On the exterior, the deterministic map $c \mapsto (c, c/\|c\|)$ has derivative norm at most $\sqrt{1 + b^{-2}}$. Split any straight segment at its intersections with the ball of radius b. Integration on each exterior portion, the interior bound, and the triangle inequality in W_2 prove the same bound globally.

The scale factors can be read directly: replacing c by $x = c/b$ multiplies input differences by $1/b$, and the two output coordinates (bH, H) multiply a direction difference by $\sqrt{1 + b^2}$. At the boundary the two decoder definitions agree in W_2. Splitting a straight segment is legitimate even if its endpoints are outside the ball: its interior and exterior subsegments have total length equal to the original segment. The same upper Lipschitz constant on every piece then gives a bound proportional to the distance between the original endpoints.

A matching inverse-scale lower bound. For the lower bound, compare inputs 0 and be_1. The second output is Dirac, so its unique coupling with the first gives

$$W_2^2(G_b(0), G_b(be_1)) = (1 + b^2)\mathbb{E}\|\Theta - e_1\|^2 = 2(1 + b^2).$$

Their input distance is b, proving the claimed lower bound.

The joint law and its distance to the target. Finally, at the first positive grid time t_1, the decoder draws $(b\Theta_1, \Theta_1)$, because $O_{t_1} c_0 = 0$. All subsequent states rotate deterministically, so their joint law is

$$\big(bO_{t_k - t_1}\Theta_1, O_{t_k - t_1}\Theta_1\big)_{k=1}^n.$$

The change of variable $\Theta = O_{-t_1}\Theta_1$, still uniform, yields the claimed grid law. These laws are realized by the continuous path $(bO_t\Theta, O_t\Theta)$ on positive times. The true flow from the initial ensemble $(0,\Theta)$ has path $(0, O_t\Theta)$. Coupling them with the same Θ has squared path cost Tb^2. Every coupling incurs at least that cost from the coarse coordinate alone: the target coarse coordinate is zero and the generated coarse coordinate always has norm b.

For each time, the coarse-coordinate contribution to squared distance is therefore exactly b^2 for every possible pairing. Integrating over $[0, T]$ gives Tb^2, and all residual-coordinate costs are nonnegative. The pairing with the same direction realizes this lower bound by making the residual coordinates identical at every time. An upper bound from a chosen coupling and a lower bound valid for every coupling thus agree. This proves equality in (64). □

Taking $b = \varepsilon/2$ gives fiber error at most ε and squared path distance $T\varepsilon^2/4$, while L_{G_b} grows as ε^{-1} as $\varepsilon \downarrow 0$. The fixed-context marginal error is exactly b in W_2, so its time-weighted marginal error is $\eta_h = b\sqrt{T}$ on every grid. Choosing $b = b_h \downarrow 0$ makes this error vanish. With the recovery bound $\varepsilon_b = 2b$, however, $L_{G_b}\varepsilon_b$ does not tend to zero. The construction therefore lies outside the

small-product condition of Corollary 8, allowing accurate joint approximation of this rotational target. It does not give a universal sharp numerical constant for all architectures.

The comparison conditions on the coarse context and averages over the unobserved initial residual. It does not recover a particular observed fine initial state. The law $Q_h(u)$ tends to $P(u)$, which can differ from δ_u outside D_b. The continuous positive-time realization and the L^2 path metric do not impose continuity from every ambient initial state at the isolated time zero. On D_b, each projection class is Dirac and the class flow rotates deterministically. The persistent random direction corresponds to the initial class selection in the classification theorem.

G Controlled temporal experiments and a learned diagnostic

The controlled experiments compare measured statistics with analytic laws. Advection illustrates Theorem 1 and Corollary 3, the Gaussian example examines the exact-composition assumption in Theorem 5, and the rotational experiment implements Proposition 9. They illustrate consequences and boundaries of the results without replacing their proofs. The learned experiment checks whether temporal and marginal diagnostics convey different information in a small trained model. All parameter values and seeds are reported, including improvements and deteriorations in the learned-model metrics.

G.1 Advection with an unresolved random phase

Purpose and relation to the theory. Theorem 1 describes the independence that remains after fixing the initial class, and Corollary 3 predicts the resulting short-time squared increments. This experiment makes both statements explicit in a field with just four Fourier coefficients. Two coefficients describe a large-scale wave, and two describe a finer wave whose phase is unresolved by the coarse observation. We compare repeatedly redrawing that phase with transporting one initial phase over time. This isolates temporal dependence while keeping the entire distribution at every individual time the same.

Fields, restriction, and the two sampling rules. On $[0, 2\pi]$ with periodic boundary conditions, write a fine field as

$$u(c, z)(\xi) = c_1 \cos \xi + c_2 \sin \xi + z_1 \cos(8\xi) + z_2 \sin(8\xi).$$

The restriction retains the two coarse coefficients, $Ru(c, z) = c$. The coarse flow rotates c through angle t, starting at $c_0 = (1, 0)$. The decoder appends $z = a(\cos \Theta, \sin \Theta)$ with $a = 0.5$ and Θ uniform on $[0, 2\pi]$. It is independent of c in the residual coordinates and exactly fiber-consistent. The coarse flow and decoder therefore give an exactly composable reconstruction kernel. All coefficients remain in compact invariant sets, and the classification assumptions hold there. We use the normalized spatial norm $\|u\|_{\mathrm{sp}}^2 = (2\pi)^{-1} \int_0^{2\pi} |u(\xi)|^2 \, d\xi$.

The spatial variable is ξ, while t denotes time. The first two basis functions complete one oscillation around the periodic domain; the other two complete eight. For any fine states with the same c, zero-time reconstruction gives the same law

$$\mu_c := \mathrm{Law}\big(u(c, a(\cos \Theta, \sin \Theta))\big).$$

Thus $P(u(c, z)) = \mu_c$ regardless of the input z. On the reconstruction support, $\mathcal{M} = \{\mu_c\}$ over the allowed coarse circle, and $F_t(\mu_c) = \mu_{\phi_t(c)}$. Fixing c_0 fixes the label μ_{c_0}, so $\rho_x = \delta_{\mu_{c_0}}$. The random fine phase Θ is a sample within that class; it is not a choice between class labels.

At each reconstruction time draw a new phase Θ_k. Restricting the resulting field removes that phase, leaving $\phi_{t_k}(c_0)$. At the next step the decoder therefore has no information about Θ_k. Theorem 1 predicts independent positive-time samples in this fixed-class case. The model is deliberately built to exhibit this structure, so the experiment checks its quantitative consequences rather than establishing the general theorem by numerical evidence.

The transport reference solves $u_t + u_\xi = 0$ with a single random initial phase:

$$z^{\mathrm{tr}}(t, \xi) = a \cos(8\xi - \Theta - 8t).$$

The independent reconstruction uses independent uniform phases at distinct positive times. Rotation preserves the uniform phase distribution, so the two models have identical full-field marginal laws

at each positive time, not merely matching moments. They also share the same coarse path. Both sample their own initial phases. Neither receives a reference test phase. This compares ensembles conditional on c_0, rather than recovery of a known fine initial state.

Derivation of the diagnostic formulas. Write z also for the residual field represented by the two fine coefficients. Orthogonality of the sine and cosine functions gives

$$\|z_1 \cos(8\xi) + z_2 \sin(8\xi)\|_{\mathrm{sp}}^2 = \tfrac{1}{2}(z_1^2 + z_2^2).$$

For a uniform phase, the residual mean field is zero and its expected squared spatial norm is $a^2/2$. Consequently the conditional full-field variance, in this norm, is $v(\mu_{\phi_t(c_0)}) = a^2/2$. Corollary 3 predicts a full-field increment limit of $2v = a^2 = 0.25$. The coarse component is deterministic and continuous, so its contribution tends to zero. Measuring only the residual removes that deterministic contribution and makes the independent-reconstruction value equal to a^2 at every positive lag.

For $t_k = kT/n$, $T = \pi/16$, $h = T/n$, define

$$I_n = \frac{1}{n-1}\sum_{k=1}^{n-1} \|z(t_{k+1},\cdot) - z(t_k,\cdot)\|_{\mathrm{sp}}^2, \qquad J_n = \frac{1}{n}\sum_{k=1}^{n} z(t_k, 0).$$

Uniform-phase integration gives

$$\mathbb{E}I_n^{\mathrm{ind}} = a^2, \qquad I_n^{\mathrm{tr}} = 2a^2 \sin^2(4h), \qquad \mathrm{Var}(J_n^{\mathrm{ind}}) = \frac{a^2}{2n}, \qquad \mathrm{Var}(J_n^{\mathrm{tr}}) = \frac{a^2}{2}\left|\frac{1}{n}\sum_{k=1}^{n} e^{8it_k}\right|^2.$$

Here I_n measures how much the residual changes between successive observations, while J_n is its average over time at the fixed spatial location zero. The phase identity

$$\mathbb{E}\{\cos(\Theta + \alpha)\cos(\Theta + \beta)\} = \tfrac{1}{2}\cos(\alpha - \beta)$$

follows by the product-to-sum formula and the vanishing average of a nonconstant sine or cosine over a full period. Independent phases instead have zero product expectation, because each cosine has mean zero. Hence each independent increment has expected squared norm $a^2/2 + a^2/2 = a^2$. For transport, the spatial inner product at separation h is $(a^2/2)\cos(8h)$, giving $a^2(1 - \cos(8h)) = 2a^2 \sin^2(4h)$.

For J_n, expand the variance of the sum into all pairs of times. In the independent case only the n diagonal terms remain, giving $a^2/(2n)$. In the transport case the sum is

$$\frac{a^2}{2n^2}\sum_{k,l}\cos(8(t_k - t_l)) = \frac{a^2}{2}\left|\frac{1}{n}\sum_{k} e^{8it_k}\right|^2.$$

The complex notation packages the sums of sines and cosines: the squared modulus is the sum of their squared averages. As the grid is refined, the average converges to $T^{-1}\int_0^T e^{8it}\,dt$, whose modulus is $|\sin(4T)/(4T)|$.

The transport variance tends to $\frac{a^2}{2}[\sin(4T)/(4T)]^2 = 1/\pi^2$. For positive lag τ, the normalized spatial residual correlation is zero for independent reconstruction and $\cos(8\tau)$ for transport. At zero lag both equal one. The endpoint residual at a fixed location has mean zero and variance $a^2/2$. In both models its density is $1/(\pi\sqrt{a^2 - z^2})$ for $|z| < a$, the scaled arcsine distribution.

The endpoint density follows from the change of variable $z = a\cos\Theta$: each interior value has two preimages in $[0, 2\pi)$, and $|dz/d\Theta| = \sqrt{a^2 - z^2}$. Adding their uniform-phase density contributions gives the displayed arcsine density. These identical endpoint distributions coexist with very different time averages: the independent variance tends to zero, whereas the transported variance remains positive. Thus matching endpoint histograms alone cannot distinguish these two temporal laws.

Sampling, checks, and reading the figures. We use $n = 16, 32, 64, 128$, seeds $0, 1, 2$, and 2,000 trajectories per seed and method. A statistic is first computed per trajectory. For a mean, its standard error uses the sample standard deviation divided by the square root of the trajectory count. Variances use the unbiased sample variance and a plug-in fourth-moment standard error. Across seeds, plotted

means average the three estimates and their standard errors combine in quadrature divided by three. Error bars show twice this standard error. All analytic checks use a prespecified investigation threshold of five standard errors, with a 10^{-12} tolerance for numerically deterministic statistics. No seed is removed.

A seed specifies one reproducible random-number stream. The independent statistical observations for these error bars are the trajectories, not the many times sampled within a trajectory. Reusing adjacent states in successive increments does not create additional independent trajectories. The five-standard-error threshold is a numerical investigation rule for discrepancies from known analytic answers; it is not a proof of a theorem or a simultaneous confidence guarantee for all reported checks.

Sampling and statistics use JAX float64. Independent checks reconstruct fields on a 256-point spatial grid, recover the coarse coefficients by a fast Fourier transform (FFT), and execute the full sequence $R \to \phi_h \to G$ from the previous fine field. Its outputs agree with direct coarse propagation and the same fresh phases to double-precision accuracy. The spatial norm agrees with the coefficient calculation, including the factor $1/2$ from the squared sine and cosine basis functions. Figure 3 shows endpoint samples and one prespecified seed-zero trajectory from each model.

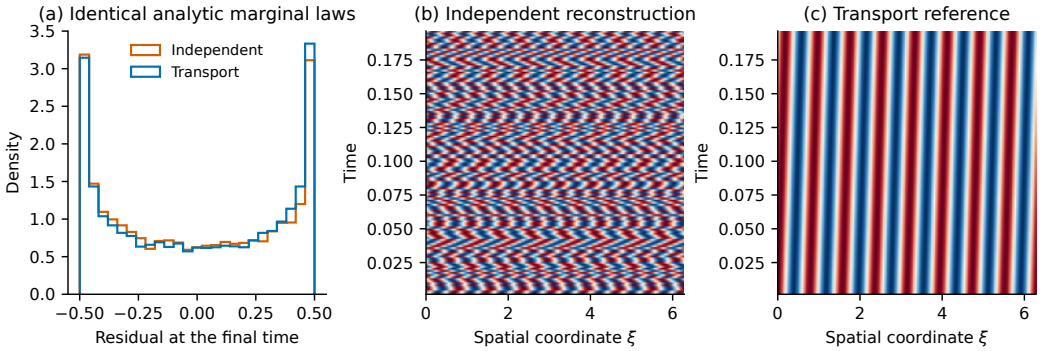

Figure 3: Advection diagnostic at $n = 128$. Left: endpoint residual histograms pool all $6{,}000$ trajectories per method. Identical marginal laws follow analytically from uniform phases. Right: the first trajectory from seed zero, without selection, shows independent reconstruction and coherent transport of the unresolved mode. Both color scales range from -0.5 to 0.5.

G.2 A BOUNDARY CHECK WITH APPROXIMATE COMPOSITION

Purpose. Theorem 5 controls the joint error using a uniform recovery bound and exact composition. Here we keep a small recovery error and a fixed decoder Lipschitz constant, while deliberately violating exact composition. Repeated reconstruction then changes the distribution of the coarse input, illustrating why the proof cannot simply reuse the original one-step recovery bound at later times. Figure 4 summarizes the comparison.

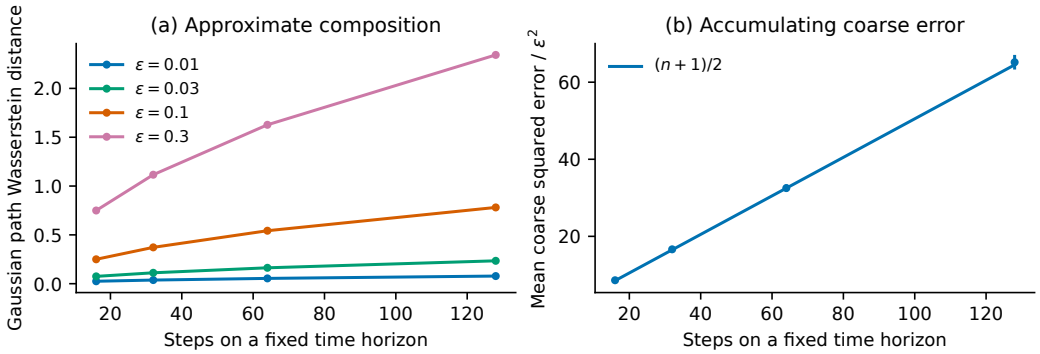

Figure 4: Approximate-composition boundary check on a fixed horizon. Left: the exact Gaussian path distance increases with the number of reconstruction steps. Right: measured coarse squared error, normalized by ε^2, follows $(n+1)/2$. Bars use three independent seeds, not the coupled repetitions across ε.

Let $X = \mathbb{R}^2$, $R(c, z) = c$, $\phi_t(c) = c$, and

$$G_\varepsilon(c) = \mathrm{Law}(c + \varepsilon\xi, \eta), \qquad \xi, \eta \text{ independent standard normal variables.}$$

The decoder has $L_G = 1$ and fiber RMS error ε. For every $s, t > 0$, the coarse variances of $Q_{s+t}(c, z)$ and $(Q_s Q_t)(c, z)$ are respectively ε^2 and $2\varepsilon^2$. The residual law is the same and independent in both. Consequently their W_2 distance is $(\sqrt{2} - 1)\varepsilon$. These kernels violate exact composition.

All positive durations use the same decoder because the coarse flow is the identity. One application adds one Gaussian perturbation to c; two applications add two independent perturbations. For centered scalar normal laws with standard deviations σ, τ, using the same standard normal variable gives squared cost $(\sigma - \tau)^2$. Every coupling has at least this cost by Cauchy–Schwarz applied to its covariance. Taking $\sigma = \varepsilon$ and $\tau = \sqrt{2}\varepsilon$ gives the stated composition defect. Likewise, translating the same noise by $c - c'$ gives $W_2(G_\varepsilon(c), G_\varepsilon(c')) = |c - c'|$: the upper bound comes from that coupling and the lower bound from the difference of means. This verifies $L_G = 1$.

Starting at coarse value zero, repeated reconstruction produces $c_k = \varepsilon \sum_{j=1}^{k} \xi_j$. On a fixed horizon $T = 1$ with $w_k = 1/n$,

$$\mathbb{E}\frac{1}{n}\sum_{k=1}^{n} c_k^2 = \varepsilon^2 \frac{n+1}{2}, \qquad \mathrm{Var}(c_n) = n\varepsilon^2.$$

The variables ξ_j have mean zero, variance one, and zero covariance across different steps. Therefore $\mathbb{E}c_k^2 = k\varepsilon^2$. Summing $k = 1, \ldots, n$ and dividing by n gives the weighted error above. Refining the grid adds more decoder calls even though the total physical time remains one; the noise size per call is unchanged. This is why increasing n changes the coarse distribution here.

The coarse path covariance is $\varepsilon^2 A_n$, where $(A_n)_{ij} = \min(i, j)$, while fresh reconstruction has covariance $\varepsilon^2 I_n$. Since the latter is scalar, diagonalizing A_n reduces the optimal Gaussian coupling to independent scalar couplings. If λ_j are its eigenvalues, the exact weighted path distance is

$$W_{2,w} = \varepsilon \left[\frac{1}{n} \sum_{j=1}^{n} (\sqrt{\lambda_j} - 1)^2 \right]^{1/2}.$$

For $i \leq j$, the sums defining c_i and c_j share exactly their first i independent perturbations, so $\mathbb{E}c_i c_j = i\varepsilon^2$. This gives the entries of A_n. A symmetric covariance matrix has an orthonormal eigenbasis. Rotating both sequences into that basis preserves squared Euclidean distance, while the fresh covariance I_n stays unchanged. The comparison then consists of scalar normal pairs with standard deviations $\varepsilon\sqrt{\lambda_j}$ and ε. Their optimal costs add, and the grid weight $1/n$ gives the displayed formula. Thus this curve is an exact calculation from Gaussian laws, with no estimated high-dimensional transport map.

The independent residual coordinates can be coupled identically and add no cost. This distance is bounded by the sequential coupling estimate $\varepsilon\sqrt{(n-1)/2}$ using the actual coarse input errors $\mathbb{E}c_{k-1}^2 = (k-1)\varepsilon^2$. Substituting a uniform error ε would instead give $\varepsilon\sqrt{(n-1)/n}$, which is inapplicable here. The discrepancy illustrates the role of exact composition in controlling marginal coarse errors. It is not a counterexample to Theorem 5 or a classification for approximate composition.

We use the same trajectory counts, seeds, and n values as advection, and $\varepsilon \in \{0.01, 0.03, 0.1, 0.3\}$. For each seed and n, the same standard-normal random walk is rescaled across ε. Monte Carlo checks compare the path coarse error and endpoint variance with the formulas above. The Gaussian path distance is computed analytically from the covariance eigenvalues, not fitted to samples. Figure 4 distinguishes these analytic and measured quantities.

G.3 Vanishing offsets with increasing sensitivity

Purpose and the two references. The question is whether vanishing coarse recovery error must erase temporal dependence. Proposition 9 gives one construction in which sensitivity grows as the error scale shrinks. We check its predicted dependence and approximation error. One reference redraws directions independently and tests temporal statistics at identical marginals. A second reference transports a direction with a zero coarse coordinate and tests approximation to the specified target. These are different comparisons, so they are reported separately. Figure 5 summarizes the comparison.

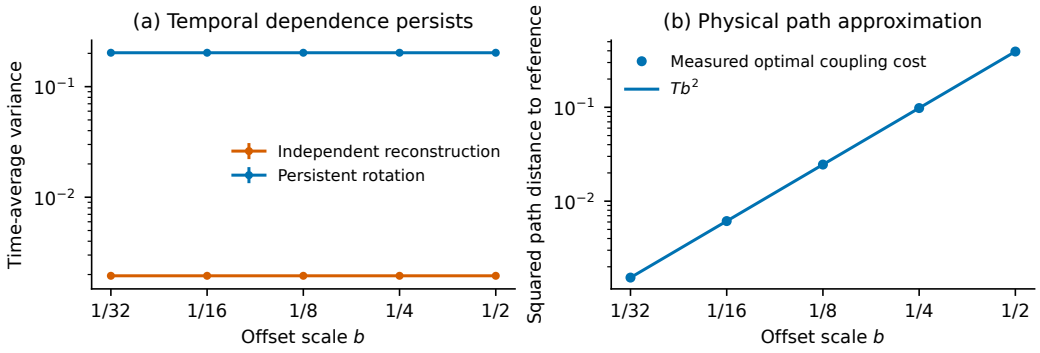

Figure 5: Sensitivity construction. Left: residual time-average variance remains separated from independent reconstruction as b decreases. Bars show twice the trajectory-based standard error across three seeds. Right: the measured cost of the provably optimal coupling to the physical ensemble equals Tb^2. The two references serve different comparisons, as specified in the text. Global sensitivity is certified by Proposition 9.

We implement Proposition 9 with $d = 4$, rotations of angular velocity one in both coordinate planes, $T = \pi/2$, and $n = 128$. For each of the three seeds we generate 2,000 independent uniform directions on the sphere. The same directions are used at every scale $b \in \{1/2, 1/4, 1/8, 1/16, 1/32\}$. Thus the curves across b are coupled evaluations, not independent estimates.

From coarse context zero, the joint law of the Markov rollout is that of $(bO_{t_k}\Theta, O_{t_k}\Theta)$ with a shared direction Θ. The independent reference draws a new uniform direction for each k, which has the same full-state marginals as that rollout. For the first residual coordinate, let $J_n = n^{-1}\sum_{k=1}^{n} z_1(t_k)$.

Its time-average variances are

$$\mathrm{Var}(J_n^{\mathrm{ind}}) = \frac{1}{dn}, \qquad \mathrm{Var}(J_n^{\mathrm{rot}}) = \frac{1}{d}\left|\frac{1}{n}\sum_{k=1}^{n} e^{it_k}\right|^2 .$$

Neither depends on b. The coarse RMS offset along the rollout is b, and the uniform decoder fiber-error bound is $2b$. The analytic bounds in Eq. (63) certify the global W_2 Lipschitz scale. The numerical experiment does not estimate that global constant from samples.

The variance formulas follow from symmetry of the uniform sphere law: $\mathbb{E}\Theta_i = 0$, $\mathbb{E}\Theta_i^2 = 1/d$, and $\mathbb{E}\Theta_i\Theta_j = 0$ for $i \neq j$. For one rotating plane the first coordinate is a linear combination of $\Theta_1 \cos t$ and $\Theta_2 \sin t$. Expanding the variance of its time average gives the squared sine and cosine sums in the displayed modulus. Independent reconstruction has zero cross-time covariances, leaving n diagonal terms. In the chosen parameters, the independent variance is $1/(4 \cdot 128)$, while the rotating variance stays near its positive continuous-time limit. Their difference persists at every value of b because the residual coordinates have not been scaled down.

A different reference, $(0, O_t\Theta)$, describes the physical ensemble with zero coarse coordinate. Any coupling to it pays at least Tb^2 in the coarse coordinates. Sharing the direction achieves that cost and eliminates the residual cost, so the squared path distance is exactly Tb^2. Figure 5 compares the measured coupling cost with this value. The positive-time initial choice of a direction is allowed by the classification theorem. We also check numerically that decoder outputs lie on D_b from Appendix F, that applying the zero-time reconstruction twice preserves its law, and that rotating inputs rotates the output law. These checks avoid the probability-zero set where normalization is undefined.

G.4 Small learned Burgers reconstruction

Purpose. The controlled examples permit exact analytic comparisons. This experiment asks how the evaluation lesson appears after fitting small neural decoders to PDE data: does adding the previous generated residual improve every aspect of the prediction? We compare single-time predictive scores, short-time changes, correlations, and uncertainty in a time average. These measure different features, so a good result for one does not determine the others.

Data and conditioning. We solve the separate one-dimensional problem $u_t + uu_\xi = 0.05u_{\xi\xi}$ on $[0, 2\pi]$ with 64 equispaced periodic points. The Fourier pseudo-spectral solver uses classical fourth-order Runge–Kutta (RK4) with step 0.002 and removes modes $|k| > 21$ after every update and from the nonlinear term. We store the initial state and 32 positive times spaced by 0.01, giving $T = 0.32$. The initial mean is uniform on $[0.8, 1.2]$. Independent sine and cosine coefficients are Gaussian with standard deviations $0.35/k^2$ for $k = 1, 2, 3$ and $0.10/\sqrt{5}$ for $k = 6, \ldots, 10$. Other modes vanish. The three splits contain 512, 128, and 128 independently generated trajectories, using data seed 271828 and separate derived random keys. Initial-state hashes show no overlap. A step-halving check on 16 test trajectories gives relative discrepancy below 1.4×10^{-7}. We check preservation of the spatial mean and reconstruction from the coarse and residual coordinates, both to tolerance 10^{-5}.

The restriction averages each block of four points, giving 16 coarse values. A fixed orthonormal basis for the zero-mean subspace of each block represents its residual with three coefficients, or 48 overall. Decoding adds the block mean to this residual. Thus coarse consistency holds algebraically, up to float32 arithmetic. Scalar coarse mean and standard deviation and residual RMS normalization use only the training split.

For clarity, the three basis columns used within a four-point block are

$$B = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} \\ -1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} \\ 0 & -2/\sqrt{6} & 1/\sqrt{12} \\ 0 & 0 & -3/\sqrt{12} \end{pmatrix} .$$

Each column sums to zero and $B^\top B = I_3$. If a block has mean c and values $u \in \mathbb{R}^4$, its coefficient vector is $q = B^\top(u - c\mathbf{1})$, and decoding gives $c\mathbf{1} + Bq$. Taking its mean returns exactly c in exact

arithmetic. Sixteen such blocks give $16 \cdot 3 = 48$ residual coefficients. This constraint is built into the representation rather than learned from a penalty.

Both models receive the same externally supplied coarse sequence $c_k = Ru(t_k)$ from each test trajectory. This removes coarse-solver error from the reconstruction comparison. It does not produce a learned closed-loop semigroup: these coarse paths need not be generated by an autonomous deterministic coarse flow, and the learned kernels are not tested for all-time composition.

Models, training, and sampling. The comparison is motivated by the use of generated history in temporal downscaling (Schillinger et al., 2026).

Both Gaussian decoders have two width-64 tanh hidden layers and 14,560 parameters, with a 64-dimensional input and 96 outputs for 48 means and log standard deviations. The independent model concatenates the coarse vector with 48 zeros. The memory model concatenates it with the preceding residual coefficients and adds those coefficients to its predicted mean. Log standard deviations are clipped to $[-4, 2]$. The inactive independent-model input weights are retained to keep the implementations comparable, though equal parameter counts do not equalize effective input information.

Each network specifies a diagonal Gaussian law for the normalized residual coefficients. If its outputs are means m_j and log standard deviations s_j, sampling uses $q_j = m_j + e^{s_j}\xi_j$ with independent standard normal ξ_j. The training loss per coefficient is

$$\frac{1}{2}(q_j^* - m_j)^2 e^{-2s_j} + s_j + \frac{1}{2}\log(2\pi),$$

averaged over the 48 coefficients and the sampled training batch. This is the negative log density of the specified Gaussian at the true normalized coefficient q_j^*. The memory model adds the previous coefficient to its network mean, so the network predicts a correction to that previous state. At evaluation this previous state is generated by the model itself.

For each seed $0, 1, 2$, models train for 4,000 Adam updates at learning rate 0.001, batch size 128, with gradient-norm clipping at one. Training samples a trajectory and a positive observation time uniformly. The memory model uses true previous residuals during training and starts with the second positive time. Validation negative log likelihood (NLL) per normalized residual coefficient is recorded every 500 updates. We report the fixed final checkpoint without validation-based selection. There is no hyperparameter search. The prespecified numerical checks require finite losses and samples, validation NLL improvement at least 0.05 from initialization, nonzero generated spread, and fiber RMS below 10^{-5}. All six final models pass. These are optimization and implementation checks, not tests of distributional calibration.

Each test context receives 16 generated paths. Both methods share the same initial positive-time sample from the independent model. Subsequent memory inputs use the previously generated residual, never the reference residual. Noise keys are paired between methods. This is free-running residual generation along a supplied coarse path.

In operational terms, both models first draw the same residual at t_1. At t_2, the independent model receives c_2 and zeros, while the memory model receives c_2 and its draw from t_1. This repeats at later times with fresh Gaussian noise. The reference fine trajectory is never passed into these later sampling steps. Its role is to supply the common coarse path and to evaluate the generated results afterward. The training protocol therefore also exposes the memory model to a difference between true training histories and generated test histories.

Metrics and statistical units. Let $r_{i,m,k,j}$ denote the generated residual at test context i, ensemble member m, time k, and spatial point j, and let $r^*_{i,k,j}$ denote its single reference trajectory. The fixed-location probe is $p_{i,m,k} = r_{i,m,k,0}$. We use the continuous ranked probability score (CRPS) and field energy score (Gneiting and Raftery, 2007), with smaller values preferred. For independent generated members, the following ensemble form is unbiased for the probe CRPS

$$\widehat{\mathrm{CRPS}}_{i,k} = \frac{1}{M}\sum_m |p_{i,m,k} - p^*_{i,k}| - \frac{1}{2M(M-1)}\sum_{m \neq m'} |p_{i,m,k} - p_{i,m',k}|, \qquad M = 16.$$

The field energy score replaces absolute value by the normalized spatial norm $\|r\|_{\mathrm{sp}} = (64^{-1}\sum_{j=0}^{63} r_j^2)^{1/2}$. Its pair term averages eight disjoint member pairs, an unbiased but noisier

estimator of the pair expectation. Scores are averaged over times within each context before aggregation. A reference is used for evaluation only, not to choose generated trajectories.

The score's first term measures distance from generated values to the reference. Its second term accounts for spread among predictions; a probabilistic forecast is evaluated as a distribution rather than by selecting its closest sample. There are $M(M-1)$ ordered pairs of distinct members, which explains the denominator in the pair term. Excluding self-pairs avoids their artificial zero distances in an unbiased estimate of the expectation for two independent draws. The energy score makes the same comparison for the whole residual field using its spatial norm.

Residual energy averages r^2 over all indices. The increment averages $(r_{i,m,k+1,j} - r_{i,m,k,j})^2$. The reported uncentered lag correlation is

$$C_\ell = \frac{\langle r_{i,m,k,j} r_{i,m,k+\ell,j} \rangle}{\sqrt{\langle r^2_{i,m,k,j} \rangle \langle r^2_{i,m,k+\ell,j} \rangle}}, \qquad \ell = 1, 8,$$

where brackets pool all valid contexts, members, times, and spatial points. In particular, nonzero correlation in the independent model can arise from the common coarse context and changing conditional means. It does not contradict conditional independence.

For example, if $r_k = m_k(c) + \epsilon_k$ with independent centered errors conditional on a supplied coarse context c, then $\mathbb{E}[r_k r_l \mid c] = m_k(c) m_l(c)$ for $k \neq l$. The uncentered numerator can therefore be nonzero even though the conditional covariance is zero. Pooling several contexts introduces further variation in those means. The reported C_ℓ is this pooled raw-product statistic, so it must be read using its stated definition. For $J_{i,m} = n^{-1} \sum_k p_{i,m,k}$ we report the variance across contexts and members, using the population divisor. The reference variance uses its 128 context-level averages. This pooled quantity differs from the fixed-context variance in the controlled tests and is not a test of Corollary 8.

Conditional sample variance first computes the unbiased variance across the 16 generated members at a fixed context, time, and spatial point, then averages these values. Probe coverage is the fraction of reference values inside the empirical 5th–95th percentile interval of the 16 members. This finite-ensemble coverage is descriptive and does not certify conditional calibration. Fiber RMS is the square root of the average squared difference between generated block means and their supplied coarse values. The saved metric files include context-level standard errors for scores and increments, and a 200-resample context bootstrap for the time-average variance, retaining all members within a selected context. Neither time points nor generated members are treated as independent test contexts. Table 2 instead displays variation across training seeds on the shared test set, explicitly as descriptive sample SD.

Results and their limits. Table 2 summarizes the main comparison. Both methods recover block averages with RMS error below 2.3×10^{-8}, while their single-time distributions differ.

Table 2: Learned Burgers diagnostic, mean $\pm$ sample standard deviation across three training seeds on one test set. CRPS is the continuous ranked probability score for the residual at spatial point zero, with lower scores preferred. Increment is the mean squared residual change between consecutive times. Lag-one correlation compares consecutive residual fields and pools contexts, times, and spatial points without subtracting their means. J is the time average of the point-zero residual, with variance across test contexts and generated members. Reference values use test trajectories. Seed deviations are descriptive, not confidence intervals.

Method	CRPS (10^{-2})	Increment (10^{-3})	Lag-one correlation	$\mathrm{Var}(J)$ (10^{-3})
Reference	—	0.022	0.9977	5.66
Independent	3.31 ± 0.04	3.342 ± 0.150	0.613 ± 0.011	4.63 ± 0.37
Memory	4.33 ± 0.31	0.368 ± 0.217	0.992 ± 0.004	21.93 ± 6.33

Tables 3 and 4 retain every seed. Memory improves the residual lag correlations and reduces excessive increments, while increasing residual energy, field energy score, probe CRPS, and time-average variance. Its time-average variance is about 3.9 times the reference value when averaged across seeds. Thus retaining history is not sufficient for accurate temporal uncertainty. The independent decoder's empirical probe coverage is only about 57%, so this experiment does not have calibrated

or identical single-time distributions. It supports evaluating marginal quality, temporal increments, and accumulated uncertainty together. The controlled advection experiment provides the separate comparison with exactly matched marginals. Neither experiment establishes a general advantage of one learned architecture.

To read the scales in Table 2, the independent increment entry 3.342 means 3.342×10^{-3}, while the reference entry 0.022 means 2.2×10^{-5}. Memory brings this quantity closer to the reference but still above it. Its lag-one correlation is also closer to the reference. Meanwhile its larger CRPS is worse under that score, and its larger variance of J indicates excessive variation in the accumulated quantity. A high lag correlation alone therefore gives an incomplete assessment. The variance of J includes variation across supplied coarse paths; it is not the conditional variance at one fixed coarse initial condition used by the analytic examples.

Table 3: All learned-model seeds: marginal scores, residual energy, time-average variance, and empirical probe interval coverage. Lower CRPS and energy score indicate better predictive distributions under their respective scores. Other quantities are compared with the reference.

Method	Seed	CRPS	Energy score	Residual energy	$\mathrm{Var}(J)$	Coverage
Reference	—	—	—	0.00448	0.00566	—
Independent	0	0.03264	0.03056	0.00441	0.00471	0.58374
Memory	0	0.04285	0.04479	0.02369	0.02124	0.88428
Independent	1	0.03321	0.03083	0.00432	0.00424	0.57324
Memory	1	0.04046	0.04157	0.01654	0.01597	0.86768
Independent	2	0.03333	0.03077	0.00423	0.00496	0.54834
Memory	2	0.04667	0.04304	0.02124	0.02858	0.89478

Table 4: Optimization and temporal diagnostics for every learned model. NLL is validation loss per normalized coefficient and is specific to each model's conditioning. Correlations are pooled and uncentered. All models complete the fixed 4,000 updates.

Method	Seed	Initial NLL	Final NLL	Increment	Correlation (1 / 8)	Fiber RMS
Independent	0	1.533	0.914	0.003426	0.6110 / 0.5883	2.171×10^{-8}
Memory	0	1.084	-2.907	0.000529	0.9895 / 0.8494	2.253×10^{-8}
Independent	1	1.592	0.898	0.003431	0.6024 / 0.5780	2.163×10^{-8}
Memory	1	1.114	-2.899	0.000121	0.9969 / 0.8850	2.227×10^{-8}
Independent	2	1.524	0.924	0.003168	0.6246 / 0.6007	2.173×10^{-8}
Memory	2	1.101	-2.889	0.000454	0.9899 / 0.8506	2.229×10^{-8}

G.5 EXECUTION, CHECKS, AND REPRODUCIBILITY

Code availability. The `anonymous-code` package accompanies this submission as supplementary material. It contains the code for all five experiment groups in Section 6 and Appendices G–H. Paths in the instructions below are relative to the extracted package root. Its `README.md` gives installation and execution instructions, `EXPERIMENTS.md` maps the experiments to figures and tables, and `VALIDATION.md` records the checks performed and their limits. The package contains source code, fixed configurations, and numerical reference values for comparison. Data and trained checkpoints are generated during reproduction and are not supplied. The files under `reference/` are comparison targets only and are not used to generate samples, train models, compute metrics, or draw figures.

Environment and measured execution. The controlled experiments and the one-dimensional Burgers learning diagnostic use JAX/jaxlib 0.11.2, Python 3.12.14, Optax 0.2.8, CUDA 13, and one NVIDIA GeForce RTX 5090 Laptop GPU under Ubuntu 22.04 on WSL2. Controlled sampling uses float64. The Burgers solver and networks use float32 with the highest JAX matrix-multiplication precision. Each run verifies a just-in-time (JIT) compiled operation on the GPU and waits for its completion, records the device, driver, software versions, source/configuration hashes, seeds, and execution command, and refuses to overwrite an existing recorded run.

The runs used for the reported results took about 40.5 seconds for the controlled suite, which passed all 366 analytic and identity checks, and 33.0 seconds for the learned suite, including data generation, compilation, training, evaluation, and saving. Its six warmed training loops together took about 2.3 seconds. These are measured times for this small diagnostic, not performance comparisons with other methods. Separate post-run checks reproduce every saved probe ensemble from its checkpoint, verify the CRPS pair normalization, and execute the full fine-field advection rollout. The independent package validation includes post-run checks and plotting, with its timings reported separately in `VALIDATION.md`. First-call compilation and steady execution are recorded separately where benchmarked. The prescribed model budget is 900 seconds per model and seed, which is never reached.

The fixed settings are in `experiments/config.json`, and the numerical, verification, reporting, and plotting scripts are in `experiments/`. The full run uses the 4,000-step training protocol and retains all seeds without selecting a favorable outcome. Smoke settings provide smaller installation checks and do not reproduce the paper results. GPU float32 training is not bitwise deterministic.

JIT compilation translates a numerical program for the device on its first use, while a warmed execution reuses that compiled program. Timing a launched GPU operation without waiting would measure dispatch rather than the completed computation; the recorded runs synchronize explicitly. The smaller warmed-loop time and the total suite time therefore answer different timing questions. Reusing a seed records the intended random-number stream, but floating-point execution on the GPU can still produce small numerical differences. A complete repeat with the same seeds preserves the reported qualitative comparisons and is retained as a reproducibility check, without adding independent statistical evidence.

Reproducing the recorded evidence. After creating the Python 3.12 environment `.venv-jax` and installing `requirements-gpu.txt` as described in the README, run the following command from the package root:

```
.venv-jax/bin/python reproduce.py \
  --suite controlled-learning \
  --output-dir outputs-controlled-learning
```

The output directory must not already exist. This command runs all three controlled experiments, generates the Burgers data, trains all six models from initialization, evaluates them, checks the results, and produces the figures and tables. Adding `--dry-run` prints the commands without running experiments or creating outputs. Omitting `--suite` selects all experiments, including the two-dimensional PDE training, and also requires the separate interpreter described in Appendix H.5.

Within the chosen output directory, `controlled/` contains the statistics and analytic checks, while `learning/` contains the generated splits, normalization values, checkpoints, training histories, and evaluation arrays. The three `learning_*.tex` files in `tables/` correspond to Tables 2–4, and `figures/` contains Figures 2–5. The files `comparison.json` and `execution.json` record numerical comparisons and execution details. Repeated runs use separate output directories and do not count as extra independent seeds.

H PDE PROTOCOLS AND DETAILED RESULTS

This appendix describes the two-dimensional PDE experiments. Appendix G covers the controlled constructions and the one-dimensional Burgers learning diagnostic. Here we explain how a coarse field is reconstructed, how the reference and coarse simulations are produced, and what each reported error measures. Saved numerical failures remain part of the evidence.

H.1 SCOPE AND CHOICE OF IMPLEMENTATIONS

The experiments evaluate coarse recovery and endpoint reconstruction for implementations of $Q_t = G\phi_t R$. The five methods range from copying coarse values to learning random reconstructions with different ways of enforcing the coarse constraint. We report the same methods and all three seeds for every PDE, including runs with poor accuracy or numerical failures. This comparison concerns the effect of these reconstruction mechanisms. It does not identify the best generative PDE solver or test the temporal classification.

For every PDE, R averages each 4×4 block and I repeats a coarse value over its block. Thus $RI = \mathrm{Id}$ and $P_K = \mathrm{Id} - IR$ removes each block's mean. Its output lies in $K = \ker R = \{z : Rz = 0\}$, the space of residuals with zero block averages. The implementations are:

- Coarse: $u = Ic$, with no learned residual.
- Projected U-Net: $u = Ic + P_K f_\theta(c)$, trained as a deterministic squared-error residual predictor.
- Unprojected diffusion: $u = Ic + z_\theta(c)$, adding Gaussian noise to the full residual space without enforcing zero block averages.
- Post-hoc projection: $u = Ic + P_K z_\theta(c)$, using the same trained ambient diffusion checkpoint as the unprojected method and projecting only the final residual. Each method uses its own sampling seed, so the comparison does not pair projections of identical noise realizations.
- Fiber-native diffusion: $u = Ic + z$, where training noise and predicted noise are projected by P_K, and each reverse sampling update is projected onto $K = \ker R$. Appendix H.5 lists the corresponding source identifiers for all five methods.

The linear algebra can be checked block by block. Copying a coarse value and averaging it again returns that value, so $RI = \mathrm{Id}$. Consequently

$$RP_K = R - RIR = 0, \qquad P_K^2 = (\mathrm{Id} - IR)^2 = \mathrm{Id} - IR = P_K.$$

If $z \in K$, then $R(Ic + z) = c$. A fiber is the set of all fine fields satisfying this same constraint for a specified c; here it is $Ic + K$. Within a 4×4 block, specifying the mean fixes one of sixteen degrees of freedom and leaves fifteen residual degrees of freedom. The symbol P_K is a deterministic linear map on arrays; it is distinct from the probability kernel $P = GR$ in the classification theorem.

The five methods impose this constraint at different stages. The projected U-Net and post-hoc method remove the final residual's block means. Fiber-native diffusion also removes such components throughout training and reverse sampling. The unprojected method permits them. These design choices control recovery of the supplied coarse values, but they do not determine whether the generated fine field or its time dependence is accurate.

These implementations use established methods for enforcing exact fiber consistency in scientific downscaling and for diffusion within a linear subspace (Harder et al., 2023; Geiss and Hardin, 2023; Hoogeboom et al., 2022; Chen et al., 2023).

All learned models use a conditional multiresolution U-Net (Ronneberger et al., 2015) with base width 32, channel multipliers $[1, 2, 4, 4]$, two residual blocks per level, and time-embedding width 256. Coarse inputs and residual targets use scalar normalization. The projected U-Net uses $50{,}000$ optimizer steps, and each diffusion model uses $100{,}000$, with Adam, batch size 32, and learning rate 2×10^{-4}. Denoising diffusion probabilistic model (DDPM) training and denoising diffusion implicit model (DDIM) sampling follow Ho et al. (2020); Song et al. (2021). Diffusion uses $1{,}000$ training noise levels with a linear variance schedule from 10^{-4} to 0.02, and 100 DDIM sampling steps. The sampler finishes at noise index zero: its last update uses index zero for both the current and next

noise level, so it returns the index-zero residual rather than the predicted clean residual. The source preserves this terminal convention.

The diffusion noise index is an internal index used to generate a single reconstruction; it is different from the physical time in ϕ_t. DDPM training teaches a network to predict added noise at those indices. DDIM sampling uses the learned predictions to move from a noisy residual toward a reconstructed one. The conditional U-Net processes the field at several spatial resolutions and combines those features; its conditioning includes the coarse input and, for diffusion, the internal noise index. A hundred denoising updates produce one fine sample, not a hundred physical-time observations.

H.2 PDEs, DISCRETIZATIONS, AND TRAINING DATA

The notation ∂_t is a time derivative, ∇ collects spatial derivatives, and $\Delta = \partial_{xx} + \partial_{yy}$ is the Laplacian. Periodic boundary conditions identify opposite edges of the square. Diffusion terms smooth spatial variation, while the nonlinear terms differ between the three equations. The following numerical update formulas describe the supplied implementations, including the internal state that matters when a solver call is restarted. They are not assumptions that the numerical solvers satisfy an exact continuous-time semigroup law.

Allen–Cahn. On the periodic unit square, we solve

$$\partial_t u = \epsilon^2 \Delta u + u - u^3, \qquad \epsilon = 0.04.$$

The implementation uses an invariant energy quadratization (IEQ) linearization with a conjugate-gradient solve, time step 0.001, tolerance 10^{-7}, and a maximum of 200 conjugate-gradient iterations. A fresh solver call initializes its auxiliary variable from the supplied field. Consequently, several solver calls can give a different final field from one uninterrupted call over the same total time, even without reconstruction.

More explicitly, set $F(u) = \frac{1}{4}(u^2-1)^2$, initialize $q = \sqrt{F(u)+1}$ pointwise, and define $H(u) = (u^3-u)/\sqrt{F(u)+1}$. With time step δt and the periodic five-point discrete Laplacian Δ_{grid}, an update solves

$$\begin{aligned}\left(1 + \tfrac{1}{2}\delta t\, H(u)^2 - \delta t\, \epsilon^2 \Delta_{\mathrm{grid}}\right) u^+ &= \left(1 + \tfrac{1}{2}\delta t\, H(u)^2\right) u - \delta t\, H(u) q,\\ q^+ &= q + \tfrac{1}{2} H(u)(u^+ - u).\end{aligned}$$

Multiplication by $H(u)$ is pointwise, and the linear system for u^+ is solved iteratively by conjugate gradients. Within a call, the algorithm carries q forward using its update. Restarting sets q from the new input field instead. The two routes can therefore differ in their internal state even if their total elapsed physical times agree. The parameter ϵ here is the Allen–Cahn equation parameter, distinct from the recovery-error bound ε in the theory.

Scalar viscous Burgers. On the periodic domain $[0, 2\pi]^2$, we solve

$$\partial_t u + \partial_x(u^2/2) + \partial_y(u^2/2) = \nu \Delta u, \qquad \nu = 0.01.$$

Advection uses a Rusanov finite-volume flux and diffusion uses a semi-implicit Fourier update, with a fixed time step 0.005. The state is scalar. This is not a two-component velocity formulation.

Writing $f(u) = u^2/2$, the numerical flux between neighboring values u_L, u_R is

$$\mathcal{F}(u_L, u_R) = \tfrac{1}{2}(f(u_L) + f(u_R)) - \tfrac{1}{2}\max(|u_L|, |u_R|)(u_R - u_L).$$

Differences of these interface fluxes in the two coordinate directions give a discrete divergence $D_{\mathrm{flux}}(u)$. The step first forms $b = u - \delta t D_{\mathrm{flux}}(u)$ and then solves $(\mathrm{Id} - \delta t\, \nu \Delta_{\mathrm{grid}})u^+ = b$. On a periodic grid, the Fourier transform diagonalizes this discrete Laplacian, so the solve is division by its mode-dependent diagonal factors. The flux uses the known field explicitly, while the diffusion solve uses the new field implicitly; this is the meaning of semi-implicit here.

Navier–Stokes. The state is scalar vorticity on the periodic domain $[0, 2\pi]^2$:

$$\partial_t \omega + v \cdot \nabla \omega = \nu \Delta \omega + 0.2 \sin(4y), \qquad -\Delta \psi = \omega - \langle \omega \rangle, \qquad v = (\partial_y \psi, -\partial_x \psi), \quad \nu = 0.002.$$

The pseudo-spectral solver uses an implicit–explicit (IMEX) Euler startup followed by a two-step Adams–Bashforth update (IMEX-AB2), with time step 0.005 and $2/3$ dealiasing of the state and nonlinear-plus-forcing term. Each new call restarts the AB2 history. The Poisson zero mode is set to zero. Velocity is reconstructed for the solver, while the generator predicts vorticity.

The streamfunction ψ supplies velocity by differentiation, and subtracting $\langle\omega\rangle$ removes the spatial mean so that the periodic Poisson equation is solvable. Fourier coefficients are denoted by a hat. Let N_j be the Fourier transform of $-v_j \cdot \nabla\omega_j + 0.2\sin(4y)$ after the stated mode filtering. At a noninitial step the code uses

$$\widehat{\omega}_{j+1}(k) = \frac{\widehat{\omega}_j(k) + \delta t(\frac{3}{2}N_j(k) - \frac{1}{2}N_{j-1}(k))}{1 + \delta t\,\nu|k|^2}.$$

The first step instead uses N_j alone in the numerator. This is the IMEX Euler startup; subsequent steps use the two-step Adams–Bashforth combination. Restarting a solver call discards the preceding nonlinear term and uses the startup rule again. The $2/3$ filter keeps modes with each integer coordinate bounded by one third of the grid size; it suppresses unresolved high-frequency contributions created by products on a discrete grid. Derivatives are calculated through Fourier coefficients, while nonlinear products are formed in physical space, which explains the term pseudo-spectral.

Data and conditioning. All PDEs use fine/coarse grid sizes $128^2/32^2$, a training horizon $T = 1$, and 26 stored times, including zero. Initial fields use random Fourier modes up to wave number 8 and decay power 2. Allen–Cahn and Burgers use amplitude 0.8 and clipping level 1. Navier–Stokes uses amplitude and clipping level 2, with maximum-velocity scaling at 1.5. Each run has $2{,}000$ training, 200 validation, and 200 test trajectories. Only the first 100 test trajectories are used for the reported evaluation.

For a fine trajectory $u_h(t)$, the stored quantities are

$$c_h(t) = Ru_h(t), \qquad c_\ell(t) = \widehat{\phi}_t(Ru_h(0)), \qquad z_h(t) = u_h(t) - Ic_h(t).$$

Here $\widehat{\phi}_t$ denotes the numerical coarse solver. Training pairs are $(c_\ell(t), z_h(t))$ at nonzero saved times. Evaluation conditions on $c_\ell(1)$ and compares generated states with $u_h(1)$. In general $c_\ell(t) \neq c_h(t)$: the reference field can have different block averages from those prescribed to the generator. These endpoint comparisons therefore do not test whether the generated conditional distribution matches the reference distribution at a fixed terminal coarse state.

The distinction is visible by decomposing the reference exactly as $u_h(t) = Ic_h(t) + z_h(t)$. A constrained generator conditioned on $c_\ell(t)$ instead produces $Ic_\ell(t) + z$. Their difference is

$$I(c_\ell(t) - c_h(t)) + (z - z_h(t)).$$

The first term comes from the different coarse evolutions, and the second from residual reconstruction. Thus even an excellent residual predictor does not remove an existing mismatch between the supplied coarse state and the block averages of the reference. This helps interpret endpoint errors without attributing every discrepancy to the decoder.

H.3 Metric definitions and aggregation

For a generated field u conditioned on c and its fine reference u^{ref}, we report

$$e_{\mathrm{fiber}} = \frac{\|Ru - c\|_2}{\|c\|_2 + 10^{-12}}, \qquad e_{L^2} = \frac{\|u - u^{\mathrm{ref}}\|_2}{\|u^{\mathrm{ref}}\|_2 + 10^{-12}}.$$

The endpoint evaluation contains $100 \times 8 = 800$ rows per method and seed. Deterministic methods repeat the same prediction over the sample index. These repeats are not independent samples. We average the endpoint L^2 values within each seed, then report the average and sample standard deviation of those three means. Fiber error is the average of the three within-seed medians. All $36{,}000$ endpoint entries are finite for both reported metrics. The relative L^2 error compares individual generated and reference fields. It does not determine whether their probability distributions agree.

The small constant 10^{-12} prevents division by zero. The first metric asks whether restriction of a reconstruction returns its supplied input. The second asks how far one generated field is from one paired reference field. For example, two independent samples from the same nondegenerate

distribution normally have positive squared distance, even though their laws agree exactly. Conversely, a single small pairwise error is not evidence that two whole distributions agree.

The count of endpoint entries is 3 PDEs times 5 methods times 3 seeds times 100 test cases times 8 sample indices, giving 36,000. The mean and median summaries serve different purposes: the mean includes the magnitude of every error, while the median is the middle value after sorting. The sample standard deviation across three seed summaries describes how those three numbers vary; it does not automatically supply a confidence interval for performance on new data.

The standard deviation across run labels is descriptive. Allen–Cahn and Burgers evaluations share substantial test content across seeds: exact tuples of saved coarse-reference diagnostic values overlap for 99, 98, and 99 of the 100 test cases in the $(0, 1)$, $(0, 2)$, and $(1, 2)$ pairs, respectively. These labels therefore do not correspond to three independent test sets, and the displayed standard deviations are not uncertainty estimates for a population ranking.

For composition, one stochastic realization per initial state is propagated over intervals of length $\tau = 0.04$. At each interval we restrict the current generated field, solve one coarse interval, and regenerate. A separate direct coarse call advances the original coarse state through the total elapsed time. With u_K the repeated output and c_K^{direct} the direct result, the diagnostic is

$$D(K) = \frac{\|Ru_K - c_K^{\mathrm{direct}}\|_2}{\|c_K^{\mathrm{direct}}\|_2 + 10^{-12}}.$$

Thus $K = 32$ and 64 correspond to $T = 1.28$ and 2.56 for all three PDEs, both beyond the training horizon $T = 1$. The first is a fixed repeated-composition diagnostic and the second a longer stress case. Neither is presented as an in-horizon accuracy measurement.

For clarity, the repeated calculation uses

$$c_k^{\mathrm{step}} = \widehat{\phi}_\tau(Ru_{k-1}), \qquad u_k \sim G(c_k^{\mathrm{step}}), \qquad c_K^{\mathrm{direct}} = \widehat{\phi}_{K\tau}(Ru_0).$$

The diagnostic compares Ru_K with the last quantity. If recovery were exact and the numerical coarse solver composed exactly, induction would make them equal. In the actual implementation, finite arithmetic and solver restart conventions also affect the comparison. Moreover, $D(K)$ discards the fine residual before measuring the error, so it cannot detect every discrepancy between fine-state distributions.

We summarize $D(32)$ by the average of the finite-case median within each seed, always reporting nonfinite cases alongside it. A finite drift value means only that the scalar diagnostic is finite. It does not certify a physically valid or accurate field. At $K = 64$, Table 5 gives the finite counts and finite-case medians separately for every seed. In particular, a median over seeds would conceal the extreme third Navier–Stokes fiber-native run.

Fiber-native Navier–Stokes has finite $D(64)$ counts $(6, 44, 4)$ out of $(100, 100, 100)$. The third run's finite median is 8.80×10^6. Thus only $54/300$ drift values remain finite after $K = 64$ reconstruction steps, a fraction of 0.18. Unprojected Navier–Stokes has counts $(97, 99, 100)$, while the remaining three methods have $(100, 100, 100)$. Fiber-native Allen–Cahn has counts $(97, 100, 97)$, and unprojected Burgers has $(99, 100, 100)$. Enforcing a linear fiber constraint does not ensure numerical rollout stability.

Nonfinite entries are values such as overflow to infinity or undefined floating-point values. They cannot be meaningfully ordered with ordinary errors to form a finite median, so the table reports their loss separately through finite counts. A method with only four surviving trajectories and a finite median is not comparable to one with all one hundred trajectories surviving unless this difference is made explicit. Even very large numbers such as 8.80×10^6 are finite and remain in the table; finiteness alone is a weak numerical condition.

The direct and repeated coarse computations also differ without generative randomness. In Navier–Stokes, resetting the AB2 history at each interval changes the numerical update relative to one uninterrupted solve, while in Allen–Cahn the auxiliary IEQ variable is reinitialized. Computing block averages and copying them back to the fine grid adds further small floating-point errors. Accordingly, the constrained methods' nonzero $D(K)$ cannot be attributed entirely to fiber error, and $D(K)$ is not a distance between the two fine-state transition kernels. These finite-grid diagnostics neither establish the exact positive-time semigroup law nor verify the regularity hypotheses of the mathematical results. They also do not establish that generated joint distributions match those of the target dynamics.

Table 5: Finite diagnostics at $K = 32, 64$ for all methods. Every count is out of 100 rollouts for the indicated seed. The $D(64)$ medians use only finite entries and must be read together with their counts. All seeds and extreme finite values are included.

Lifter	Seed	Finite $D(32)$	Finite $D(64)$	Median finite $D(64)$
Allen–Cahn				
Coarse	0	100/100	100/100	2.81×10^{-6}
	1	100/100	100/100	2.82×10^{-6}
	2	100/100	100/100	2.82×10^{-6}
Projected U-Net	0	100/100	100/100	2.81×10^{-6}
	1	100/100	100/100	2.82×10^{-6}
	2	100/100	100/100	2.82×10^{-6}
Unprojected diffusion	0	100/100	100/100	1.53×10^{-2}
	1	100/100	100/100	1.65×10^{-2}
	2	100/100	100/100	7.60×10^{-3}
Post-hoc projection	0	100/100	100/100	2.81×10^{-6}
	1	100/100	100/100	2.82×10^{-6}
	2	100/100	100/100	2.82×10^{-6}
Fiber-native diffusion	0	100/100	97/100	2.82×10^{-6}
	1	100/100	100/100	2.82×10^{-6}
	2	100/100	97/100	2.82×10^{-6}
Burgers				
Coarse	0	100/100	100/100	6.67×10^{-8}
	1	100/100	100/100	6.67×10^{-8}
	2	100/100	100/100	6.67×10^{-8}
Projected U-Net	0	100/100	100/100	1.11×10^{-7}
	1	100/100	100/100	1.08×10^{-7}
	2	100/100	100/100	1.10×10^{-7}
Unprojected diffusion	0	100/100	99/100	1.13×10^{-1}
	1	100/100	100/100	1.84×10^{-2}
	2	100/100	100/100	2.42×10^{-2}
Post-hoc projection	0	100/100	100/100	1.08×10^{-7}
	1	100/100	100/100	1.09×10^{-7}
	2	100/100	100/100	1.09×10^{-7}
Fiber-native diffusion	0	100/100	100/100	1.09×10^{-7}
	1	100/100	100/100	1.09×10^{-7}
	2	100/100	100/100	1.09×10^{-7}
Navier–Stokes				
Coarse	0	100/100	100/100	4.22×10^{-4}
	1	100/100	100/100	4.20×10^{-4}
	2	100/100	100/100	4.33×10^{-4}
Projected U-Net	0	100/100	100/100	4.22×10^{-4}
	1	100/100	100/100	4.20×10^{-4}
	2	100/100	100/100	4.33×10^{-4}
Unprojected diffusion	0	100/100	97/100	9.33×10^{-2}
	1	100/100	99/100	2.86×10^{-2}
	2	100/100	100/100	3.15×10^{-2}
Post-hoc projection	0	100/100	100/100	4.22×10^{-4}
	1	100/100	100/100	4.20×10^{-4}
	2	100/100	100/100	4.33×10^{-4}
Fiber-native diffusion	0	99/100	6/100	3.87×10^{-4}
	1	100/100	44/100	3.91×10^{-4}
	2	99/100	4/100	8.80×10^{6}

H.4 ENDPOINT TABLE AND ALL RECORDED ROLLOUT HORIZONS

Table 6 reports the endpoint and $D(32)$ summaries. Figure 6 analyzes the saved CSVs at all recorded $K \in \{1, 2, 4, 8, 16, 32, 64\}$, with interval $\tau = 0.04$. It includes all five methods and all three seeds for each PDE, giving 315 method–seed–horizon groups of 100 initial states. All horizons use the same trained models. Every finite-case median is shown alongside its finite fraction. Curves therefore reveal when a small surviving subset accompanies a large or apparently moderate drift statistic. The training horizon is marked separately from later stress evaluations.

Table 6: PDE evaluations. Fiber error is the average of the three within-seed medians at $T = 1$. Relative L^2 is the mean $\pm$ sample SD of the three within-seed means. $D(32)$ averages within-seed medians of finite rollout diagnostics at $T = 1.28$. These seed summaries are descriptive, not confidence intervals. All endpoint entries are finite. Navier–Stokes fiber-native diffusion has $298/300$ finite $D(32)$ entries.

PDE	Lifter	Fiber error	Relative L^2 ($\times 10^{-3}$)	$D(32)$
Allen–Cahn	Coarse	0	64.273 ± 0.054	2.22×10^{-6}
	Projected U-Net	3.30×10^{-8}	2.643 ± 0.009	2.23×10^{-6}
	Unprojected diffusion	3.61×10^{-4}	2.753 ± 0.072	7.42×10^{-3}
	Post-hoc projection	3.31×10^{-8}	2.764 ± 0.056	2.23×10^{-6}
	Fiber-native diffusion	3.31×10^{-8}	2.751 ± 0.063	2.22×10^{-6}
Burgers	Coarse	0	96.466 ± 0.152	6.03×10^{-8}
	Projected U-Net	3.60×10^{-8}	56.711 ± 0.056	9.58×10^{-8}
	Unprojected diffusion	7.18×10^{-4}	56.787 ± 0.095	1.80×10^{-2}
	Post-hoc projection	3.58×10^{-8}	56.747 ± 0.039	9.60×10^{-8}
	Fiber-native diffusion	3.57×10^{-8}	56.756 ± 0.045	9.59×10^{-8}
Navier–Stokes	Coarse	0	84.562 ± 0.426	2.24×10^{-4}
	Projected U-Net	3.44×10^{-8}	39.461 ± 0.617	2.24×10^{-4}
	Unprojected diffusion	9.13×10^{-4}	39.442 ± 0.661	2.51×10^{-2}
	Post-hoc projection	3.44×10^{-8}	39.478 ± 0.620	2.24×10^{-4}
	Fiber-native diffusion	3.44×10^{-8}	39.496 ± 0.636	2.23×10^{-4}

How to read the endpoint and horizon results together. In Table 6, constrained learned methods have recovery errors near floating-point precision, while unprojected diffusion has larger recovery errors. Their relative L^2 columns do not show a consistent advantage for one constrained diffusion choice across all three equations. A small $D(32)$ also does not settle longer-horizon behavior: Figure 6 must be read together with its finite-fraction panels and Table 5. In particular, the loss of finite Navier–Stokes trajectories at $K = 64$ is part of the result. The evidence supports separately reporting constraint satisfaction, field error, and numerical survival over time.

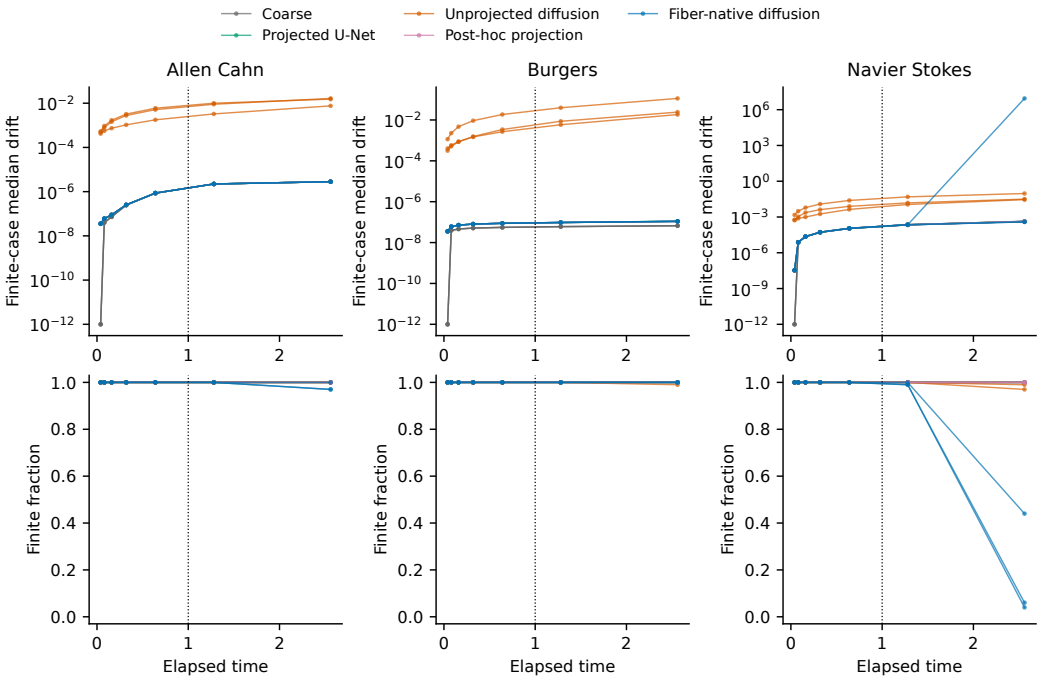

Figure 6: All recorded PDE rollout horizons. Each method has three lines, one per seed. Top: finite-case median drift $D(K)$. Values below 10^{-12}, including exact zeros, are drawn at 10^{-12} for the logarithmic axis. Bottom: fraction of finite diagnostics. The vertical dotted line marks the training horizon $T = 1$. The plot retains extreme finite values and failures, including the severe loss of finite fiber-native Navier–Stokes diagnostics at $K = 64$.

H.5 Reproduction and release scope

The same `anonymous-code` package described in Appendix G.5 contains these experiments under `pde/`. Its three benchmark directories, `allen_cahn`, `burgers`, and `navier_stokes`, are under `pde/benchmarks/`. Each contains the solver, data generation, training, sampling, and evaluation code, with the paper settings in its `configs/` directory. These experiments use JAX solvers and PyTorch models. The reference environment used Python 3.10.12, NumPy 1.26.4, JAX/jaxlib 0.4.34, PyTorch 2.5.1 with CUDA 12.4, and an NVIDIA RTX A6000. The package specifies this environment separately from the Python 3.12 environment used in Appendix G.

After following the root README to create `.venv-pde` with `pde/requirements-cuda12.txt`, use both environments from the package root:

```
.venv-jax/bin/python reproduce.py --suite pde \
  --python-pde .venv-pde/bin/python --output-dir outputs-pde
```

The output directory must not already exist. The command runs all three PDEs and seeds, generating the data and training the 27 required checkpoints before sampling and evaluation. Each PDE and seed trains a projected U-Net, an ambient diffusion model, and a fiber-native diffusion model in the prescribed order. Post-hoc projection uses the ambient checkpoint. This order preserves the random-key sequence. The top-level runner then uses the Python 3.12 environment for horizon analysis and plotting. The file `pde/README.md` explains individual stages and table aggregation.

The deterministic lifters use the source identifiers `coarse` and `projected_unet`. The diffusion variants use `unprojected_diffusion`, `posthoc_projection`, and `ours_fiber_native`. The final identifier is part of the random-key and checkpoint conventions.

Within the chosen output directory, `pde/` contains the generated data, checkpoints, samples, and evaluation CSVs. The files `pde/paper_tables/empirical_stress_table.tex` and `pde/paper_tables/empirical_main_table.tex` correspond to Tables 5 and 6, respectively. The file `pde_analysis/all_horizons.csv` contains all 315 method–seed–horizon groups, and `figures/pde_horizons.pdf` reproduces Figure 6. Once a run has generated its evaluation metrics, the analysis scripts can regenerate its tables and plots without training again. The supplied package contains no datasets or trained weights, so its full reproduction command generates them from the stated configurations.

The package validation checked table aggregation and all 315 horizon groups against the recorded measurements. It did not rerun the full two-dimensional PDE data generation, training, and sampling workload. The executed checks and the separately validated controlled and Burgers experiments are documented in `VALIDATION.md`.