CONDITIONAL FLOW MATCHING FOR TRANSPORT BETWEEN MARKOV PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by sequence-to-sequence transport in the context time-series domain adaptation, we study the problem of transportation between trajectories of Markov processes. Given a limited number of trajectories from source distribution and the target distribution, we formulate a flow matching based algorithm which learns a transport map from the source to target trajectory distribution, while preserving the Markov structure. We show that this is consistent in the population limit and derive finite-sample error bounds under mixing time assumptions, following the analysis of classical statistical problems including regression (Nagaraj et al., 2020), principal component analysis (Kumar & Sarkar, 2023), and matrix concentration (Neeman et al., 2024) in the Markov setting. We complement that with a lower-bound construction showing that a mixing-time dependent sample complexity is unavoidable even with regular Gaussian conditional transitions. We evaluate on synthetic and real-world data. For image retrieval from electroencephalography (EEG) on THINGS-EEG2 (Gifford et al., 2022), the task is to identify the viewed image from EEG signals captured from human subjects, which suffers from high inter subject variability. We augment the ENIGMA decoder (Kneeland et al., 2026) with a conditional flow before its subject-specific temporal map. This improves mean top-5 retrieval accuracy from 43.87% to 49.05%, an 11.82% relative improvement.

1 INTRODUCTION

Sequence-to-sequence tasks seek to learn a map from an input sequence to an output sequence. In time-series adaptation, such a map can transform observations from a source process to match the statistical behavior of a target process. Related transport and alignment problems arise in sensor calibration (Kim et al., 2020), climate-model bias correction (Robin et al., 2019), and time-series domain adaptation (Painblanc et al., 2023; Mishra et al., 2026).

In many of these cases, the sequences themselves are long, but the number of sequences available from the source and target distributions is limited. A practical concern is how much data must be collected to learn this map for a subject or device. Sensor calibration can require repeated, costly data collection (Kim et al., 2020), while neural decoding motivates adapting pretrained models using limited recordings from each new subject. ENIGMA and MindEye2 pursue this objective for EEG and functional magnetic resonance imaging (fMRI), respectively (Kneeland et al., 2026; Scotti et al., 2024). These constraints motivate studying adaptation from one or a few observed trajectories. Although each trajectory contains many observations, those observations are temporally dependent. This raises the question:

*Can we learn, from **one** source and **one** target trajectory, a map that transforms a fresh source trajectory into one that follows the target dynamics?*

In this paper, we study this problem for Markovian trajectories. When both the source and target follow Markov processes, we want the learned mapping to preserve the target's transition law and hence its Markov structure in the generated trajectory. At first glance, this task seems daunting: treating an entire trajectory as one high-dimensional datapoint leaves us with only one sample from each trajectory distribution. However, the Markov property and time homogeneity let us learn a common transition rule from successive observations. Under stationarity, these observations share a common transition-pair distribution, although they remain dependent. We exploit this structure to give a principled answer to the question above.

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.

In particular, we propose an algorithm based on conditional flow matching for transporting a source trajectory to a target trajectory, accounting for the source history through the current source state. We prove exact population-level transport under suitable assumptions, provide finite-sample error bounds, and evaluate the algorithm on real and simulated data.

Transport between Markov chains has also been studied through cost-minimizing joint distributions over transitions and trajectories (O'Connor et al., 2022; Moulos, 2021). Our objective is to learn a recursively reusable generator from one source and one target trajectory.

Formally, let $(\mathbf{x}_i)_{i\geq 0}$ and $(\mathbf{y}_i)_{i\geq 0}$ be independent, time-homogeneous Markov processes on $\mathbb{R}^d$ with transition kernels P_X and P_Y. We observe trajectories $\mathbf{x}_{0:m}$ and $\mathbf{y}_{0:m}$. At test time, we receive a fresh source trajectory $\mathbf{x}'_{0:n}$ with transition kernel P_X and seek to map it to a trajectory $\widehat{\mathbf{y}}_{0:n}$ with transition kernel P_Y. Given the source trajectory, we map each next source state conditional on the current source and generated target states. Lemma 1 shows that, conditional on their joint history, the next source state still follows the source transition kernel. Thus maps that transport each source transition to the corresponding target transition can be applied recursively. We learn a single family of maps $\Phi_{\mathbf{x},\mathbf{y}} : \mathbb{R}^d \to \mathbb{R}^d$ and generate recursively by

$$\widehat{\mathbf{y}}_{i+1} = \Phi_{\mathbf{x}'_i,\widehat{\mathbf{y}}_i}(\mathbf{x}'_{i+1}). \tag{1}$$

Histogram or marginal matching matches state frequencies alone and need not preserve the transition law. For e.g., consider two Markov processes on $\{0, 1\}$ that are equally likely to be in either state. If the source repeats its current state with probability 0.9 while the target switches with probability 0.9, their state frequencies agree, with very different transition behaviour.

Our approach. To learn the transition map, we use conditional flow matching (CFM) (Lipman et al., 2023; Liu et al., 2023), regressing the next-state displacement $\mathbf{y}^+ - \mathbf{x}^+$ on $\mathbf{z}_t = (1-t)\mathbf{x}^+ + t\mathbf{y}^+$, with flow time $t \in [0, 1]$ and current states $(\mathbf{x}, \mathbf{y})$ as conditioning inputs. Here $\mathbf{x}^+ \sim P_X(\mathbf{x}, \cdot)$ and $\mathbf{y}^+ \sim P_Y(\mathbf{y}, \cdot)$ are independent given the current states. At test time, integrating the learned velocity from $\mathbf{x}'_{i+1}$ with $(\mathbf{x}'_i, \widehat{\mathbf{y}}_i)$ fixed implements the recursive update (1). Our analysis quantifies how transition mixing time affects the required training trajectory length. Section 3 defines the conditional velocity and its ODE map, and Figure 1 summarizes the algorithm.

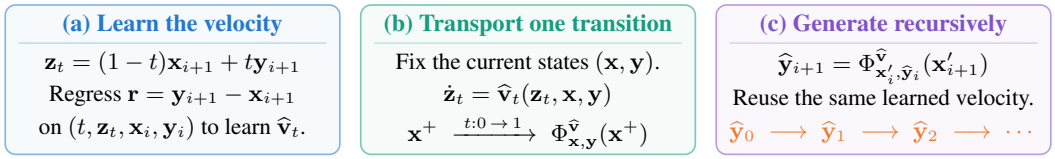


Figure 1: Each learned flow transports one source transition to one target transition. The index i denotes time in the Markov process, while $t \in [0, 1]$ denotes flow time within a single transition.

1.1 Our results

Population correctness. Propositions 1 and 2 prove exact generation under suitable assumptions. The population flow maps each source transition $P_X(\mathbf{x}, \cdot)$ to the target transition $P_Y(\mathbf{y}, \cdot)$. Initialized from the stationary target distribution independently of the fresh source trajectory, its recursive application in (1) produces the exact target trajectory law.

Sample complexity. Theorem 1 bounds the squared velocity error of least-squares estimation, integrated over flow time, from two independent stationary trajectories. For a finite candidate class $\mathcal{V}$ containing the population velocity, under suitable assumptions, error at most $\epsilon > 0$ holds with probability at least $1 - \delta$, for $\delta \in (0, 1)$, if

$$m \geq CK \max\{1, \epsilon^{-1}\} \log\left(\frac{6K|\mathcal{V}|}{\delta}\right), \tag{2}$$

where m counts transitions per trajectory, $K = \tau_{\text{mix}}(\delta/(3m))$ is the observation spacing in Theorem 1, and C depends only on the assumption parameters. Compared with independent-sample analyses (Zhou & Liu, 2025; Gaur et al., 2025), dependence replaces the sample size m by m/K,

up to logarithmic factors, as in Markovian regression (Nagaraj et al., 2020), streaming principal component analysis (Kumar & Sarkar, 2023), and matrix concentration (Neeman et al., 2024).

Lower bound. Proposition 3 adapts the lower-bound argument of Nagaraj et al. (2020, Appendix C.1) to Gaussian conditional transitions. Even with only two candidate velocities, achieving small squared velocity error with high probability requires a number of transitions proportional to the mixing time.

Accuracy of generated states and trajectories. Theorem 2 in Appendix A.5 bounds generation error under a coverage condition on generated state pairs. For contractive target dynamics and coverage uniform over the horizon, the state-marginal error bound has a decaying initialization term and an estimation term bounded independently of trajectory length. Proposition 5 in the same appendix bounds trajectory Wasserstein distance with transport cost equal to the sum of squared state distances.

Empirical evaluation. Section 5 evaluates Gaussian and nonlinear transition recovery and supervised EEG-to-image retrieval. For EEG, we freeze the ENIGMA encoder and projection network pretrained on other subjects (Kneeland et al., 2026; Alljoined, 2025). Using 4,096 calibration images per subject, we compare adapting only the subject-specific temporal map with jointly training that map and a preceding width-64 flow conditioned on the 32 previous EEG samples. This raises mean top-5 retrieval accuracy from 43.87% to 49.05%, an 11.82% relative improvement.

1.2 Related work

Kerrigan et al. (2024a) and Chemseddine et al. (2025) construct flows between conditional distributions. For Markov chains, Moulos (2021) and O'Connor et al. (2022) optimize trajectory and transition couplings through dynamic programming, while O'Connor et al. (2024) estimate stationary transport plans and their costs from observations. We use conditional flow matching to learn a continuous transition map from one source and one target trajectory. Conditioning on current source and target states lets us reuse this map to generate from a fresh source trajectory.

Zhou & Liu (2025) and Gaur et al. (2025) give finite-sample error bounds for neural flow matching trained on independent samples. For sequential data, Lim et al. (2026) express the empirical flow-matching solution as a similarity-weighted average of observed transition velocities and prove convergence as the trajectory grows. We give finite-sample bounds for learning from dependent transitions: for a finite candidate class, we bound the number of transitions needed for a prescribed velocity error in terms of the transition mixing time, and prove that a mixing-dependent sample threshold is necessary. Appendix F surveys further related work.

2 Preliminaries and notation

Notation. Here, a Markov process is discrete-time, possibly continuous state-space, and a trajectory is a finite sequence of its states. Vectors and vector-valued functions use bold lowercase letters, e.g. $\mathbf{x} \in \mathbb{R}^d$ and the velocity field $\mathbf{v}_t$. Matrices use bold uppercase letters, e.g. $\mathbf{A} \in \mathbb{R}^{d\times d}$. Scalars, indices, scalar-valued functions, and Markov kernels use ordinary italics. We use $\|\cdot\|$ for the Euclidean norm for vectors and $\|\mathbf{A}\|_{\mathrm{F}} = (\sum_{j,k} A_{jk}^2)^{1/2}$ for the Frobenius norm of a matrix. For a random vector and $p \geq 1$, $\|\mathbf{u}\|_{L^p} = (\mathbb{E}\|\mathbf{u}\|^p)^{1/p}$. Distributions are probability measures; densities are not assumed to exist. We write $\mathrm{Law}(\mathbf{u})$ for the law of $\mathbf{u}$, so $\mathrm{Law}(\mathbf{u})(A) = \mathbb{P}(\mathbf{u} \in A)$ for every measurable set A, and $g_{\#}\mu$ for the law of $g(\mathbf{u})$ when $\mathbf{u} \sim \mu$. The quadratic Wasserstein distance $W_2(\mu, \nu)$ is the square root of the minimum $\mathbb{E}\,\|\mathbf{u} - \mathbf{w}\|^2$ over joint distributions with $\mathbf{u} \sim \mu$ and $\mathbf{w} \sim \nu$; such a joint distribution is called a *coupling* (Villani, 2009, Chapter 6). We also write $\|\mu - \nu\|_{\mathrm{TV}} = \sup_A |\mu(A) - \nu(A)|$ for total-variation distance. Time is indexed by i, while $t \in [0, 1]$ denotes flow time within one transition. We write $\mathcal{U}(0, 1)$ for the uniform distribution on $[0, 1]$.

Markov processes and mixing. For a Markov kernel P, $P(\mathbf{x}, \cdot)$ is the distribution of the next state given the current state $\mathbf{x}$. The first-order Markov property means $\mathrm{Law}(\mathbf{x}_{i+1} \mid \mathbf{x}_{0:i}) = P(\mathbf{x}_i, \cdot)$; time homogeneity means that P does not depend on time index i. A distribution π is stationary when $\int P(\mathbf{x}, A)\pi(\,\mathrm{d}\mathbf{x}) = \pi(A)$ for every set A. For a stationary Markov process $(\mathbf{x}_i)_{i\geq 0}$, initialized with $\mathbf{x}_0 \sim \pi$, each $\mathbf{x}_i$ has distribution π. Using standard definitions of mixing (Berbee, 1979; Yu, 1994; Bradley, 2005), let

$$\beta_\ell = \mathbb{E}_{\mathbf{x}\sim\pi} \left\|\mathrm{Law}(\mathbf{x}_\ell \mid \mathbf{x}_0 = \mathbf{x}) - \pi\right\|_{\mathrm{TV}}, \qquad \tau_{\mathrm{mix}}(\alpha) = \min\{\ell \geq 1 : \beta_\ell \leq \alpha\}. \tag{3}$$

We write $\tau_{\mathrm{mix}} = \tau_{\mathrm{mix}}(1/4)$ and call it the mixing time; $\tau_{\mathrm{mix}}(\alpha)$ denotes the mixing time at any other tolerance $\alpha \in (0,1)$. If $\beta_\ell \le c_0 e^{-\ell/\tau_0}$, then $\tau_{\mathrm{mix}}(\alpha) \le \max\{1, \lceil \tau_0 \log(c_0/\alpha) \rceil\}$.

Remark (Comparison with worst-case mixing). Let $d_\ell^{\mathrm{wc}} = \sup_{\mathbf{x}} \|\mathrm{Law}(\mathbf{x}_\ell \mid \mathbf{x}_0 = \mathbf{x}) - \pi\|_{\mathrm{TV}}$ and $\tau_{\mathrm{wc}}(\alpha) = \min\{\ell \ge 1 : d_\ell^{\mathrm{wc}} \le \alpha\}$, the standard worst-case total-variation mixing time (Levin & Peres, 2017, Chapter 4). Since an average is at most a supremum, $\beta_\ell \le d_\ell^{\mathrm{wc}}$ and therefore $\tau_{\mathrm{mix}}(\alpha) \le \tau_{\mathrm{wc}}(\alpha)$.

Flow matching. Let μ_0 and μ_1 be distributions on $\mathbb{R}^d$ with finite second moments. Draw $\mathbf{u}_0 \sim \mu_0$ and $\mathbf{u}_1 \sim \mu_1$ independently, and let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be any smooth compactly supported scalar function.

$$\mathbf{z}_t = (1-t)\mathbf{u}_0 + t\mathbf{u}_1, \qquad \mathbf{v}_t^\star(\mathbf{z}) = \mathbb{E}[\mathbf{u}_1 - \mathbf{u}_0 \mid \mathbf{z}_t = \mathbf{z}], \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\varphi(\mathbf{z}_t)] = \mathbb{E}\left[\langle \nabla\varphi(\mathbf{z}_t), \mathbf{v}_t^\star(\mathbf{z}_t)\rangle\right]. \tag{4}$$

The interpolation $\mathbf{z}_t$ joins the sampled endpoints by a straight line. The velocity $\mathbf{v}_t^\star(\mathbf{z})$ averages the endpoint displacements of paths passing through $\mathbf{z}$ at time t, giving the local direction and speed of transport. The function φ tests the evolving interpolation law through its expectation; smooth compact support means that φ is infinitely differentiable and vanishes outside a bounded set. Under the regularity in Assumption 1, uniqueness for this continuity equation identifies the interpolation law with the law generated by the velocity's ordinary differential equation (ODE) (Ambrosio et al., 2008, Proposition 8.1.7). Specifically, for each $\mathbf{a} \in \mathbb{R}^d$, define $\Phi_t(\mathbf{a})$ by

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(\mathbf{a}) = \mathbf{v}_t^\star(\Phi_t(\mathbf{a})), \qquad \Phi_0(\mathbf{a}) = \mathbf{a}.$$

Then $(\Phi_1)_{\#}\mu_0 = \mu_1$, so Φ_1 transports the initial distribution to the target distribution (Lipman et al., 2023; Liu et al., 2023). We apply this identity to conditional transition laws in Proposition 1.

3 Problem and Model

The model separates training trajectories from the fresh source trajectory used for generation.

Model 1 (Markov process transport)**.** Fix $d, m, n \in \mathbb{N}$. Let $(\mathbf{x}_i)_{i\ge 0}$ and $(\mathbf{y}_i)_{i\ge 0}$ be independent, time-homogeneous Markov processes on $\mathbb{R}^d$, with transition kernels P_X and P_Y. We observe one source trajectory $\mathbf{x}_{0:m}$ and one target trajectory $\mathbf{y}_{0:m}$. At generation time, we receive a fresh source trajectory $\mathbf{x}'_{0:n}$ with transition kernel P_X, independent of both training trajectories. Its initial distribution is arbitrary.

The finite-sample analysis additionally assumes stationary, mixing training data.

Model 2 (Stationary training trajectories)**.** Instantiate Model 1 with stationary distributions π_X and π_Y, and initialize the training trajectories independently as $\mathbf{x}_0 \sim \pi_X$ and $\mathbf{y}_0 \sim \pi_Y$. Write (β_ℓ^X, τ_X) and (β_ℓ^Y, τ_Y) for their respective mixing profiles and times in (3). Assume $\beta_\ell^X \to 0$ and $\beta_\ell^Y \to 0$ as $\ell \to \infty$. The two processes need not have the same mixing profile or mixing time.

The generation objective specifies the conditional law of each output transition.

Problem 1 (Transport between Markov processes)**.** Under Model 1, given the two training trajectories and a fresh source trajectory $\mathbf{x}'_{0:n}$, learn a single family of maps $(\Phi_{\mathbf{x},\mathbf{y}})$ such that the recursion

$$\widehat{\mathbf{y}}_{i+1} = \Phi_{\mathbf{x}'_i, \widehat{\mathbf{y}}_i}(\mathbf{x}'_{i+1}) \tag{5}$$

has transition kernel P_Y. If P_Y has stationary distribution π_Y and $\widehat{\mathbf{y}}_0 \sim \pi_Y$ independently of $\mathbf{x}'$, the desired output is a stationary target trajectory. We call (5) a *rollout*.

Conditional transport problem. Fix a time index i and condition on the current states $\mathbf{x}_i = \mathbf{x}$ and $\mathbf{y}_i = \mathbf{y}$. Here $\mathbf{x}_i, \mathbf{y}_i$ are random states, whereas $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ are their fixed values. To construct the map $\Phi_{\mathbf{x},\mathbf{y}}$ in (5), write $\mathbf{x}^+ = \mathbf{x}_{i+1}$ and $\mathbf{y}^+ = \mathbf{y}_{i+1}$. Their conditional laws are

$$\mathrm{Law}(\mathbf{x}^+ \mid \mathbf{x}_i = \mathbf{x}, \mathbf{y}_i = \mathbf{y}) = P_X(\mathbf{x}, \cdot), \qquad \mathrm{Law}(\mathbf{y}^+ \mid \mathbf{x}_i = \mathbf{x}, \mathbf{y}_i = \mathbf{y}) = P_Y(\mathbf{y}, \cdot), \tag{6}$$

and they are conditionally independent. Define

$$\mathbf{z}_t = (1-t)\mathbf{x}^+ + t\mathbf{y}^+, \qquad \mathbf{v}_t^\star(\mathbf{z}, \mathbf{x}, \mathbf{y}) = \mathbb{E}\left[\mathbf{y}^+ - \mathbf{x}^+ \mid \mathbf{z}_t = \mathbf{z}, \mathbf{x}_i = \mathbf{x}, \mathbf{y}_i = \mathbf{y}\right]. \tag{7}$$

Write $q_{t|\mathbf{x},\mathbf{y}} = \mathrm{Law}(\mathbf{z}_t \mid \mathbf{x}_i = \mathbf{x}, \mathbf{y}_i = \mathbf{y})$. For any velocity function $\mathbf{v}$, let $\Phi^{\mathbf{v}}_{\mathbf{x},\mathbf{y},t}(\mathbf{a})$ denote the solution at flow time t of

$$\frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}t} = \mathbf{v}_t(\mathbf{z}_t, \mathbf{x}, \mathbf{y}), \qquad \mathbf{z}_0 = \mathbf{a}, \tag{8}$$

and write $\Phi^{\mathbf{v}}_{\mathbf{x},\mathbf{y}} = \Phi^{\mathbf{v}}_{\mathbf{x},\mathbf{y},1}$. The required one-transition identity is

$$(\Phi_{\mathbf{x},\mathbf{y}})_{\#} P_X(\mathbf{x}, \cdot) = P_Y(\mathbf{y}, \cdot). \tag{9}$$

The regression in (7) applies standard flow matching (Lipman et al., 2023; Liu et al., 2023) conditional on $(\mathbf{x}, \mathbf{y})$. The ODE requires the following regularity.

Assumption 1 (Flow regularity)**.** At every current-state pair $(\mathbf{x}, \mathbf{y})$ where a population flow is invoked, the transition distributions have finite second moments. The population velocity is measurable in t and satisfies, for all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$ and almost every $t \in [0, 1]$,

$$\|\mathbf{v}^\star_t(\mathbf{z}, \mathbf{x}, \mathbf{y}) - \mathbf{v}^\star_t(\mathbf{z}', \mathbf{x}, \mathbf{y})\| \le L_t(\mathbf{x}, \mathbf{y}) \|\mathbf{z} - \mathbf{z}'\|, \qquad \|\mathbf{v}^\star_t(\mathbf{z}, \mathbf{x}, \mathbf{y})\| \le G_t(\mathbf{x}, \mathbf{y})(1 + \|\mathbf{z}\|), \tag{10}$$

where $L_\cdot(\mathbf{x}, \mathbf{y})$ and $G_\cdot(\mathbf{x}, \mathbf{y})$ are nonnegative and integrable on $[0, 1]$. Whenever a learned flow is invoked, the learned velocity satisfies the same conditions with its own nonnegative integrable functions L_t and G_t.

These bounds give a unique solution on $[0, 1]$ by Ambrosio et al. (2008, Lemma 8.1.4) and Gronwall's inequality, ensuring the ODE uniqueness required in Liu (2022, Definition 3.1). Bansal et al. (2026, Assumption 1(b)) similarly require the learned velocity to be Lipschitz, with a constant uniform in flow time. Appendix B verifies the bounds for Gaussian transitions. Algorithm 1 trains the velocity and generates the target trajectory. Exact stationary generation uses $\widehat{\mu}_0 = \pi_Y$; the finite-sample bounds retain the error of approximate initialization.

4 Main Results

Transport requires learning target transitions and preserving them under recursive application. We first establish these properties for the population velocity. This reduces the statistical problem to estimating that velocity from dependent observations, for which we give an upper bound and a mixing-dependent lower bound. We then bound generation error under suitable stability assumptions.

4.1 From one transition to a trajectory

For fixed current states $(\mathbf{x}, \mathbf{y})$, applying the flow-matching identity (4) to $P_X(\mathbf{x}, \cdot)$ and $P_Y(\mathbf{y}, \cdot)$ shows that the population ODE follows the conditional interpolation law and reaches the target transition.

Proposition 1 (Conditional transport)**.** *Under Model 1, fix current-state values $(\mathbf{x}, \mathbf{y})$ at which Assumption 1 holds. For the ODE map $\Phi^{\mathbf{v}^\star}_{\mathbf{x},\mathbf{y},t}$ in (8) and every $t \in [0, 1]$,*

$$(\Phi^{\mathbf{v}^\star}_{\mathbf{x},\mathbf{y},t})_{\#} P_X(\mathbf{x}, \cdot) = q_{t|\mathbf{x},\mathbf{y}}. \tag{11}$$

In particular, $(\Phi^{\mathbf{v}^\star}_{\mathbf{x},\mathbf{y}})_{\#} P_X(\mathbf{x}, \cdot) = P_Y(\mathbf{y}, \cdot)$.

Algorithm 1 Conditional flow matching for transport between Markov processes

Require: Training trajectories $\mathbf{x}_{0:m}, \mathbf{y}_{0:m}$; candidate velocity class $\mathcal{V}$; fresh source trajectory $\mathbf{x}'_{0:n}$; initial distribution $\widehat{\mu}_0$.

Training

for $i = 0, \dots, m-1$ **do**
  Set $\mathbf{r}_i = \mathbf{y}_{i+1} - \mathbf{x}_{i+1}$ and $\mathbf{z}_i(t) = (1-t)\mathbf{x}_{i+1} + t\mathbf{y}_{i+1}$ for $t \in [0, 1]$.
end for
Minimize $\frac{1}{m}\sum_{i=0}^{m-1} \int_0^1 \|\mathbf{v}_t(\mathbf{z}_i(t), \mathbf{x}_i, \mathbf{y}_i) - \mathbf{r}_i\|^2 \, \mathrm{d}t$ over $\mathbf{v} \in \mathcal{V}$ to obtain $\widehat{\mathbf{v}}$.

Generation

Draw $\widehat{\mathbf{y}}_0 \sim \widehat{\mu}_0$ independently of $\mathbf{x}'$.
for $i = 0, \dots, n-1$ **do**
  Solve $\dot{\mathbf{z}}_t = \widehat{\mathbf{v}}_t(\mathbf{z}_t, \mathbf{x}'_i, \widehat{\mathbf{y}}_i)$ on $[0, 1]$ from $\mathbf{z}_0 = \mathbf{x}'_{i+1}$.
  Set $\widehat{\mathbf{y}}_{i+1} = \mathbf{z}_1$.
end for
return $\widehat{\mathbf{y}}_{0:n}$.

During recursive generation, the current target state depends on past source states. We therefore need the transport identity to remain valid after conditioning on this shared history. The source Markov property ensures that the next source state still has law $P_X(\mathbf{x}'_i, \cdot)$, to which we can apply the transport map.

Lemma 1 (Conditional transitions during generation)**.** *Under Model 1, let $\widehat{\mathbf{y}}_0 \sim \widehat{\mu}_0$ be independent of $\mathbf{x}'$, and suppose each $\widehat{\mathbf{y}}_{i+1}$ is a fixed measurable function of $(\mathbf{x}'_{0:i+1}, \widehat{\mathbf{y}}_{0:i})$. Then, for every $i \geq 0$, $\mathrm{Law}(\mathbf{x}'_{i+1} \mid \mathbf{x}'_{0:i}, \widehat{\mathbf{y}}_{0:i}) = P_X(\mathbf{x}'_i, \cdot)$ almost surely. If the updates follow (5) with maps satisfying (9) at every visited state pair, then, for every $i \geq 0$ and measurable $B \subseteq \mathbb{R}^d$, almost surely,*

$$\mathbb{P}\left(\widehat{\mathbf{y}}_{i+1} \in B \mid \mathbf{x}'_{0:i}, \widehat{\mathbf{y}}_{0:i}\right) = P_Y(\widehat{\mathbf{y}}_i, B). \tag{12}$$

The generated process thus has the target transition kernel conditional on its full history. Its initial distribution determines its trajectory law; choosing π_Y gives the stationary target law. Appendix A.1 gives the conditioning argument and proves the following consequence.

Proposition 2 (Exact trajectory generation)**.** *Under Model 1, suppose P_Y has stationary distribution π_Y, draw $\widehat{\mathbf{y}}_0 \sim \pi_Y$ independently of $\mathbf{x}'$, and generate by (5). If (9) holds at every visited current-state pair, then $\widehat{\mathbf{y}}_{0:n}$ has the stationary target trajectory law for every $n \geq 0$.*

4.2 Velocity estimation from two dependent trajectories

Algorithm 1 estimates the population velocity by regressing next-state displacements on interpolated states and the current-state pair. The statistical difficulty is dependence between successive regression pairs from the two trajectories. For each observed transition and flow time $t \in [0, 1]$, set

$$\mathbf{z}_i(t) = (1-t)\mathbf{x}_{i+1} + t\mathbf{y}_{i+1}, \quad \mathbf{r}_i = \mathbf{y}_{i+1} - \mathbf{x}_{i+1}, \quad \mathbf{h}_i(t) = (t, \mathbf{z}_i(t), \mathbf{x}_i, \mathbf{y}_i). \tag{13}$$

Regression noise. Let $(\mathbf{h}, \mathbf{r})$ be obtained from a fresh stationary transition pair and an independent flow time $t \sim \mathcal{U}(0, 1)$ using (13). The population velocity in (7) satisfies $\mathbf{v}^\star(\mathbf{h}) = \mathbb{E}[\mathbf{r} \mid \mathbf{h}]$. Define the regression noise by $\boldsymbol{\xi} = \mathbf{v}^\star(\mathbf{h}) - \mathbf{r}$ and $\boldsymbol{\xi}_i(t) = \mathbf{v}^\star(\mathbf{h}_i(t)) - \mathbf{r}_i$. The following assumption controls this regression noise.

Assumption 2 (Conditional sub-Gaussian noise)**.** There exists $\nu \geq 0$ such that, almost surely, for every $\mathbf{a} \in \mathbb{R}^d$ and $\theta \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\theta\langle \mathbf{a}, \boldsymbol{\xi}\rangle} \mid \mathbf{h}\right] \leq e^{\theta^2\nu^2\|\mathbf{a}\|^2/2}. \tag{14}$$

By stationarity, the same bound holds for $(\mathbf{h}_i(t), \mathbf{r}_i)$ when t is drawn independently from $\mathcal{U}(0, 1)$, for every i.

Remark. Assumption 2 holds with $\nu = B$ whenever $\|\mathbf{r}\| \leq B$ almost surely. This includes source and target processes on bounded state spaces. It also holds for transitions with uniformly bounded additive noise and for Gaussian transition kernels with uniformly bounded covariance; both allow nonlinear transition means. Appendix C verifies these cases.

We fit this conditional mean over a nonempty finite class $\mathcal{V}$ of measurable functions $\mathbf{v} : [0, 1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$. The finite-class assumption reflects finite-precision storage of neural-network parameters. For a fixed architecture with p parameters, each taking at most q values, we have $|\mathcal{V}| \leq q^p$ and $\log|\mathcal{V}| \leq p \log q$. Kumar et al. (2025, Section 3) use the same interpretation. We average the regression loss over flow time:

$$\widehat{L}(\mathbf{v}) = \frac{1}{m}\sum_{i=0}^{m-1}\int_0^1 \|\mathbf{v}(\mathbf{h}_i(t)) - \mathbf{r}_i\|^2 \, \mathrm{d}t, \qquad \mathbf{v} \in \mathcal{V}. \tag{15}$$

Uniformly sampling transition indices and fresh independent flow times gives unbiased minibatch estimates, as in Kumar et al. (2025, Remark 2). Appendix D specifies this procedure for our experiments. The statistical analysis concerns empirical minimization; finite-budget optimization is a separate source of error. We minimize this loss and measure squared error relative to the population velocity:

$$\widehat{\mathbf{v}} \in \arg\min_{\mathbf{v}\in\mathcal{V}} \widehat{L}(\mathbf{v}), \qquad R(\mathbf{v}, \mathbf{v}^\star) = \mathbb{E}\,\|\mathbf{v}(\mathbf{h}) - \mathbf{v}^\star(\mathbf{h})\|^2, \tag{16}$$

where $\mathbf{h}$ is a fresh draw from the stationary training distribution. To relate empirical fit to population error, we assume realizability and a fourth-moment bound on prediction errors.

Assumption 3 (Realizability and fourth-moment bound)**.** The class is realizable, meaning $\mathbf{v}^\star \in \mathcal{V}$. There is a constant $C_{4\to 2} \geq 1$ such that, for every $\mathbf{v} \in \mathcal{V}$, $R(\mathbf{v}, \mathbf{v}^\star) < \infty$ and

$$\left(\mathbb{E}\,\|\mathbf{v}(\mathbf{h}) - \mathbf{v}^\star(\mathbf{h})\|^4\right)^{1/4} \leq C_{4\to 2}\sqrt{R(\mathbf{v}, \mathbf{v}^\star)}. \tag{17}$$

The moment comparison parallels Kumar et al. (2025, Assumption 2) and the assumption in Wainwright (2019, Section 14.2), here under the joint training distribution. Appendix B verifies realizability and $C_{4\to 2}^4 \leq 8$ for the finite Gaussian velocity class.

To account for training dependence, note that the function $t \mapsto (\mathbf{h}_i(t), \mathbf{r}_i)$ is determined by the transition observation $\mathbf{o}_i = (\mathbf{x}_i, \mathbf{x}_{i+1}, \mathbf{y}_i, \mathbf{y}_{i+1})$. Consecutive observations $\mathbf{o}_i$ and $\mathbf{o}_{i+1}$ contain the same states $\mathbf{x}_{i+1}$ and $\mathbf{y}_{i+1}$. Under Model 2, this sequence is stationary and Markov by Lemma 2. With stationary law π_O, its observation mixing profile and mixing time are

$$\beta_\ell = \mathbb{E}_{\mathbf{o}_0 \sim \pi_O} \left\|\mathrm{Law}(\mathbf{o}_\ell \mid \mathbf{o}_0) - \pi_O\right\|_{\mathrm{TV}}, \qquad \tau_{\mathrm{mix}}(\alpha) = \min\{\ell \geq 1 : \beta_\ell \leq \alpha\}, \tag{18}$$

for integers $\ell \geq 1$ and $\alpha \in (0, 1)$, specializing (3) to $(\mathbf{o}_i)$. Given $\mathbf{o}_0$, the states $(\mathbf{x}_\ell, \mathbf{y}_\ell)$ evolve from $(\mathbf{x}_1, \mathbf{y}_1)$ for $\ell - 1$ steps. Independence bounds their joint total-variation distance by the sum of the two chains' distances; drawing the next states cannot increase that distance. Lemma 2 in Appendix A.2 therefore gives, for every $\ell \geq 2$ and $\alpha \in (0, 1)$,

$$\beta_\ell \leq \beta_{\ell-1}^X + \beta_{\ell-1}^Y, \qquad \tau_{\mathrm{mix}}(\alpha) \leq 1 + \max\{\tau_X(\alpha/2), \tau_Y(\alpha/2)\}. \tag{19}$$

The slower process controls this upper bound; the mixing times may differ. We combine the noise, moment, and mixing assumptions. The resulting risk bound scales as K/m, up to class complexity and logarithmic factors, where K is the observation mixing time at tolerance $\delta/(3m)$.

Theorem 1 (Velocity estimation from dependent trajectories)**.** *Under Model 2, assume Assumptions 3 and 2. Fix $\delta \in (0, 1)$ and set $K = \tau_{\mathrm{mix}}(\delta/(3m))$ for the transition-observation sequence $(\mathbf{o}_i)$, as defined in (18). Suppose*

$$m \geq 128 C_{4\to 2}^4 K \log\left(\frac{6K|\mathcal{V}|}{\delta}\right). \tag{20}$$

Then the minimizer $\widehat{\mathbf{v}}$ of (15) satisfies, with probability at least $1 - \delta$ over the two training trajectories,

$$R(\widehat{\mathbf{v}}, \mathbf{v}^\star) \leq 128 C_{4\to 2}^4 \nu^2 \frac{K \log(6K|\mathcal{V}|/\delta)}{m}. \tag{21}$$

The same guarantee holds when each time integral is replaced by the average of any fixed number $J \geq 1$ of independent uniform flow-time samples per transition, with probability also over these samples (Corollary 1). The sample size m counts observed transitions in both cases.

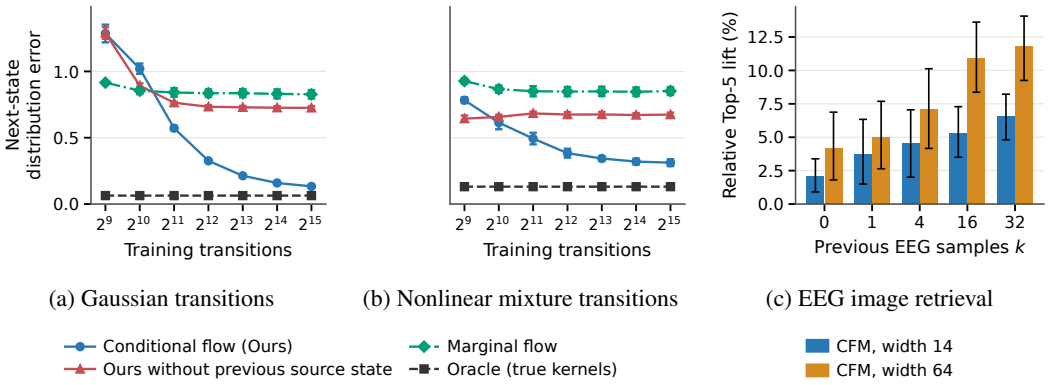
To see how these ingredients enter, put $\mathbf{g}_i(\mathbf{v}, t) = \mathbf{v}(\mathbf{h}_i(t)) - \mathbf{v}^\star(\mathbf{h}_i(t))$ and decompose the excess loss:

$$\widehat{L}(\mathbf{v}) - \widehat{L}(\mathbf{v}^\star) = \frac{1}{m}\sum_i \int_0^1 \left(\|\mathbf{g}_i(\mathbf{v}, t)\|^2 + 2\,\langle \mathbf{g}_i(\mathbf{v}, t), \boldsymbol{\xi}_i(t)\rangle\right) \mathrm{d}t. \tag{22}$$

At the empirical minimizer, the left side is nonpositive by realizability. Conditional sub-Gaussianity and the fourth-moment assumption bound the exponential moment of the negative excess loss at a random flow time. Jensen's inequality preserves this bound when we average over flow time. To handle dependence, we split the indices into subsequences $a, a + K, a + 2K, \ldots$ for $a \in \{0, \ldots, K - 1\}$. We couple each subsequence to independent observations and bound the probability of any mismatch. Hölder's inequality combines the exponential bounds for these subsequences without assuming independence between them. In Appendix A.2, we give the full proof and derive the sample complexity in (2) by combining (20) and (21).

To establish the necessity of mixing-dependent sample complexity, we construct two models that no estimator can reliably distinguish from a short trajectory. We do so by adapting the unvisited-state argument of Nagaraj et al. (2020, Appendix C.1): a slowly mixing chain may avoid the region where two candidate velocities differ, leaving their models indistinguishable on the observed trajectories.

Proposition 3 (Mixing-time lower bound)**.** *For every integer $\tau \geq 1$ and $\epsilon > 0$, there exist two instances of Model 2 which satisfy Assumptions 2, 1, and 3 with universal constants $\nu, C_{4\to 2}$, share a class $\mathcal{V}$ of two velocities, and both have observation mixing time $\Theta(\tau)$, such that for any*



(a) Gaussian transitions (b) Nonlinear mixture transitions (c) EEG image retrieval

Figure 2: Synthetic transition recovery and EEG image retrieval. (a–b) Next-state distribution error for $d = 10$ (mean $\pm$ standard error over three seeds; lower is better). The ablation omits $\mathbf{x}_i$; marginal flow omits both context states; the oracle uses true kernels. (c) Relative Top-5 retrieval accuracy lift over ENIGMA with only $\mathbf{W}_s$ adapted, versus EEG context length k; both flow networks precede $\mathbf{W}_s$ and have two hidden layers: width 64 and width 14. Accuracies average three seeds per subject, then ten subjects; error bars show 95% subject-bootstrap percentile intervals.

$\delta \in (0, 1/(2\pi))$ and any estimator $\widetilde{\mathbf{v}} \in \mathcal{V}$ which achieves $R(\widetilde{\mathbf{v}}, \mathbf{v}^\star) < \epsilon$ with probability at least $1 - \delta$ under both instances using m observed transitions from each training trajectory,

$$m \geq \frac{1}{3}\tau_{\mathrm{mix}} \log \frac{1}{2\pi\delta}. \tag{23}$$

Here τ_{mix} denotes either instance's observation mixing time at tolerance $1/4$.

Remark (Comparing the upper and lower bounds). The lower bound uses $\tau_{\mathrm{mix}} = \tau_{\mathrm{mix}}(1/4)$, whereas the upper bound uses $K = \tau_{\mathrm{mix}}(\delta/(3m))$. We compare these rates under geometric mixing, i.e, $\beta_\ell \leq c_0 e^{-c_1 \ell/\tau_{\mathrm{mix}}}$ for all $\ell \geq 1$ and universal $c_0 \geq 1, c_1 > 0$. Then $K = O(\tau_{\mathrm{mix}} \log(3m/\delta))$, so the bounds agree in mixing-time dependence up to logarithmic factors. Appendix A.4 derives this comparison, verifies the exponential profile for our Gaussian construction, and gives a variant under finite worst-case mixing time using total-variation contraction (Levin & Peres, 2017).

5 Experiments

We test Gaussian and nonlinear transition recovery and EEG image retrieval. Lemma 1 identifies the source law as $P_X(\mathbf{x}_i, \cdot)$, so conditioning also on $\mathbf{x}_i$ gives the exact trajectory guarantee of Proposition 2. We compare our flow, an ablation omitting $\mathbf{x}_i$, an unconditional marginal flow, and an oracle. Marginal flow learns a single map T between state distributions and outputs $T(\mathbf{x}_{i+1})$ without either context input. Appendices D and E specify the protocols. Figure 2(a–b) varies training size at $d = 10$. We measure conditional next-state error by sliced W_2 (Rowland et al., 2019), defined in Appendix D. Both oracles transport the supplied source next state using the true transition laws. Their positive error reflects finite-sample evaluation.

Gaussian transitions. We simulate independent processes in $\mathbb{R}^{10}$ with Gaussian transitions $P_X(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{A}_X\mathbf{x}, \mathbf{Q}_X)$ and $P_Y(\mathbf{y}, \cdot) = \mathcal{N}(\mathbf{A}_Y\mathbf{y}, \mathbf{Q}_Y)$, where $\mathbf{A}_X, \mathbf{A}_Y$ are transition matrices and $\mathbf{Q}_X, \mathbf{Q}_Y$ are positive-definite noise covariances. Source and target transition eigenvalue magnitudes lie in $[0.72, 0.94]$ and $[0.15, 0.58]$, respectively; all four matrices have independent random orthogonal eigenbases. Each process supplies $m \in \{2^9, 2^{10}, \ldots, 2^{15}\}$ training transitions. Figure 2(a) shows that our error falls from 1.286 at $m = 512$ to 0.133 at $m = 32768$, versus 0.725 for the ablation and 0.827 for marginal flow. Conditioning improves recovery with more data.

Nonlinear mixture transitions. For $d = 10$, let $\mathbf{R}_X, \mathbf{R}_Y$ be independent random orthogonal matrices. Define $\mathbf{f}(\mathbf{x}) = 0.85\mathbf{R}_X \tanh(\mathbf{R}_X^\top \mathbf{x})$, $\mathbf{g}(\mathbf{y}) = \mathbf{R}_Y[0.55 \tanh(\mathbf{R}_Y^\top \mathbf{y}) + 0.2 \sin(2\mathbf{R}_Y^\top \mathbf{y})]$, and $\mathbf{u} = d^{-1/2}\mathbf{R}_Y(1, \ldots, 1)^\top$, where $\tanh$ and $\sin$ act coordinatewise. Each map applies scalar

nonlinearities in rotated coordinates, then transforms back, coupling the observed coordinates. We use $P_X(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{f}(\mathbf{x}), 0.6^2\mathbf{I})$ and $P_Y(\mathbf{y}, \cdot) = \frac{1}{2}\mathcal{N}(\mathbf{g}(\mathbf{y}) - 0.8\sqrt{d}\,\mathbf{u}, 0.25^2\mathbf{I}) + \frac{1}{2}\mathcal{N}(\mathbf{g}(\mathbf{y}) + 0.8\sqrt{d}\,\mathbf{u}, 0.25^2\mathbf{I})$. Each target noise vector uses one mixture sign, giving two modes along $\mathbf{u}$. We use the same training-size sweep as in panel (a). In Figure 2(b), conditioning improves with more data: at $m = 32768$, our error is 0.312 versus 0.675 for the ablation, 0.852 for marginal flow, and 0.131 for the oracle. The ablation performs better at the smallest sample sizes, while our gains taper at larger sizes under the fixed training budget. Appendix D.3 reports the corresponding $d = 100$ results.

Supervised EEG adaptation. In THINGS-EEG2 (Gifford et al., 2022), participants view images while electroencephalography (EEG) records electrical activity from 63 scalp channels. Each recording spans one second after an image appears. The task is to learn a decoder which predicts the viewed image among the 200 candidate images. Responses to the same image vary across people, motivating adaptation of a decoder for new subjects. We test whether a flow conditioned on preceding EEG measurements improves adaptation by transforming the recording before decoding.

For subject s viewing image u, let $\mathbf{A} = \mathbf{E}_s(u) = (\mathbf{a}_1, \ldots, \mathbf{a}_{250}) \in \mathbb{R}^{63\times 250}$. Our ENIGMA baseline (Kneeland et al., 2026; Alljoined, 2025) predicts $D(\mathbf{W}_s(\mathbf{A}))$, adapting the subject-specific temporal map $\mathbf{W}_s$ while freezing the encoder and projection network D pretrained on other subjects. We predict $D(\mathbf{W}_s(F_{\boldsymbol{\theta}}(\mathbf{A})))$, jointly training the flow $F_{\boldsymbol{\theta}}$ and $\mathbf{W}_s$. Both methods share 4,096 calibration and 1,000 validation images per subject; evaluation averages 80 EEG repetitions per image.

Training pairs $\mathbf{A}$ with each reference subject's recording $\mathbf{B}_q = \mathbf{E}_q(u)$ of the same image, with columns $\mathbf{b}_{qj}$. For $0 \le k \le 32$, the history $\mathbf{h}_k(\mathbf{A}, j) = (\mathbf{a}_{j-1}, \ldots, \mathbf{a}_{j-k}, \mathbf{0}, \ldots, \mathbf{0}) \in \mathbb{R}^{63\cdot 32}$ contains the previous k samples, with missing samples and unused slots set to zero; define $\mathbf{h}_k(\mathbf{B}_q, j)$ similarly. This keeps network size fixed as k varies. With $t \sim \mathcal{U}(0,1)$ and $\mathbf{z}_{qj} = (1-t)\mathbf{a}_j + t\mathbf{b}_{qj}$, we combine ENIGMA's supervised objective with the mean flow-matching loss

$$\left\|\mathbf{v}_{\boldsymbol{\theta}}\left(t, \mathbf{z}_{qj}; \mathbf{h}_k(\mathbf{A}, j), \mathbf{h}_k(\mathbf{B}_q, j)\right) - (\mathbf{b}_{qj} - \mathbf{a}_j)\right\|^2 .$$

At inference, we generate $\widehat{\mathbf{B}} = F_{\boldsymbol{\theta}}(\mathbf{A})$ sequentially by integrating from each $\mathbf{a}_j$, using previously generated target history $\mathbf{h}_k(\widehat{\mathbf{B}}, j)$ and holding both histories fixed during each solve. Only the source recording is required. Remaining details are in Appendix E.

Figure 2c compares widths 64 and 14 at $k \in \{0, 1, 4, 16, 32\}$. Top-5 accuracy is the frequency of ranking the viewed image among the top five candidates. At $k = 32$, our width-64 model achieves 49.05% accuracy versus 43.87% for the baseline, an 11.82% relative lift (95% subject-bootstrap interval $[9.25, 14.06]\%$), up from 4.18% lift at $k = 0$. The width-14 model similarly reaches 6.53% lift at $k = 32$. Both widths benefit from temporal context and improve on the baseline for all ten subjects after averaging seeds.

6 Conclusion

We prove that conditional flow matching transforms a fresh source trajectory into the exact target trajectory distribution at the population level. For a finite candidate class, we prove a sample-complexity upper bound proportional to mixing time up to logarithmic factors, a lower bound showing that this dependence is unavoidable, and a finite-sample rollout bound. Our synthetic experiments show the value of conditioning on the previous source state, and our EEG study shows improved adaptation between subjects.

Extending the analysis beyond finite classes, identifying when accuracy on observed transitions carries over to generated trajectories, and handling longer memory remain open directions. Generation can be slow because it requires sequential ODE solves along the trajectory. Reducing this latency is another interesting direction, potentially through parallel Picard iteration (Shih et al., 2023) or training flows to generate several future states jointly.

ETHICS STATEMENT

Our EEG experiments reuse the publicly released THINGS-EEG2 dataset (Gifford et al., 2022); we collect no new human-subject data. The dataset is distributed under a CC BY 4.0 license. EEG-based image retrieval raises privacy concerns because it seeks to infer viewed stimuli from neural recordings. Our evaluation is limited to ranking a fixed set of candidate images in a controlled experimental setting. Any use beyond this setting requires separate validation and safeguards for participant consent and neural-data privacy.

REPRODUCIBILITY STATEMENT

Appendix A contains the proofs, and Appendices D and E describe the experimental procedures.

AI USE STATEMENT

In this work, we used generative AI tools to explore ideas for the rollout proof and the lower bound, assist with writing and revising proofs, and support experiment implementation. Additionally, we used these tools for manuscript editing, literature searches and citation checking, code development, and figure preparation. The authors reviewed the AI-assisted work by checking the proofs, testing the code, and verifying the references. The authors take responsibility for the final manuscript, mathematical claims, code, figures, and experimental results, including content produced with the assistance of generative AI.

REFERENCES

Alljoined. ENIGMA released implementation. GitHub repository, commit 460af4b6a0ca, 2025. URL https://github.com/Alljoined/ENIGMA/tree/460af4b6a0ca6eee11aeaf35c5431e8e3d55c11a.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, second edition, 2008. doi: 10.1007/978-3-7643-8722-8.

Frank Aurzada and Christoph Baumgarten. Survival probabilities of weighted random walks. *ALEA. Latin American Journal of Probability and Mathematical Statistics*, 8:235–258, 2011. URL https://alea.impa.br/articles/v8/08-13.pdf.

Julio Backhoff, Mathias Beiglböck, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017. doi: 10.1137/16M1080197.

Vansh Bansal, Saptarshi Roy, Alessandro Rinaldo, and Purnamrita Sarkar. On the convergence and straightness of rectified flow. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300, pp. 4636–4644, 2026.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024.

H. C. P. Berbee. *Random Walks with Stationary Increments and Renewal Theory*, volume 112 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1979. URL https://ir.cwi.nl/pub/12694.

Richard C. Bradley. Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

Clément L. Canonne. A short note on an inequality between KL and TV, 2022. URL https://arxiv.org/abs/2202.07198.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional wasserstein distances with applications in bayesian OT flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025.

Zhenyao Cui, Siyuan Kan, Siyang Li, Ziwei Wang, and Dongrui Wu. SCORE: Subject coordinate recovery for label-free cross-subject EEG-to-image retrieval. *arXiv preprint arXiv:2608.19134*, 2026.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2701–2709, 2020. URL https://proceedings.mlr.press/v119/duan20b.html.

Yuan Gao, Jian Huang, and Yuling Jiao. Gaussian interpolation flows. *Journal of Machine Learning Research*, 25(253):1–52, 2024.

Mudit Gaur, Prashant Trivedi, Shuchin Aeron, Amrit Singh Bedi, George K. Atia, and Vaneet Aggarwal. Generative modeling with continuous flows: Sample complexity of flow matching. *arXiv preprint arXiv:2512.01286*, 2025. URL https://arxiv.org/abs/2512.01286v1. Version 1, 1 December 2025.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. doi: 10.1016/j.neuroimage.2022.119754.

Mohamed Hamdouche, Pierre Henry-Labordère, and Huyên Pham. Nonparametric generative modeling for time series via Schrödinger bridge. *Journal of Machine Learning Research*, 27(58):1–23, 2026.

Qunjie Huang and Weina Zhu. SATTC: Structure-aware label-free test-time calibration for cross-subject EEG-to-image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Dynamic conditional optimal transport through simulation-free flows. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-2968.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Functional flow matching. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 3934–3942, 2024b.

DongWook Kim, Junghan Kwon, Byungjun Jeon, and Yong-Lae Park. Adaptive calibration of soft sensors using optimal transportation transfer learning for mass production and long-term usage. *Advanced Intelligent Systems*, 2(6):1900178, 2020. doi: 10.1002/aisy.201900178.

Reese Kneeland, Wangshu Jiang, Ugo Bruzadin Nunes, Paul Steven Scotti, Arnaud Delorme, and Jonathan Xu. ENIGMA: EEG-to-image in 15 minutes using less than 1% of the parameters. *arXiv preprint arXiv:2602.10361*, 2026. URL https://arxiv.org/abs/2602.10361v1. Version 1, 10 February 2026.

Syamantak Kumar and Purnamrita Sarkar. Streaming PCA for markovian data. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Syamantak Kumar, Dheeraj Nagaraj, and Purnamrita Sarkar. Dimension-free score matching and time bootstrapping for diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1082.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, second edition, 2017. doi: 10.1090/mbk/107. With contributions by Elizabeth L. Wilmer.

Soon Hoe Lim, Shizheng Lin, Michael W. Mahoney, and N. Benjamin Erichson. Is flow matching just trajectory replay for sequential data? *arXiv preprint arXiv:2602.08318*, 2026. URL https://arxiv.org/abs/2602.08318v2. Version 2, 7 May 2026.

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

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Khushboo Mishra, Varun Trivedi, and Tanima Dutta. Towards uncertainty-aware unsupervised domain adaptation for videos and time-series with causal optimal transport. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29421–29430, 2026.

Vrettos Moulos. Bicausal optimal transport for Markov chains via dynamic programming. In *IEEE International Symposium on Information Theory*, pp. 1688–1693, 2021. doi: 10.1109/ISIT45174.2021.9517977.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c22abfa379f38b5b0411bc11fa9bf92f-Abstract.html.

Joe Neeman, Bobby Shi, and Rachel Ward. Concentration inequalities for sums of markov-dependent random matrices. *Information and Inference: A Journal of the IMA*, 13(4):iaae032, 2024. doi: 10.1093/imaiai/iaae032.

Kevin O'Connor, Kevin McGoff, and Andrew B. Nobel. Optimal transport for stationary markov chains via policy iteration. *Journal of Machine Learning Research*, 23(45):1–52, 2022.

Kevin O'Connor, Andrew Nobel, and Kevin McGoff. Estimation of stationary optimal transport plans. *Information and Inference: A Journal of the IMA*, 13(2):iaae006, 2024. doi: 10.1093/imaiai/iaae006.

François Painblanc, Laetitia Chapel, Nicolas Courty, Chloé Friguet, Charlotte Pelletier, and Romain Tavenard. Match-and-deform: Time series domain adaptation through optimal transport and temporal alignment. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 341–356, 2023. doi: 10.1007/978-3-031-43424-2_21.

Yoann Robin, Mathieu Vrac, Philippe Naveau, and Pascal Yiou. Multivariate stochastic bias corrections with optimal transport. *Hydrology and Earth System Sciences*, 23:773–786, 2019. doi: 10.5194/hess-23-773-2019.

Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. Orthogonal estimation of Wasserstein distances. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 186–195, 2019. URL https://proceedings.mlr.press/v89/rowland19a.html.

Abhishek Roy, Krishnakumar Balasubramanian, and Murat A. Erdogdu. On empirical risk minimization with dependent and heavy-tailed data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 2018. URL https://arxiv.org/abs/1503.04123.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059, 2024. URL https://proceedings.mlr.press/v235/scotti24a.html.

Andy Shih, Suneel Belkhale, Stefano Ermon, Dorsa Sadigh, and Nima Anari. Parallel sampling of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0d1986a61e30e5fa408c81216a616e20-Abstract.html.

Marta Gentiloni Silveri, Giovanni Conforti, and Alain Durmus. Theoretical guarantees in KL for diffusion flow matching. In *Advances in Neural Information Processing Systems*, volume 37, pp. 138432–138473, 2024. doi: 10.52202/079017-4393.

Alexander Y. Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free Schrödinger bridges via score and flow matching. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 1279–1287, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019. URL https://www.cambridge.org/core/books/highdimensional-statistics/8A91ECEEC38F46DAB53E9FF8757C7A4E.

Ganchao Wei and Li Ma. Stream-level flow matching with Gaussian processes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 66137–66154, 2025.

Tianlin Xu, Li Kevin Wenliang, Michael Munn, and Beatrice Acciaio. COT-GAN: Generating sequential data via causal optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

Andrea Zanette and Martin J. Wainwright. Bellman residual orthogonalization for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0227.

Zhengyu Zhou and Weiwei Liu. An error analysis of flow matching for deep generative modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78903–78932, 2025.

Ingvar Ziemann, Stephen Tu, George J. Pappas, and Nikolai Matni. Sharp rates in dependent learning theory: Avoiding sample size deflation for the square loss. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62779–62802. PMLR, 2024. URL https://proceedings.mlr.press/v235/ziemann24a.html.

APPENDIX CONTENTS

A Proofs . . . 16
A.1 Conditional transport and trajectory generation . . . 16
A.2 Dependent velocity estimation . . . 16
A.3 The sample-size lower bound . . . 20
A.4 Comparison of mixing tolerances . . . 22
A.5 Accuracy of generated states and trajectories . . . 22
B Gaussian transitions, estimation, and rollout . . . 25
C Examples for the conditional noise assumption . . . 27
D Experimental details . . . 28
D.1 Gaussian transition experiments . . . 29
D.2 Nonlinear mixture benchmark . . . 29
D.3 Results at dimension 100 . . . 29
E Ten-subject EEG study . . . 30
E.1 Protocol . . . 30
E.2 Adapter configurations . . . 30
E.3 Additional results . . . 31
E.4 Flow after the temporal map . . . 31
F Further comparisons with prior work . . . 32
F.1 Sequential transport and generation . . . 32
F.2 Statistical error and dependence . . . 33
F.3 Distribution mismatch and rollout stability . . . 33

A PROOFS

A.1 CONDITIONAL TRANSPORT AND TRAJECTORY GENERATION

Proposition 1 identifies the law of a single population flow. Lemma 1 verifies that its source input retains the required conditional law during generation; together they give Proposition 2.

Proof of Proposition 1. We show that the interpolation law solves the continuity equation, then use uniqueness to identify it with the ODE flow law. Condition on $\mathbf{x}_i = \mathbf{x}$ and $\mathbf{y}_i = \mathbf{y}$. Throughout this proof, all laws and expectations are under this conditioning, which we suppress from the notation. Let $\mathbf{z}_t = (1-t)\mathbf{x}^+ + t\mathbf{y}^+$, $q_t = \mathrm{Law}(\mathbf{z}_t)$, and $\mathbf{r} = \mathbf{y}^+ - \mathbf{x}^+$. For every smooth compactly supported $\varphi : \mathbb{R}^d \to \mathbb{R}$, finite first moments and conditioning on $\mathbf{z}_t$ give

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(\mathbf{z}) q_t(\,\mathrm{d}\mathbf{z}) &= \mathbb{E}\,\langle \nabla\varphi(\mathbf{z}_t), \mathbf{r}\rangle = \mathbb{E}\,\langle \nabla\varphi(\mathbf{z}_t), \mathbb{E}[\mathbf{r} \mid \mathbf{z}_t]\rangle \\ &= \int \langle \nabla\varphi(\mathbf{z}), \mathbf{v}_t^\star(\mathbf{z})\rangle\, q_t(\,\mathrm{d}\mathbf{z}).\end{aligned}$$

Thus q_t solves the continuity equation with velocity $\mathbf{v}_t^\star$ and initial law $P_X(\mathbf{x}, \cdot)$. To identify this solution, observe that the ODE flow in (8) has the same velocity and initial law. Uniqueness under Assumption 1 (Ambrosio et al., 2008, Proposition 8.1.7) therefore gives $\mathrm{Law}(\Phi^{\mathbf{v}^\star}_{\mathbf{x},\mathbf{y},t}(\mathbf{x}^+)) = q_{t|\mathbf{x},\mathbf{y}}$. At $t = 1$, this is $P_Y(\mathbf{y}, \cdot)$. □

Proof of Lemma 1. We first identify the source law conditional on the generated history, then apply the transport identity to obtain the target transition. Independence of $\widehat{\mathbf{y}}_0$ and $\mathbf{x}'$, followed by the source Markov property, gives

$$\mathrm{Law}\left(\mathbf{x}'_{i+1} \mid \mathbf{x}'_{0:i}, \widehat{\mathbf{y}}_0\right) = \mathrm{Law}\left(\mathbf{x}'_{i+1} \mid \mathbf{x}'_{0:i}\right) = P_X(\mathbf{x}'_i, \cdot).$$

By recursion, $\widehat{\mathbf{y}}_{0:i}$ is a deterministic function of $(\mathbf{x}'_{0:i}, \widehat{\mathbf{y}}_0)$. Since the generated history includes $\widehat{\mathbf{y}}_0$, these two histories generate the same sigma-field:

$$\sigma(\mathbf{x}'_{0:i}, \widehat{\mathbf{y}}_{0:i}) = \sigma(\mathbf{x}'_{0:i}, \widehat{\mathbf{y}}_0).$$

This proves the first assertion. For updates of the form (5), the transport identity (9) at $(\mathbf{x}'_i, \widehat{\mathbf{y}}_i)$ now gives, for every measurable $B \subseteq \mathbb{R}^d$,

$$\begin{aligned}\mathbb{P}(\widehat{\mathbf{y}}_{i+1} \in B \mid \mathbf{x}'_{0:i}, \widehat{\mathbf{y}}_{0:i}) &= \int \mathbf{1}\{\Phi_{\mathbf{x}'_i, \widehat{\mathbf{y}}_i}(\mathbf{u}) \in B\} P_X(\mathbf{x}'_i, \,\mathrm{d}\mathbf{u}) \\ &= P_Y(\widehat{\mathbf{y}}_i, B).\end{aligned}$$

□

Proof of Proposition 2. By Lemma 1 and the tower property, $\mathrm{Law}(\widehat{\mathbf{y}}_{i+1} \mid \widehat{\mathbf{y}}_{0:i}) = P_Y(\widehat{\mathbf{y}}_i, \cdot)$. Starting from $\widehat{\mathbf{y}}_0 \sim \pi_Y$, iteration gives the joint law

$$\pi_Y(\,\mathrm{d}\mathbf{y}_0) \prod_{j=0}^{n-1} P_Y(\mathbf{y}_j, \,\mathrm{d}\mathbf{y}_{j+1}),$$

which is the stationary target trajectory distribution. □

A.2 DEPENDENT VELOCITY ESTIMATION

Lemma 2 bounds the mixing of the transition observations. Lemma 3 then couples regularly spaced observations to independent copies. Lemma 4 bounds the exponential moment of each excess loss, and Lemma 5 combines these bounds to control the risk of an empirical minimizer.

Lemma 2 (Mixing of transition observations)**.** *Under Model 2, let $(\mathbf{o}_i)_{i\ge 0}$ be the transition observations in (18). Then $(\mathbf{o}_i)$ is a stationary, time-homogeneous Markov process. For every integer $\ell \ge 2$ and $\alpha \in (0,1)$,*

$$\beta_\ell \le \beta^X_{\ell-1} + \beta^Y_{\ell-1}, \qquad \tau_{\mathrm{mix}}(\alpha) \le 1 + \max\{\tau_X(\alpha/2), \tau_Y(\alpha/2)\}.$$

Proof. We express each observation as a random update of the current-state pair. Total-variation contraction then reduces its mixing bound to those of the two independent chains.

For probability laws μ, ν and a Markov kernel Q, write $(\mu Q)(A) = \int Q(u, A)\mu(\,\mathrm{d}u)$. Since $Q(u, A) \in [0, 1]$,

$$\begin{aligned}|(\mu Q)(A) - (\nu Q)(A)| &= \left| \int_0^1 (\mu\{u : Q(u, A) > s\} - \nu\{u : Q(u, A) > s\})\,\mathrm{d}s \right| \\ &\le \|\mu - \nu\|_{\mathrm{TV}}\,.\end{aligned}$$

Taking the supremum over A proves

$$\|\mu Q - \nu Q\|_{\mathrm{TV}} \le \|\mu - \nu\|_{\mathrm{TV}}\,. \tag{24}$$

This contraction also applies to a kernel that appends an independent draw. Combining it with the triangle inequality, using intermediate law $\nu_X \otimes \mu_Y$, gives

$$\begin{aligned}\|\mu_X \otimes \mu_Y - \nu_X \otimes \nu_Y\|_{\mathrm{TV}} &\le \|\mu_X \otimes \mu_Y - \nu_X \otimes \mu_Y\|_{\mathrm{TV}} + \|\nu_X \otimes \mu_Y - \nu_X \otimes \nu_Y\|_{\mathrm{TV}} \\ &\le \|\mu_X - \nu_X\|_{\mathrm{TV}} + \|\mu_Y - \nu_Y\|_{\mathrm{TV}}\,.\end{aligned}$$

To apply these inequalities, define the observation kernel as follows. Let $Q((\mathbf{x}, \mathbf{y}), \cdot)$ be the law of $(\mathbf{x}, \mathbf{x}^+, \mathbf{y}, \mathbf{y}^+)$ obtained by drawing $\mathbf{x}^+ \sim P_X(\mathbf{x}, \cdot)$ and $\mathbf{y}^+ \sim P_Y(\mathbf{y}, \cdot)$ independently. The Markov property and independence of the chains give

$$\mathrm{Law}(\mathbf{o}_{i+1} \mid \mathbf{o}_0, \ldots, \mathbf{o}_i) = Q((\mathbf{x}_{i+1}, \mathbf{y}_{i+1}), \cdot).$$

Thus $(\mathbf{o}_i)$ is time-homogeneous and Markov. Stationarity of the two trajectories gives its stationary law $\pi_O = (\pi_X \otimes \pi_Y)Q$.

Conditioning on $\mathbf{o}_0$ fixes $\mathbf{x}_1$ and $\mathbf{y}_1$. For $\ell \ge 2$, the two chains then evolve independently for $\ell - 1$ steps before the operation Q constructs $\mathbf{o}_\ell$. Therefore, almost surely,

$$\mathrm{Law}(\mathbf{o}_\ell \mid \mathbf{o}_0) = \left(P_X^{\ell-1}(\mathbf{x}_1, \cdot) \otimes P_Y^{\ell-1}(\mathbf{y}_1, \cdot)\right) Q.$$

Applying the two total-variation inequalities above yields

$$\begin{aligned}\|\mathrm{Law}(\mathbf{o}_\ell \mid \mathbf{o}_0) - \pi_O\|_{\mathrm{TV}} &\le \left\|P_X^{\ell-1}(\mathbf{x}_1, \cdot) \otimes P_Y^{\ell-1}(\mathbf{y}_1, \cdot) - \pi_X \otimes \pi_Y\right\|_{\mathrm{TV}} \\ &\le \left\|P_X^{\ell-1}(\mathbf{x}_1, \cdot) - \pi_X\right\|_{\mathrm{TV}} + \left\|P_Y^{\ell-1}(\mathbf{y}_1, \cdot) - \pi_Y\right\|_{\mathrm{TV}}\,.\end{aligned}$$

Averaging over $\mathbf{o}_0$, whose next-state marginals are π_X and π_Y, gives

$$\beta_\ell = \mathbb{E}\,\|\mathrm{Law}(\mathbf{o}_\ell \mid \mathbf{o}_0) - \pi_O\|_{\mathrm{TV}} \le \beta_{\ell-1}^X + \beta_{\ell-1}^Y.$$

For the mixing-time bound, (24) also implies monotonicity: for a Markov kernel P with stationary law π,

$$\left\|P^{k+1}(u, \cdot) - \pi\right\|_{\mathrm{TV}} = \left\|P^k(u, \cdot)P - \pi P\right\|_{\mathrm{TV}} \le \left\|P^k(u, \cdot) - \pi\right\|_{\mathrm{TV}}\,.$$

Hence each stationary-average mixing profile is nonincreasing. For $k = \max\{\tau_X(\alpha/2), \tau_Y(\alpha/2)\}$, we obtain $\beta_{k+1} \le \beta_k^X + \beta_k^Y \le \alpha$, proving the mixing-time bound. □

For an integer spacing $1 \le K \le m$, write $I_a = \{a + jK : j \ge 0,\ a + jK < m\}$ for $0 \le a < K$. The next lemma supplies independent observations within each I_a while preserving the joint law of the original data.

Lemma 3 (Residue-class coupling)**.** *Let $(\mathbf{o}_i)$ be a stationary, time-homogeneous Markov process on a Euclidean space with stationary law π_O and mixing profile β_ℓ defined by (18). For every integer $1 \le K \le m$, there are copies $(\overline{\mathbf{o}}_i)_{i=0}^{m-1}$ jointly coupled with $(\mathbf{o}_i)$ such that $(\overline{\mathbf{o}}_i)_{i \in I_a}$ are iid with law π_O for each a, and*

$$\mathbb{P}(\overline{\mathbf{o}}_i \ne \mathbf{o}_i \text{ for at least one } i) \le m\beta_K. \tag{25}$$

The coupling preserves the joint law of $(\mathbf{o}_i)$; the copies in different residue classes may be dependent.

Proof. We construct independent copies sequentially within each residue class, then bound the probability of any mismatch by a union bound. The construction uses maximal coupling (Berbee, 1979, Corollary 4.2.5): given $\mathbf{u} \sim \mu$, auxiliary randomness produces $\overline{\mathbf{u}} \sim \nu$ with $\mathbb{P}(\overline{\mathbf{u}} \neq \mathbf{u}) = \|\mu - \nu\|_{\mathrm{TV}}$.

Write P_O for the transition kernel. Fix a residue class $I_a = \{a, a+K, \dots\} \cap \{0, \dots, m-1\}$, with indices $i_1 < \cdots < i_r$, and keep $\overline{\mathbf{o}}_{i_1} = \mathbf{o}_{i_1} \sim \pi_O$. For $j \geq 2$, suppose the earlier copies have been constructed from the earlier observations in this class and independent auxiliary randomness. The Markov property gives

$$\mathrm{Law}\left(\mathbf{o}_{i_j} \mid \mathbf{o}_{i_1}, \dots, \mathbf{o}_{i_{j-1}}, \overline{\mathbf{o}}_{i_1}, \dots, \overline{\mathbf{o}}_{i_{j-1}}\right) = P_O^K(\mathbf{o}_{i_{j-1}}, \cdot).$$

Apply maximal coupling conditionally to this law and π_O, using fresh auxiliary randomness to construct $\overline{\mathbf{o}}_{i_j}$ from $\mathbf{o}_{i_j}$. The conditional law of the new copy is π_O, so it is independent of all earlier copies in this class. Its mismatch probability is

$$\begin{aligned}\mathbb{P}(\overline{\mathbf{o}}_{i_j} \neq \mathbf{o}_{i_j}) &= \mathbb{E}\left\|P_O^K(\mathbf{o}_{i_{j-1}}, \cdot) - \pi_O\right\|_{\mathrm{TV}} \\ &= \mathbb{E}_{\mathbf{o}\sim\pi_O}\left\|P_O^K(\mathbf{o}, \cdot) - \pi_O\right\|_{\mathrm{TV}} = \beta_K.\end{aligned}$$

Repeat this construction separately for each class, using independent auxiliary randomness. The original observations remain unchanged, and a union bound yields

$$\mathbb{P}(\exists i : \overline{\mathbf{o}}_i \neq \mathbf{o}_i) \leq \sum_{a=0}^{K-1} (|I_a| - 1)\beta_K = (m-K)\beta_K \leq m\beta_K.$$

□

For $\mathbf{v} \in \mathcal{V}$, recall the prediction error $\mathbf{g}_i(\mathbf{v}, t) = \mathbf{v}(\mathbf{h}_i(t)) - \mathbf{v}^\star(\mathbf{h}_i(t))$ from (22). Define the excess loss and its average over flow time by

$$\begin{aligned}\Delta_i(\mathbf{v}, t) &= \|\mathbf{v}(\mathbf{h}_i(t)) - \mathbf{r}_i\|^2 - \|\mathbf{v}^\star(\mathbf{h}_i(t)) - \mathbf{r}_i\|^2 \\ &= \|\mathbf{g}_i(\mathbf{v}, t)\|^2 + 2\left\langle \mathbf{g}_i(\mathbf{v}, t), \boldsymbol{\xi}_i(t)\right\rangle, \qquad A_i(\mathbf{v}) = \int_0^1 \Delta_i(\mathbf{v}, t)\,\mathrm{d}t.\end{aligned} \tag{26}$$

Thus $m^{-1}\sum_i A_i(\mathbf{v}) = \widehat{L}(\mathbf{v}) - \widehat{L}(\mathbf{v}^\star)$. The noise and moment assumptions ensure that the integrated losses are finite almost surely: $\mathbb{E}\|\boldsymbol{\xi}\|^2 \leq d\nu^2$ and every candidate has finite prediction risk. All time evaluations of a transition share its endpoints; we keep them together when bounding the excess loss.

Lemma 4 (Exponential bounds for excess loss)**.** *Under Assumptions 2 and 3, for a stationary transition and $\nu > 0$, define*

$$\lambda = \frac{1}{4\nu^2}, \qquad \Psi(u) = \frac{1}{4C_{4\to 2}^4}\left(1 - e^{-u/(16\nu^2)}\right), \quad u \geq 0. \tag{27}$$

Let $t \sim \mathcal{U}(0,1)$ be independent of the transition. For every $\mathbf{v} \in \mathcal{V}$ and transition index i,

$$\max\left\{\mathbb{E}e^{-\lambda\Delta_i(\mathbf{v},t)}, \mathbb{E}e^{-\lambda A_i(\mathbf{v})}\right\} \leq e^{-\Psi(R(\mathbf{v},\mathbf{v}^\star))}, \tag{28}$$

where Δ_i and A_i are defined in (26).

Proof. We first bound the exponential moment at a random flow time using the conditional noise assumption and the fourth-moment bound. Jensen's inequality then transfers this bound to the integral over flow time.

Fix $\mathbf{v}$ and draw a stationary transition and an independent uniform flow time, giving $(\mathbf{h}, \mathbf{r})$. Write $\mathbf{g} = \mathbf{v}(\mathbf{h}) - \mathbf{v}^\star(\mathbf{h})$, $G = \|\mathbf{g}\|^2$, and $R = R(\mathbf{v}, \mathbf{v}^\star) = \mathbb{E}G$. Use $\lambda = 1/(4\nu^2)$ from (27). Conditional on $\mathbf{h}$, the vector $\mathbf{g}$ is fixed, so Assumption 2 gives

$$\mathbb{E}\left[e^{-\lambda(G + 2\langle \mathbf{g}, \boldsymbol{\xi}\rangle)} \mid \mathbf{h}\right] \leq e^{-\lambda G + 2\lambda^2\nu^2 G} = e^{-G/(8\nu^2)}. \tag{29}$$

If $R > 0$, Cauchy–Schwarz and (17) yield

$$R \leq \frac{R}{2} + \mathbb{E}[G\mathbf{1}\{G \geq R/2\}] \leq \frac{R}{2} + \sqrt{\mathbb{E}G^2\,\mathbb{P}(G \geq R/2)}, \qquad \mathbb{P}(G \geq R/2) \geq \frac{R^2}{4\mathbb{E}G^2} \geq \frac{1}{4C_{4\to 2}^4}.$$

Averaging the conditional bound (29) and using this probability bound gives

$$\mathbb{E}e^{-\lambda(G+2\langle \mathbf{g},\boldsymbol{\xi}\rangle)} \le \mathbb{E}e^{-G/(8\nu^2)} \le 1 - \frac{1-e^{-R/(16\nu^2)}}{4C_{4\to 2}^4} \le e^{-\Psi(R)}.$$

For $R = 0$, (29) directly gives the same bound. This proves the random-time bound. For the integrated loss, Jensen's inequality gives, pointwise in the transition observation,

$$e^{-\lambda A_i(\mathbf{v})} = \exp\left(-\lambda \int_0^1 \Delta_i(\mathbf{v},t)\,\mathrm{d}t\right) \le \int_0^1 e^{-\lambda\Delta_i(\mathbf{v},t)}\,\mathrm{d}t.$$

Taking expectations and applying Tonelli's theorem proves (28). □

The next lemma uses only the exponential bound for a stationary loss and the mixing of its observations. We state it for a general loss $A_i(\mathbf{v}) = A(\mathbf{v}, \mathbf{o}_i)$ so that it applies to both integrated and sampled flow times.

Lemma 5 (Risk from empirical excess loss)**.** *Let $(\mathbf{o}_i)$ be a stationary Markov process as in Lemma 3, and let $\mathcal{V}$ be finite. Fix $\nu > 0$, $C_{4\to 2} \ge 1$, and use λ, Ψ from (27), with finite risks R as in (16). Suppose real-valued measurable losses $A_i(\mathbf{v}) = A(\mathbf{v}, \mathbf{o}_i)$ satisfy*

$$\mathbb{E}e^{-\lambda A_i(\mathbf{v})} \le e^{-\Psi(R(\mathbf{v},\mathbf{v}^\star))} \quad \textit{for every } \mathbf{v} \in \mathcal{V}. \tag{30}$$

Fix an integer $1 \le K \le m$ and $\varepsilon_{\mathrm{opt}} \ge 0$. Any possibly data-dependent $\widetilde{\mathbf{v}} \in \mathcal{V}$ with $m^{-1}\sum_i A_i(\widetilde{\mathbf{v}}) \le \varepsilon_{\mathrm{opt}}$ almost surely satisfies, for every $u > 0$,

$$\mathbb{P}\left(R(\widetilde{\mathbf{v}}, \mathbf{v}^\star) \ge u\right) \le m\beta_K + |\mathcal{V}| \exp\left\{\frac{m}{K}\left(\lambda\varepsilon_{\mathrm{opt}} - \Psi(u)\right)\right\}. \tag{31}$$

Proof. Apply Lemma 3 with spacing K to obtain copies $(\overline{\mathbf{o}}_i)$. Let $\overline{A}_i(\mathbf{v})$ be $A(\mathbf{v}, \overline{\mathbf{o}}_i)$. Each copied loss is a function of its copied observation. Thus, for fixed $\mathbf{v}$, independence within I_a and the stationary exponential bound (30) give

$$\mathbb{E}\exp\left(-\lambda \sum_{i\in I_a} \overline{A}_i(\mathbf{v})\right) \le e^{-|I_a|\Psi(R(\mathbf{v},\mathbf{v}^\star))}.$$

To combine the classes, apply Hölder's inequality with all K exponents equal to K (the identity when $K = 1$):

$$\begin{aligned}\mathbb{E}\exp\left(-\frac{\lambda}{K}\sum_i \overline{A}_i(\mathbf{v})\right) &\le \prod_{a=0}^{K-1}\left[\mathbb{E}\exp\left(-\lambda\sum_{i\in I_a}\overline{A}_i(\mathbf{v})\right)\right]^{1/K} \\ &\le \exp\left(-\frac{m}{K}\Psi(R(\mathbf{v},\mathbf{v}^\star))\right).\end{aligned}$$

This step requires no independence across residue classes. Markov's inequality now gives

$$\mathbb{P}\left(\frac{1}{m}\sum_i \overline{A}_i(\mathbf{v}) \le \varepsilon_{\mathrm{opt}}\right) \le \exp\left\{\frac{m}{K}\left(\lambda\varepsilon_{\mathrm{opt}} - \Psi(R(\mathbf{v},\mathbf{v}^\star))\right)\right\}.$$

Since Ψ is increasing, a union bound over candidates with $R(\mathbf{v}, \mathbf{v}^\star) \ge u$ bounds the probability that any such candidate has copied excess loss at most $\varepsilon_{\mathrm{opt}}$ by the exponential term in (31). On the event that all copies agree with the data, $\overline{A}_i(\mathbf{v}) = A_i(\mathbf{v})$ for every $i, \mathbf{v}$. The assumed empirical loss bound for $\widetilde{\mathbf{v}}$ then places it among the candidates just considered whenever its risk is at least u. Adding the coupling failure probability $m\beta_K$ proves the claim. □

Proof of Theorem 1. We apply Lemma 5 with the integrated excess losses in (26). Lemma 2 supplies the stationary Markov observations, and Lemma 4 supplies the exponential bound (30). It remains to choose the risk threshold. First suppose $\nu > 0$, and set

$$K = \tau_{\mathrm{mix}}(\delta/(3m)), \qquad L = \log(6K|\mathcal{V}|/\delta), \qquad u = 128C_{4\to 2}^4\nu^2 KL/m.$$

The sample-size condition (20) implies $K \le m$ and $u/(16\nu^2) = 8C_{4\to2}^4 KL/m \le 1/16$. By the definition of Ψ in (27) and $1 - e^{-s} \ge s/2$ for $s \in [0,1]$,

$$\Psi(u) \ge \frac{1}{4C_{4\to2}^4} \frac{u}{32\nu^2} = \frac{KL}{m}.$$

Realizability and empirical optimality give $m^{-1}\sum_i A_i(\widehat{\mathbf{v}}) = \widehat{L}(\widehat{\mathbf{v}}) - \widehat{L}(\mathbf{v}^\star) \le 0$. Lemma 5 with $\varepsilon_{\mathrm{opt}} = 0$ therefore yields

$$\mathbb{P}\left(R(\widehat{\mathbf{v}}, \mathbf{v}^\star) \ge u\right) \le m\beta_K + |\mathcal{V}|e^{-L} \le \frac{\delta}{3} + \frac{\delta}{6K} \le \frac{\delta}{2}. \tag{32}$$

This proves the stated $1-\delta$ guarantee. More generally, the same calculation holds for any $1 \le K \le m$ with $m\beta_K \le \delta/3$ and (20).

If $\nu = 0$, the conditional noise bound also holds with every positive parameter η. The preceding argument then bounds $\mathbb{P}(R(\widehat{\mathbf{v}}, \mathbf{v}^\star) > 128C_{4\to2}^4\eta^2 KL/m)$ by $\delta/2$. As $\eta \downarrow 0$, these events increase to $\{R(\widehat{\mathbf{v}}, \mathbf{v}^\star) > 0\}$. Continuity of probability proves the zero-risk conclusion. □

Corollary 1 (Sampled flow times)**.** *Under the assumptions and sample-size condition of Theorem 1, fix an integer $J \ge 1$ and draw $(t_{ij})_{0\le i<m,\, 1\le j\le J}$ iid from $\mathcal{U}(0,1)$, independently of the trajectories. Any minimizer $\widehat{\mathbf{v}}_J$ of*

$$\widehat{L}_J(\mathbf{v}) = \frac{1}{mJ}\sum_{i=0}^{m-1}\sum_{j=1}^{J} \left\|\mathbf{v}(\mathbf{h}_i(t_{ij})) - \mathbf{r}_i\right\|^2 \tag{33}$$

obeys (21), with $\widehat{\mathbf{v}}_J$ in place of $\widehat{\mathbf{v}}$, with probability at least $1-\delta$ over the trajectories and flow-time samples.

Proof. First suppose $\nu > 0$. Append the independent time vector $(t_{i1}, \dots, t_{iJ})$ to $\mathbf{o}_i$. The resulting process is stationary and Markov. Appending independent uniform draws cannot increase total variation by (24); contraction under projection onto $\mathbf{o}_i$ gives the reverse inequality. Its mixing profile is therefore exactly β_ℓ. Apply Lemma 5 with the loss $A_i(\mathbf{v}) = J^{-1}\sum_j \Delta_i(\mathbf{v}, t_{ij})$. To verify its exponential hypothesis, convexity gives

$$\exp\left(-\frac{\lambda}{J}\sum_j \Delta_i(\mathbf{v}, t_{ij})\right) \le \frac{1}{J}\sum_j e^{-\lambda\Delta_i(\mathbf{v}, t_{ij})}.$$

Taking expectations and applying the random-time bound in Lemma 4 to each summand gives $\mathbb{E}\exp(-\lambda J^{-1}\sum_j \Delta_i(\mathbf{v}, t_{ij})) \le e^{-\Psi(R(\mathbf{v},\mathbf{v}^\star))}$, with λ, Ψ from (27). This verifies (30). Realizability and empirical optimality give $m^{-1}\sum_i A_i(\widehat{\mathbf{v}}_J) \le 0$, so Lemma 5 applies with $\varepsilon_{\mathrm{opt}} = 0$. Use the same K, L, u as in the proof of Theorem 1; its limiting argument also covers $\nu = 0$. □

The corollary controls risk integrated over flow time, for each fixed J, without treating mJ evaluations as independent transitions. A deterministic time grid instead requires moment and noise assumptions for its time sampling distribution; transferring grid-averaged risk to integrated risk also requires control of the time-discretization error. For approximate minimization of the integrated loss, $\widehat{L}(\widetilde{\mathbf{v}}) \le \min_{\mathbf{v}\in\mathcal{V}} \widehat{L}(\mathbf{v}) + \varepsilon_{\mathrm{opt}}$ gives $m^{-1}\sum_i A_i(\widetilde{\mathbf{v}}) \le \varepsilon_{\mathrm{opt}}$ by realizability. Lemma 5 applies with the integrated excess loss in (26). Thus (31) also records the effect of an empirical optimization error; we do not certify this error for AdamW.

For (2), set $C = 128C_{4\to2}^4 \max\{1, \nu^2\}$ and $L = \log(6K|\mathcal{V}|/\delta)$. The condition $m \ge CKL\max\{1, \epsilon^{-1}\}$ implies both (20) and $128C_{4\to2}^4\nu^2 KL/m \le \epsilon$. Theorem 1 therefore gives risk at most ϵ with probability at least $1-\delta$.

A.3 The sample-size lower bound

Proof of Proposition 3. We construct two models with separated velocities whose observations agree when a slowly mixing Gaussian coordinate stays nonnegative. We first verify the flow and moment assumptions and compute the velocity separation. We then bound the probability that the data agree

and apply a two-point testing argument, as in Nagaraj et al. (2020, Appendix C.1, proof of Theorem 1). Finally, we relate the Gaussian coordinate's time scale to the observation mixing time.

For the prescribed integer $\tau \geq 1$, set $a = e^{-1/\tau}$ and $\sigma_\tau^2 = (1-a^2)^{-1}$. Fix a smooth function $\psi : \mathbb{R} \to [0,1]$ equal to 1 on $(-\infty, -1]$ and to 0 on $[0, \infty)$. For $G \sim N(0,1)$, write $M_2 = \mathbb{E}\psi(G)^2 \geq \mathbb{P}(G \leq -1) > 0$ and set $h = \sqrt{\epsilon/(\pi M_2)}$. Under $\theta \in \{-h, h\}$, let

$$\mathbf{y}_i = (C_i, U_i), \qquad C_{i+1} = aC_i + \eta_{i+1}, \qquad U_{i+1} = \theta\psi(C_i/\sigma_\tau) + \zeta_{i+1}, \tag{34}$$

where (η_i) and (ζ_i) are independent iid standard Gaussian sequences. Use a stationary extension of (C_i) to index -1 and set $U_0 = \theta\psi(C_{-1}/\sigma_\tau) + \zeta_0$. Then $\mathbf{y}$ is stationary and $C_i/\sigma_\tau \sim N(0,1)$. Independently, take $\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{I}_2)$. For $\mathbf{y} = (c,u)$, put $\mathbf{g}_\theta(\mathbf{y}) = (ac, \theta\psi(c/\sigma_\tau))$. The conditional transitions are $P_X(\mathbf{x}, \cdot) = N(\mathbf{0}, \mathbf{I}_2)$ and $P_{Y,\theta}(\mathbf{y}, \cdot) = N(\mathbf{g}_\theta(\mathbf{y}), \mathbf{I}_2)$.

For fixed $(\mathbf{x}, \mathbf{y})$, apply Lemma 6 with means $\mathbf{0}, \mathbf{g}_\theta(\mathbf{y})$ in dimension two. With $s_t = (1-t)^2 + t^2$, it gives

$$\mathbf{v}^\star_{\theta,t}(\mathbf{z}, \mathbf{x}, \mathbf{y}) = \frac{2t-1}{s_t}\mathbf{z} + \frac{1-t}{s_t}\mathbf{g}_\theta(\mathbf{y}), \qquad \mathcal{V} = \{\mathbf{v}^\star_{-h}, \mathbf{v}^\star_h\}. \tag{35}$$

Its flow is $\Phi^{\mathbf{v}^\star_\theta}_{\mathbf{x},\mathbf{y},t}(\mathbf{z}_0) = t\mathbf{g}_\theta(\mathbf{y}) + \sqrt{s_t}\,\mathbf{z}_0$. Lemma 6 verifies Assumption 1 for both candidates, with $L_t = |2t-1|/s_t$ and $\int_0^1 L_t\,\mathrm{d}t = \log 2$. It also gives $\boldsymbol{\xi} \mid \mathbf{h} \sim N(\mathbf{0}, \mathbf{I}_2/s_t)$, verifying Assumption 2 with $\nu^2 = 2$.

Write R_θ for risk under model θ. Only the second velocity coordinate differs, by $2h(1-t)\psi(c/\sigma_\tau)/s_t$ in absolute value. Using the integral (51),

$$R_\theta(\mathbf{v}^\star_{-\theta}, \mathbf{v}^\star_\theta) = 4h^2 M_2 \int_0^1 \frac{(1-t)^2}{s_t^2}\,\mathrm{d}t = \pi h^2 M_2 = \epsilon. \tag{36}$$

The same class contains the truth under either model. Since the candidate difference has norm at most $4h$,

$$\mathbb{E}_\theta \left\|\mathbf{v}^\star_{-\theta}(\mathbf{h}) - \mathbf{v}^\star_\theta(\mathbf{h})\right\|^4 \leq 16h^2 R_\theta(\mathbf{v}^\star_{-\theta}, \mathbf{v}^\star_\theta) = \frac{16}{\pi M_2} R_\theta(\mathbf{v}^\star_{-\theta}, \mathbf{v}^\star_\theta)^2.$$

This verifies Assumption 3 with the universal constant $C^4_{4\to 2} = 16/(\pi M_2)$.

The candidates thus satisfy the flow and moment assumptions and have risk separation ϵ. To show that a short trajectory cannot reliably distinguish them, we exhibit an event on which their observations coincide. Couple the two models using the same source trajectory, Gaussian noises, C-trajectory, and auxiliary randomness of the estimator. On $E = \{C_{-1}, C_0, \ldots, C_{m-1} \geq 0\}$, the observed $(U_0, \ldots, U_m)$ equal $(\zeta_0, \ldots, \zeta_m)$ under both models. Thus the data agree on E, including the initial observation U_0. Let $(Z_s)_{s\geq 0}$ be a stationary Ornstein–Uhlenbeck process with covariance $\mathbb{E}[Z_s Z_t] = e^{-|s-t|}$. The vectors $(C_{-1}, \ldots, C_{m-1})/\sigma_\tau$ and $(Z_0, Z_{1/\tau}, \ldots, Z_{m/\tau})$ have the same centered Gaussian law. Continuous nonnegativity implies nonnegativity at these sampled times. The survival formula in Aurzada & Baumgarten (2011, Equation (2.5)), evaluated at time $2m/\tau$ to account for their covariance $e^{-|s-t|/2}$, gives

$$\mathbb{P}(E) \geq \mathbb{P}(Z_s \geq 0 \text{ for all } s \in [0, m/\tau]) = \frac{1}{\pi}\arcsin(e^{-m/\tau}) \geq \frac{1}{\pi}e^{-m/\tau}. \tag{37}$$

Since the estimator takes values in $\mathcal{V} = \{\mathbf{v}^\star_{-h}, \mathbf{v}^\star_h\}$, let p be its conditional probability of returning $\mathbf{v}^\star_h$ given E. The conditional data law is the same under both models, so the conditional error probabilities are p under $\theta = -h$ and $1-p$ under $\theta = h$. Their maximum is at least $1/2$. Together with (36), the two-point argument gives

$$\max_{\theta \in \{-h,h\}} \mathbb{P}_\theta\left(R_\theta(\widetilde{\mathbf{v}}, \mathbf{v}^\star_\theta) \geq \epsilon\right) \geq \frac{1}{2}\mathbb{P}(E) \geq \frac{1}{2\pi}e^{-m/\tau}. \tag{38}$$

It remains to verify that τ is proportional to the observation mixing time. Write $\tau_{\text{mix},\theta}(\alpha)$ for the observation mixing time under model θ, as defined in (18). Let $\pi_C = N(0, \sigma_\tau^2)$ and let β_k^C be the stationary-average coefficient of (C_i). The joint-law identity (58), applied to this chain, gives $\beta_k^C = \|\mathrm{Law}(C_0, C_k) - \pi_C \otimes \pi_C\|_{\mathrm{TV}}$. For $r = a^k$, the relative entropy of the joint standardized Gaussian pair $(C_0/\sigma_\tau, C_k/\sigma_\tau)$ with respect to two independent standard Gaussians is $-\frac{1}{2}\log(1-r^2)$,

as computed in (59). Pinsker's inequality (Canonne, 2022, Lemma 2) gives the upper bound below. For the lower bound, the bivariate Gaussian sign formula (Aurzada & Baumgarten, 2011, proof of Lemma 2.11) shows that the probability that C_0, C_k have the same sign exceeds its value under independence by $\arcsin(a^k)/\pi$. Thus, for $k \geq 1$,

$$\frac{1}{\pi}\arcsin(a^k) \leq \beta_k^C \leq \frac{1}{2}\sqrt{-\log(1-a^{2k})}.$$

For $\ell \geq 3$, conditional on $\mathbf{o}_0$, the law of $C_{\ell-1}$ depends only on C_1 and is the $(\ell-2)$-step Gaussian transition law. Given $C_{\ell-1}$, draw $C_\ell, C_{\ell+1}$ and the independent noises defining $U_\ell, U_{\ell+1}$, as well as the independent source pair. This constructs $\mathbf{o}_\ell$ by a common Markov kernel. The contraction bound (24) gives the upper bound below; projecting $(\mathbf{o}_0, \mathbf{o}_\ell)$ onto (C_1, C_ℓ) gives the lower bound:

$$\beta_{\ell-1}^C \leq \beta_\ell \leq \beta_{\ell-2}^C. \tag{39}$$

The lower bound also holds for $\ell = 2$, and $\beta_1 = 1$ since adjacent observations share the continuous coordinate C_1. Since $\arcsin(e^{-1/4})/\pi > 1/4$ and $\sqrt{-\log(1-e^{-2})}/2 < 1/4$, these bounds imply, under both models,

$$\frac{\tau}{4} \leq \tau_{\mathrm{mix},\theta}(1/4) \leq \tau + 2 \leq 3\tau.$$

The source is iid, and the target is a projection of the mixing observation process, so both satisfy Model 2. With the mixing-time comparison established, an estimator that succeeds under both models with probability at least $1-\delta$ must satisfy, by (38),

$$m \geq \tau \log\frac{1}{2\pi\delta} \geq \frac{1}{3}\tau_{\mathrm{mix},\theta}(1/4)\log\frac{1}{2\pi\delta} \quad \text{for either } \theta,$$

which proves (23). □

A.4 Comparison of mixing tolerances

Remark (Comparing the upper and lower bounds). To compare the bounds, we relate the estimation spacing $K = \tau_{\mathrm{mix}}(\delta/(3m))$ to τ_{mix}. If $\beta_\ell \leq c_0\exp(-c_1\ell/\tau_{\mathrm{mix}})$ for all $\ell \geq 1$ and universal $c_0 \geq 1, c_1 > 0$, then

$$\tau_{\mathrm{mix}}(\delta/(3m)) \leq \left\lceil \frac{\tau_{\mathrm{mix}}}{c_1}\log\frac{3c_0 m}{\delta} \right\rceil.$$

If we strengthen the mixing assumption in Model 2 to require finite worst-case mixing time $\tau_{\mathrm{wc}}(1/4)$ for the observation process $(\mathbf{o}_i)$, with τ_{wc} defined in Section 2, then

$$\tau_{\mathrm{mix}}(\delta/(3m)) \leq \tau_{\mathrm{wc}}(\delta/(3m)) \leq \underbrace{\tau_{\mathrm{wc}}(1/4)\left\lceil \log_2\frac{3m}{\delta}\right\rceil}_{K}$$

by total-variation contraction (Levin & Peres, 2017, Lemmas 4.10–4.11 and (4.34)). This choice satisfies $m\beta_K \leq \delta/3$. The proof of Theorem 1 therefore applies with this K in both its risk bound and sample-size condition. The Gaussian models in Proposition 3 satisfy the exponential stationary-average profile above, by (39) and its Gaussian bounds. Their worst-case mixing time is infinite: conditional on $C_1 = c$, the mean of C_ℓ is $a^{\ell-1}c$, which is unbounded as $|c|$ grows, while its conditional variance is fixed. Thus $d_\ell^{\mathrm{wc}} = 1$ at every finite lag.

A.5 Accuracy of generated states and trajectories

The estimation theorem controls velocity error at stationary training pairs, but the learned recursion visits pairs drawn from its own rollout distribution. We connect these errors through the concentrability condition (41), then use contraction of the target transitions to control their accumulation across steps. Throughout this subsection, laws and expectations involving a fitted velocity are taken with the training data held fixed. For a candidate velocity $\mathbf{v} \in \mathcal{V}$, let $\widehat{\mathbf{y}}_i^{\mathbf{v}}$ be its rollout in (5) using $\Phi^{\mathbf{v}}$ and an initialization $\widehat{\mathbf{y}}_0 \sim \widehat{\mu}_0$ independent of the fresh source and training data. Write $\Gamma_i^{\mathbf{v}} = \mathrm{Law}(\mathbf{x}_i', \widehat{\mathbf{y}}_i^{\mathbf{v}})$ for the distribution of the source and generated state pair. The velocity error at a fixed pair is

$$\ell_{\mathbf{v}}(\mathbf{x},\mathbf{y}) = \int_0^1 \mathbb{E}_{q_{t|\mathbf{x},\mathbf{y}}}\left\|\mathbf{v}_t(\mathbf{z},\mathbf{x},\mathbf{y}) - \mathbf{v}_t^\star(\mathbf{z},\mathbf{x},\mathbf{y})\right\|^2 \mathrm{d}t, \qquad R(\mathbf{v},\mathbf{v}^\star) = \mathbb{E}_{\pi_X\otimes\pi_Y}\ell_{\mathbf{v}}. \tag{40}$$

The next theorem assumes that the error under $\Gamma_i^{\mathbf{v}}$ is controlled by its training average and combines this with target contraction.

Theorem 2 (Finite-sample rollout). *Under the assumptions of Theorem 1 and Assumption 1, fix $\epsilon > 0$, $\delta \in (0,1)$, and an integer $n \geq 1$. Suppose the candidate rollouts are well-defined, every candidate is $\widehat{L}_t$-Lipschitz in $\mathbf{z}$, uniformly over current-state pairs, for a common nonnegative function $\widehat{L}_t$ with $\Lambda = \int_0^1 \widehat{L}_t \, \mathrm{d}t < \infty$. Suppose, for some finite $C_{\mathrm{roll}} \geq 1$,*

$$\mathbb{E}_{\Gamma_i^{\mathbf{v}}} \ell_{\mathbf{v}} \leq C_{\mathrm{roll}} R(\mathbf{v}, \mathbf{v}^\star) \qquad \text{for all } \mathbf{v} \in \mathcal{V}, \quad 0 \leq i < n. \tag{41}$$

Suppose the target transitions and π_Y have finite second moments, $a_0 = W_2(\widehat{\mu}_0, \pi_Y) < \infty$, and $W_2(P_Y(\mathbf{y}, \cdot), P_Y(\mathbf{y}', \cdot)) \leq \rho \, \|\mathbf{y} - \mathbf{y}'\|$ for some $0 \leq \rho < 1$. If m satisfies (2), then, with probability at least $1 - \delta$ over the training trajectories, the rollout using the exact ODE flow of $\widehat{\mathbf{v}}$ satisfies

$$W_2(\mathrm{Law}(\widehat{\mathbf{y}}_n), \pi_Y) \leq \rho^n a_0 + \frac{e^{\Lambda} \sqrt{C_{\mathrm{roll}} \epsilon}}{1 - \rho}. \tag{42}$$

The bound separates the decaying initialization error from the accumulated estimation error, whose amplification by the target dynamics is at most $1/(1-\rho)$. The concentrability condition (41) transfers training accuracy to generated state pairs, as in offline reinforcement learning (Zanette & Wainwright, 2022, Section 4.3, (16b)).

Proposition 4 converts conditional velocity error to transition error, and Proposition 5 propagates this error along the trajectory. We condition on the fitted velocity when applying these deterministic bounds, then use Theorem 1 to prove the finite-sample guarantee.

Assumption 4 (Learned velocity Lipschitz bound). There is a nonnegative integrable function $\widehat{L}_t$ such that, for every current-state pair used in the result, all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$, and almost every $t \in [0,1]$,

$$\|\widehat{\mathbf{v}}_t(\mathbf{z}, \mathbf{x}, \mathbf{y}) - \widehat{\mathbf{v}}_t(\mathbf{z}', \mathbf{x}, \mathbf{y})\| \leq \widehat{L}_t \, \|\mathbf{z} - \mathbf{z}'\|, \qquad \Lambda = \int_0^1 \widehat{L}_t \, \mathrm{d}t < \infty. \tag{43}$$

A uniform Lipschitz bound on the learned velocity is also used by Bansal et al. (2026, Assumption 1(b)); here the bound may vary with flow time. Every candidate in Appendix B satisfies it with $\Lambda = \log 2$.

Proposition 4 (One-transition error). *Under Assumptions 1 and 4 at $(\mathbf{x}, \mathbf{y})$, the learned transition map satisfies*

$$W_2^2 \left((\Phi^{\widehat{\mathbf{v}}}_{\mathbf{x},\mathbf{y}})_{\#} P_X(\mathbf{x}, \cdot), P_Y(\mathbf{y}, \cdot) \right) \leq e^{2\Lambda} \ell_{\widehat{\mathbf{v}}}(\mathbf{x}, \mathbf{y}), \tag{44}$$

where $\ell_{\mathbf{v}}$ is the conditional risk in (40).

Proof of Proposition 4. We start the two flows from the same source draw and bound their separation using Gronwall's inequality. The population transport identity then converts this coupling cost to the claimed Wasserstein bound.

Fix current-state values $(\mathbf{x}, \mathbf{y})$ and start both flows from the same $\mathbf{x}^+ \sim P_X(\mathbf{x}, \cdot)$. All laws and expectations below are over this source draw, with $(\mathbf{x}, \mathbf{y})$ held fixed. Write $\mathbf{z}_t^\star = \Phi^{\mathbf{v}^\star}_{\mathbf{x},\mathbf{y},t}(\mathbf{x}^+)$, $\widehat{\mathbf{z}}_t = \Phi^{\widehat{\mathbf{v}}}_{\mathbf{x},\mathbf{y},t}(\mathbf{x}^+)$, and $b_t = \|\widehat{\mathbf{v}}_t(\mathbf{z}_t^\star, \mathbf{x}, \mathbf{y}) - \mathbf{v}_t^\star(\mathbf{z}_t^\star, \mathbf{x}, \mathbf{y})\|$. Subtract the two ODEs and apply Assumption 4 to obtain

$$\|\widehat{\mathbf{z}}_t - \mathbf{z}_t^\star\| \leq \int_0^t \left(\widehat{L}_s \, \|\widehat{\mathbf{z}}_s - \mathbf{z}_s^\star\| + b_s \right) \mathrm{d}s.$$

Gronwall's inequality and Cauchy–Schwarz give

$$\mathbb{E} \, \|\widehat{\mathbf{z}}_1 - \mathbf{z}_1^\star\|^2 \leq \mathbb{E} \left(\int_0^1 e^{\int_t^1 \widehat{L}_s \, \mathrm{d}s} b_t \, \mathrm{d}t \right)^2 \leq e^{2\Lambda} \int_0^1 \mathbb{E} b_t^2 \, \mathrm{d}t.$$

To express the right side in terms of conditional velocity error, Proposition 1 gives $\mathrm{Law}(\mathbf{z}_t^\star) = q_{t|\mathbf{x},\mathbf{y}}$, so the right side is $e^{2\Lambda} \ell_{\widehat{\mathbf{v}}}(\mathbf{x}, \mathbf{y})$. At $t = 1$, the coupled endpoints have the two laws in (44), proving the bound. □

For a fixed measurable transition map $\Phi_{\mathbf{x},\mathbf{y}}$ and its rollout (5), let $\widetilde{P}_{\mathbf{x},\mathbf{y}} = (\Phi_{\mathbf{x},\mathbf{y}})_{\#} P_X(\mathbf{x}, \cdot)$ and define

$$e_i^2 = \mathbb{E} W_2^2 \left(\widetilde{P}_{\mathbf{x}_i', \widehat{\mathbf{y}}_i}, P_Y(\widehat{\mathbf{y}}_i, \cdot) \right). \tag{45}$$

To propagate this error, we assume that the target kernel is Lipschitz in W_2.

Assumption 5 (Target kernel Lipschitz bound)**.** The target transition distributions have finite second moments, and there is a finite constant $\rho \geq 0$ such that, for all $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^d$,

$$W_2(P_Y(\mathbf{y}, \cdot), P_Y(\mathbf{y}', \cdot)) \leq \rho \|\mathbf{y} - \mathbf{y}'\| . \tag{46}$$

For example, consider transitions $\mathbf{g}(\mathbf{y}) + \boldsymbol{\eta}$, where $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ is ρ-Lipschitz and the noise law is independent of $\mathbf{y}$ with $\mathbb{E}\|\boldsymbol{\eta}\|^2 < \infty$. Use the same noise at both states. This gives $W_2(P_Y(\mathbf{y}, \cdot), P_Y(\mathbf{y}', \cdot)) \leq \|\mathbf{g}(\mathbf{y}) - \mathbf{g}(\mathbf{y}')\| \leq \rho\|\mathbf{y} - \mathbf{y}'\|$.

Proposition 5 (Rollout recursion)**.** *Under Model 1 and Assumption 5, suppose P_Y has a stationary distribution π_Y with finite second moment. Fix measurable transition maps $\Phi_{\mathbf{x},\mathbf{y}}$ and generate by (5), with $\widehat{\mathbf{y}}_0 \sim \widehat{\mu}_0$ independently of the fresh source. Suppose $a_0 = W_2(\widehat{\mu}_0, \pi_Y) < \infty$ and the errors e_i in (45) are finite. With $a_{i+1} = \rho a_i + e_i$,*

$$W_2(\mathrm{Law}(\widehat{\mathbf{y}}_n), \pi_Y) \leq a_n = \rho^n a_0 + \sum_{i=0}^{n-1} \rho^{n-1-i} e_i.$$

For a stationary target trajectory $\mathbf{y}_{0:n}$, the joint laws satisfy

$$W^2_{2,d_n}\left(\mathrm{Law}(\widehat{\mathbf{y}}_{0:n}), \mathrm{Law}(\mathbf{y}_{0:n})\right) \leq \sum_{i=0}^{n} a_i^2, \qquad d_n^2(\mathbf{y}, \mathbf{z}) = \sum_{i=0}^{n} \|\mathbf{y}_i - \mathbf{z}_i\|^2 , \tag{47}$$

where W_{2,d_n} uses the trajectory metric d_n.

Proof of Proposition 5. We couple the generated and stationary target trajectories while preserving the source transitions. We then bound the coupling cost recursively, iterate for the marginal guarantee, and sum the costs for the trajectory guarantee.

Couple $(\widehat{\mathbf{y}}_0, \mathbf{y}_0)$ optimally, independently of the fresh source. At each step, let $\mathcal{H}_i$ be the joint history of the source, generated, and comparison trajectories through time i, including the auxiliary randomness already used. We construct this history from $\mathbf{x}'_{0:i}$ and randomness independent of the full source trajectory. The Markov and tower-property argument of Lemma 1 therefore gives

$$\mathrm{Law}(\mathbf{x}'_{i+1} \mid \mathcal{H}_i) = P_X(\mathbf{x}'_i, \cdot), \qquad \mathrm{Law}(\widehat{\mathbf{y}}_{i+1} \mid \mathcal{H}_i) = \widetilde{P}_{\mathbf{x}'_i, \widehat{\mathbf{y}}_i}.$$

Keep this source transition and its generated output. Conditional on $\mathcal{H}_i$, couple that output optimally to $\mathbf{u}_{i+1} \sim P_Y(\widehat{\mathbf{y}}_i, \cdot)$, and couple $\mathbf{u}_{i+1}$ optimally to $\mathbf{y}_{i+1} \sim P_Y(\mathbf{y}_i, \cdot)$. Measurable optimal couplings, disintegration, and gluing on $\mathbb{R}^d$ justify these conditional draws (Villani, 2009, Chapter 1 and Corollary 5.22). They use only the history, the generated output, and fresh auxiliary randomness, so future source transitions are unchanged. The comparison process has conditional transition law $P_Y(\mathbf{y}_i, \cdot)$ and hence the stationary target trajectory law.

To bound the cost of this coupling, let $\Delta_i = (\mathbb{E}\,\|\widehat{\mathbf{y}}_i - \mathbf{y}_i\|^2)^{1/2}$. By the definition of e_i in (45) and Assumption 5, the two costs satisfy

$$\mathbb{E}\,\|\widehat{\mathbf{y}}_{i+1} - \mathbf{u}_{i+1}\|^2 = e_i^2, \qquad \mathbb{E}\left[\|\mathbf{u}_{i+1} - \mathbf{y}_{i+1}\|^2 \mid \mathcal{H}_i\right] \leq \rho^2 \|\widehat{\mathbf{y}}_i - \mathbf{y}_i\|^2 .$$

Minkowski's inequality now gives

$$\begin{aligned}\Delta_{i+1} &\leq \left(\mathbb{E}\,\|\widehat{\mathbf{y}}_{i+1} - \mathbf{u}_{i+1}\|^2\right)^{1/2} + \left(\mathbb{E}\,\|\mathbf{u}_{i+1} - \mathbf{y}_{i+1}\|^2\right)^{1/2} \\ &\leq e_i + \rho\Delta_i.\end{aligned}$$

Since $\Delta_0 = a_0$, induction gives $\Delta_i \leq a_i$, and solving the recursion for a_i yields the marginal bound. For the joint law, the same coupling has expected trajectory cost $\mathbb{E}\sum_{i=0}^n \|\widehat{\mathbf{y}}_i - \mathbf{y}_i\|^2 = \sum_{i=0}^n \Delta_i^2 \leq \sum_{i=0}^n a_i^2$, which proves (47). □

The concentrability condition (41) is uniform over $\mathcal{V}$, so it applies to the fitted candidate. The constant C_{roll} may depend on the horizon, source initialization, and target initialization; a bound uniform in the horizon requires the same constant for all i. Appendix F.3 gives the precise comparison with off-policy evaluation. Verifying this condition for broader model classes remains an open question.

For example, if $\Gamma_i^{\mathbf{v}}$ has density at most M relative to $\pi_X \otimes \pi_Y$, uniformly over candidates and times, then (41) holds with $C_{\mathrm{roll}} = M$.

Proof of Theorem 2. We use the estimation theorem to bound population risk, then concentrability to control transition error at generated states. The rollout recursion and a geometric sum give the result.

Under (2), Theorem 1 gives $R(\widehat{\mathbf{v}}, \mathbf{v}^\star) \le \epsilon$ with probability at least $1-\delta$, as verified in Appendix A.2. Fix any training sample in this event and use the maps $\Phi_{\mathbf{x},\mathbf{y}} = \Phi^{\widehat{\mathbf{v}}}_{\mathbf{x},\mathbf{y}}$ in (45). The common candidate Lipschitz bound gives Assumption 4 for the selected velocity. Proposition 4 and the concentrability condition (41) give, for every $i < n$,

$$e_i^2 \le e^{2\Lambda} \mathbb{E}_{\Gamma_i^{\widehat{\mathbf{v}}}} \ell_{\widehat{\mathbf{v}}}(\mathbf{x}, \mathbf{y}) \le e^{2\Lambda} C_{\mathrm{roll}} R(\widehat{\mathbf{v}}, \mathbf{v}^\star). \tag{48}$$

These errors are finite at every $i < n$, and the target contraction assumption gives Assumption 5. Thus Proposition 5, with the same source and initialization, gives

$$W_2(\mathrm{Law}(\widehat{\mathbf{y}}_n), \pi_Y) \le \rho^n a_0 + e^{\Lambda} \sqrt{C_{\mathrm{roll}} R(\widehat{\mathbf{v}}, \mathbf{v}^\star)} \sum_{j=0}^{n-1} \rho^j. \tag{49}$$

For $0 \le \rho < 1$, the geometric sum is $(1-\rho^n)/(1-\rho) \le (1-\rho)^{-1}$. Substituting $R(\widehat{\mathbf{v}}, \mathbf{v}^\star) \le \epsilon$ proves (42). No additional failure probability is incurred. □

B Gaussian transitions, estimation, and rollout

This example verifies the estimation assumptions explicitly and bounds rollout error by coupling the two recursions with the same Gaussian noise. We first record the conditional calculation also used in the lower bound.

Lemma 6 (Flow between Gaussian transitions). *Fix $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, draw independent $\mathbf{x}^+ \sim N(\mathbf{a}, \mathbf{I}_d)$ and $\mathbf{y}^+ \sim N(\mathbf{b}, \mathbf{I}_d)$, and set $\mathbf{z}_t = (1-t)\mathbf{x}^+ + t\mathbf{y}^+$, $\mathbf{r} = \mathbf{y}^+ - \mathbf{x}^+$, and $s_t = (1-t)^2 + t^2$. The conditional mean velocity and its ODE flow are*

$$\mathbf{v}_t^\star(\mathbf{z}) = \frac{2t-1}{s_t}\mathbf{z} - \frac{t}{s_t}\mathbf{a} + \frac{1-t}{s_t}\mathbf{b}, \qquad \Phi_t(\mathbf{z}_0) = (1-t)\mathbf{a} + t\mathbf{b} + \sqrt{s_t}(\mathbf{z}_0 - \mathbf{a}). \tag{50}$$

For each $t \in [0,1]$, the residual $\mathbf{v}_t^\star(\mathbf{z}_t) - \mathbf{r}$ conditional on $\mathbf{z}_t$ has law $N(\mathbf{0}, \mathbf{I}_d/s_t)$. The velocity satisfies Assumption 1. For each $t \in [0,1]$, the map $\mathbf{z} \mapsto \mathbf{v}_t^\star(\mathbf{z})$ is Lipschitz with constant $L_t = |2t-1|/s_t$, and $\int_0^1 L_t \, \mathrm{d}t = \log 2$.

Proof. Write $\mathbf{m}_t = (1-t)\mathbf{a} + t\mathbf{b}$. The pair $(\mathbf{r}, \mathbf{z}_t)$ is jointly Gaussian, with means $\mathbf{b} - \mathbf{a}, \mathbf{m}_t$, variances $2\mathbf{I}_d, s_t\mathbf{I}_d$, and cross covariance $(2t-1)\mathbf{I}_d$. Gaussian conditioning gives

$$\mathbb{E}[\mathbf{r} \mid \mathbf{z}_t = \mathbf{z}] = \mathbf{b} - \mathbf{a} + \frac{2t-1}{s_t}(\mathbf{z} - \mathbf{m}_t), \qquad \mathrm{Cov}(\mathbf{r} \mid \mathbf{z}_t) = \left(2 - \frac{(2t-1)^2}{s_t}\right)\mathbf{I}_d = \frac{\mathbf{I}_d}{s_t}.$$

Expanding the mean proves the velocity formula. Since $\dot{s}_t = 2(2t-1)$, differentiating the proposed flow gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(\mathbf{z}_0) = \mathbf{b} - \mathbf{a} + \frac{2t-1}{s_t}(\Phi_t(\mathbf{z}_0) - \mathbf{m}_t) = \mathbf{v}_t^\star(\Phi_t(\mathbf{z}_0)), \qquad \Phi_0(\mathbf{z}_0) = \mathbf{z}_0.$$

For all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$,

$$\|\mathbf{v}_t^\star(\mathbf{z}) - \mathbf{v}_t^\star(\mathbf{z}')\| = \frac{|2t-1|}{s_t}\|\mathbf{z} - \mathbf{z}'\|.$$

The affine coefficients are bounded because $s_t \ge 1/2$, and

$$\int_0^1 \frac{|2t-1|}{s_t}\,\mathrm{d}t = 2\int_{1/2}^1 \frac{2t-1}{s_t}\,\mathrm{d}t = \log s_1 - \log s_{1/2} = \log 2.$$

These bounds verify Assumption 1. □

For $s_t = (1-t)^2 + t^2$, symmetry under $t \mapsto 1-t$ gives the first two equalities below, and the substitution $q = 2t-1$ gives the last:

$$\int_0^1 \frac{t^2}{s_t^2}\,\mathrm{d}t = \int_0^1 \frac{(1-t)^2}{s_t^2}\,\mathrm{d}t = \frac{1}{2}\int_0^1 \frac{\mathrm{d}t}{s_t} = \frac{\pi}{4}. \tag{51}$$

Consider independent stationary processes $X_{i+1} = aX_i + \epsilon_{i+1}$ and $Y_{i+1} = bY_i + \zeta_{i+1}$ with independent standard Gaussian noise. The unknown (a, b) belongs to a known finite set $\Theta_0 \subset [-r, r]^2$, where $0 < r < 1$. With $s_t = (1-t)^2 + t^2$, define

$$\mathbf{v}_t^{u,w}(z, x, y) = \frac{2t-1}{s_t} z - \frac{t}{s_t} ux + \frac{1-t}{s_t} wy, \qquad (u, w) \in \Theta_0. \tag{52}$$

Its exact time-one map is $\Phi_{x,y}^{u,w}(z) = z - ux + wy$.

Proposition 6 (Gaussian estimation and rollout)**.** *In the Gaussian model above, fix $\delta \in (0, 1)$, set $K = \tau_{\mathrm{mix}}(\delta/(3m))$, and suppose $m \geq 1024K \log(6K|\Theta_0|/\delta)$. With probability at least $1 - \delta$, the minimizer of (15) over (52) satisfies*

$$R(\widehat{\mathbf{v}}, \mathbf{v}^\star) \leq 2048 \frac{K \log(6K|\Theta_0|/\delta)}{m}. \tag{53}$$

On this event, conditional on the training data, the rollout using its exact ODE flow, a fresh stationary source, and an independent initialization $\widehat{Y}_0 \sim \widehat{\mu}_0$ with $a_0 = W_2(\widehat{\mu}_0, \pi_Y) < \infty$ satisfies, for every integer $n \geq 0$,

$$W_2(\mathrm{Law}(\widehat{Y}_n), \pi_Y) \leq r^n a_0 + \sqrt{\frac{8}{\pi} R(\widehat{\mathbf{v}}, \mathbf{v}^\star)} \sum_{j=0}^{n-1} r^j. \tag{54}$$

More generally, for $P_X(x, \cdot) = N(ax, \sigma_X^2)$ and $P_Y(y, \cdot) = N(by, \sigma_Y^2)$ with $\sigma_X, \sigma_Y > 0$, the population map is

$$\Phi_{x,y}(z) = by + \frac{\sigma_Y}{\sigma_X}(z - ax). \tag{55}$$

The scalar population flow is increasing and transports the source law to the target law, so it agrees with this quantile map.

Proof of Proposition 6. We first compute the population flow, then verify the moment and mixing assumptions to apply Theorem 1. For the rollout bound, we couple the recursions using the same Gaussian noise and use the risk identity to control their separation.

Apply Lemma 6 in dimension one with means ax, by. The population velocity is (52) with $(u, w) = (a, b)$, and the conditional noise variance is $1/s_t \leq 2$. Applying the flow formula with means ux, wy gives, for every candidate,

$$z_t = (1-t)ux + twy + \sqrt{s_t}(z_0 - ux), \qquad z_1 = z_0 - ux + wy. \tag{56}$$

Lemma 6 also verifies flow regularity and Assumption 4 for every candidate, with $\Lambda = \log 2$.

To check the fourth-moment assumption, fix $(u, w) \in \Theta_0$ and write $R = R(\mathbf{v}^{u,w}, \mathbf{v}^{a,b})$, $A = (u-a)^2/(1-a^2)$, and $B = (w-b)^2/(1-b^2)$. At fixed t, the prediction error under $\pi_X \otimes \pi_Y$ is centered Gaussian with variance $f_t = [t^2 A + (1-t)^2 B]/s_t^2$. The integrals in (51) give

$$R(\mathbf{v}^{u,w}, \mathbf{v}^{a,b}) = \frac{\pi}{4}(A + B). \tag{57}$$

To bound the fourth moment by R^2, use $s_t \geq 1/2$ and $t^2, (1-t)^2 \leq s_t$ to obtain $f_t \leq 2(A + B)$. Thus

$$\mathbb{E}|\mathbf{v}^{u,w}(\mathbf{h}) - \mathbf{v}^{a,b}(\mathbf{h})|^4 = 3 \int_0^1 f_t^2 \, \mathrm{d}t \leq 6(A + B)R = \frac{24}{\pi} R^2 \leq 8R^2.$$

Together with the conditional noise calculation above, this verifies Assumptions 2 and 3 with $\nu^2 = 2$ and $C_{4\to 2}^4 = 8$; realizability follows from $(a, b) \in \Theta_0$.

It remains to check stationary mixing before applying the estimation theorem. Lemma 2 gives $\beta_\ell \leq \beta_{\ell-1}^X + \beta_{\ell-1}^Y$ for $\ell \geq 2$. To bound β_ℓ^X, use the identity

$$\beta_\ell^X = \|\mathrm{Law}(X_0, X_\ell) - \pi_X \otimes \pi_X\|_{\mathrm{TV}}. \tag{58}$$

Indeed, the two joint laws have the same first marginal π_X; integrating the total variation of their conditional second-coordinate laws gives the definition of β_ℓ^X. Write $r_\ell = a^\ell$. Standardizing the

two coordinates gives a centered Gaussian pair with covariance $\mathbf{\Sigma}_\ell = \left(\begin{smallmatrix} 1 & r_\ell \\ r_\ell & 1 \end{smallmatrix}\right)$. For laws μ, ν, write $D_{\mathrm{KL}}(\mu\|\nu)$ for relative entropy. The Gaussian density formula gives

$$D_{\mathrm{KL}}\left(N(\mathbf{0}, \mathbf{\Sigma}_\ell)\|N(\mathbf{0}, \mathbf{I}_2)\right) = \frac{1}{2}\left(\operatorname{tr}\mathbf{\Sigma}_\ell - 2 - \log\det\mathbf{\Sigma}_\ell\right) = -\frac{1}{2}\log(1 - r_\ell^2). \tag{59}$$

Pinsker's inequality, $\|\mu - \nu\|_{\mathrm{TV}}^2 \le D_{\mathrm{KL}}(\mu\|\nu)/2$ (Canonne, 2022, Lemma 2), and $-\log(1-q) \le q/(1-q)$ give

$$\beta_\ell^X \le \frac{1}{2}\sqrt{-\log(1 - a^{2\ell})} \le \frac{r^\ell}{2\sqrt{1 - r^2}}.$$

Applying the same calculation to Y yields $\beta_\ell \le r^{\ell-1}/\sqrt{1-r^2}$ for $\ell \ge 2$. All estimation assumptions therefore hold. Theorem 1, with $\nu^2 = 2$ and $C_{4\to 2}^4 = 8$, now gives (53).

With the risk bound established, we turn to rollout error. Condition on the fitted parameters $(\widehat{a}, \widehat{b})$ and write $R = R(\widehat{\mathbf{v}}, \mathbf{v}^\star)$. Couple $\widehat{Y}_0 \sim \widehat{\mu}_0$ and $Y_0 \sim \pi_Y$ so that $\|\widehat{Y}_0 - Y_0\|_{L^2} = a_0$, independently of the fresh stationary source. The source noise variables $\epsilon_{i+1} = X'_{i+1} - aX'_i$ are iid standard Gaussian, with ϵ_{i+1} independent of the histories through time i. Using them to drive both recursions, (56) gives

$$\begin{aligned} \widehat{Y}_{i+1} &= \widehat{b}\widehat{Y}_i + (a - \widehat{a})X'_i + \epsilon_{i+1}, \\ Y_{i+1} &= bY_i + \epsilon_{i+1}. \end{aligned}$$

The comparison trajectory has the stationary target law. Subtracting the updates cancels the shared noise and gives

$$\widehat{Y}_{i+1} - Y_{i+1} = \widehat{b}(\widehat{Y}_i - Y_i) + (\widehat{b} - b)Y_i + (a - \widehat{a})X'_i.$$

Minkowski's inequality, $|\widehat{b}| \le r$, and (57) therefore give

$$\begin{aligned} \|\widehat{Y}_{i+1} - Y_{i+1}\|_{L^2} &\le r\|\widehat{Y}_i - Y_i\|_{L^2} + \frac{|\widehat{b} - b|}{\sqrt{1 - b^2}} + \frac{|\widehat{a} - a|}{\sqrt{1 - a^2}} \\ &\le r\|\widehat{Y}_i - Y_i\|_{L^2} + \sqrt{8R/\pi}. \end{aligned}$$

This estimate uses only the stationary second moments of X'_i and Y_i; they may be dependent under the coupling. Iterating from the initial cost a_0 proves (54). □

C Examples for the conditional noise assumption

Bounded displacements, bounded additive noise, and Gaussian transition kernels satisfy Assumption 2 under the conditions below. Each calculation conditions on the full regression input $\mathbf{h} = (t, \mathbf{z}_t, \mathbf{x}, \mathbf{y})$ and verifies only the conditional noise assumption.

Bounded displacements. Recall $\mathbf{r} = \mathbf{y}^+ - \mathbf{x}^+$ and $\boldsymbol{\xi} = \mathbb{E}[\mathbf{r} \mid \mathbf{h}] - \mathbf{r}$ from Section 4.2. Suppose $\|\mathbf{r}\| \le B$ almost surely. For every fixed direction $\mathbf{a} \in \mathbb{R}^d$, the scalar $\langle \mathbf{a}, \mathbf{r}\rangle$ lies in $[-B\|\mathbf{a}\|, B\|\mathbf{a}\|]$, including after conditioning on $\mathbf{h}$. Since $\langle \mathbf{a}, \boldsymbol{\xi}\rangle = \mathbb{E}[\langle \mathbf{a}, \mathbf{r}\rangle \mid \mathbf{h}] - \langle \mathbf{a}, \mathbf{r}\rangle$, conditional Hoeffding's lemma (Vershynin, 2018, Section 2.2) gives

$$\mathbb{E}\left[e^{\theta\langle \mathbf{a}, \boldsymbol{\xi}\rangle} \mid \mathbf{h}\right] \le \exp\left(\frac{\theta^2(2B\|\mathbf{a}\|)^2}{8}\right) = \exp\left(\frac{\theta^2 B^2\|\mathbf{a}\|^2}{2}\right).$$

Thus $\nu = B$ suffices. In particular, if both processes take values in a ball of radius B, then $\|\mathbf{y}^+ - \mathbf{x}^+\| \le 2B$, so $\nu = 2B$ suffices for either discrete or continuous state distributions.

Bounded additive noise. Subtracting any function of $(\mathbf{x}, \mathbf{y})$ from the displacement leaves its conditional residual unchanged. This allows the preceding Hoeffding bound to apply when only the additive noise is bounded. Consider

$$\mathbf{x}^+ = \mathbf{f}(\mathbf{x}) + \boldsymbol{\eta}^X, \qquad \mathbf{y}^+ = \mathbf{g}(\mathbf{y}) + \boldsymbol{\eta}^Y, \qquad \|\boldsymbol{\eta}^X\|, \|\boldsymbol{\eta}^Y\| \le B \quad \text{almost surely},$$

where $\mathbf{f}, \mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ may be nonlinear and the noise distributions may depend on the current states. Conditional on $\mathbf{h}$, the term $\mathbf{g}(\mathbf{y}) - \mathbf{f}(\mathbf{x})$ is fixed. Consequently,

$$\boldsymbol{\xi} = \mathbb{E}[\boldsymbol{\eta}^Y - \boldsymbol{\eta}^X \mid \mathbf{h}] - (\boldsymbol{\eta}^Y - \boldsymbol{\eta}^X).$$

The noise difference $\boldsymbol{\eta}^Y - \boldsymbol{\eta}^X$ has norm at most $2B$, so the preceding calculation gives $\nu = 2B$. The states and transition means themselves need not be bounded.

Gaussian transitions. Suppose the two transition kernels are Gaussian with arbitrary mean functions and covariance matrices bounded above by $\sigma^2\mathbf{I}$, uniformly in the current states. Fix $(\mathbf{x}, \mathbf{y}, t)$. The endpoints are independent under (6), and $(\mathbf{r}, \mathbf{z}_t)$ is jointly Gaussian. Gaussian conditioning therefore gives a centered Gaussian residual with

$$\mathrm{Var}(\langle \mathbf{a}, \boldsymbol{\xi} \rangle \mid \mathbf{h}) = \mathrm{Var}(\langle \mathbf{a}, \mathbf{r} \rangle \mid \mathbf{h}) \le \mathrm{Var}(\langle \mathbf{a}, \mathbf{r} \rangle \mid \mathbf{x}, \mathbf{y}, t) \le 2\sigma^2 \|\mathbf{a}\|^2 .$$

The Gaussian moment-generating function yields (14) with $\nu = \sqrt{2}\sigma$, including for nonlinear conditional means.

For other sub-Gaussian transition models, Assumption 2 requires a bound conditional on the full input $\mathbf{h}$, including $\mathbf{z}_t = (1-t)\mathbf{x}^+ + t\mathbf{y}^+$. Sub-Gaussian bounds on the stationary states, or on additive transition noise conditional only on the current states, do not by themselves give the conditional bound used in Lemma 4.

D Experimental details

The Gaussian and nonlinear experiments use seeds $0, 1, 2$, with $d = 10$ for the main panels and $d = 100$ as a stress test. The main panels use $m \in \{512, 1024, 2048, 4096, 8192, 16384, 32768\}$ training transitions; the stress test uses $m \in \{256, 512, 1024, 2048, 4096, 8192\}$. Architecture and optimization settings stay fixed across training sizes. Each seed produces one complete experiment run. We report the mean across these three runs, with standard error equal to the sample standard deviation divided by $\sqrt{3}$. Each training trajectory starts from zero, discards 1,000 states, and retains 32,769 states at $d = 10$ or 8,193 at $d = 100$. Training sets are nested prefixes; process parameters and evaluation draws stay fixed across training sizes within each seed. Independent random streams generate parameters, source and target training trajectories, evaluation contexts, projection directions, and evaluation noise.

At each training update, we sample transition indices uniformly with replacement and draw one fresh independent $t \sim \mathcal{U}(0, 1)$ per sampled transition. Each repeated occurrence of a transition receives a new flow time, including repeated indices within a minibatch. Source and target transitions retain their paired indices. The minibatch loss is an unbiased estimate, conditional on the training trajectories and current network, of $\widehat{L}(\mathbf{v})/d$, where (15) defines the time-integrated objective. The factor $1/d$ averages over coordinates. The velocity network takes $(t, \mathbf{z}, \mathbf{x}, \mathbf{y})$ and outputs a vector in $\mathbb{R}^d$. It has three SiLU hidden layers, width 64 at $d = 10$ and 256 at $d = 100$, and a zero-initialized final layer. Training uses 2,400 AdamW updates with batch size 256, initial learning rate 2×10^{-3}, cosine decay, weight decay 10^{-5}, and gradient norm clipping at 5. We compare the full conditional flow with two learned baselines. All three learned flows share initialization, minibatch indices, and flow times. The source-state ablation sets $\mathbf{x}_i$ to zero. Training still uses $\mathbf{x}_{i+1}$ to form the interpolation and conditions on $\mathbf{y}_i$; generation starts from $\mathbf{x}'_{i+1}$ and conditions on $\widehat{\mathbf{y}}_i$. The marginal-flow baseline zeros both context inputs and fits an unconditional flow from the same source and target next-state observations. Its ODE defines a single map T applied pointwise as $\widehat{\mathbf{y}}_{i+1} = T(\mathbf{x}'_{i+1})$. It is trained to match the marginal state distributions and uses the same architecture, training budget, and solver. Every learned flow uses the final training update; no test metric selects a checkpoint or hyperparameter. For a fixed current-state pair $(\mathbf{x}, \mathbf{y})$, generation initializes $\mathbf{z}_0 = \mathbf{x}^+$ and applies 32 explicit midpoint steps, with $h = 1/32$ and $k = 0, \ldots, 31$:

$$\mathbf{a}_k = \widehat{\mathbf{v}}_{kh}(\mathbf{z}_k, \mathbf{x}, \mathbf{y}), \qquad \mathbf{z}_{k+1} = \mathbf{z}_k + h\widehat{\mathbf{v}}_{(k+1/2)h}(\mathbf{z}_k + h\mathbf{a}_k/2, \mathbf{x}, \mathbf{y}). \tag{60}$$

We evaluate the next-state distribution at 128 fresh current-state pairs. For each pair c, each method produces 512 next states. We project them onto $J_{\mathrm{dir}} = 64$ independent directions drawn uniformly from the unit sphere. Let $\widehat{q}_{c,j,k}$ be the kth smallest generated projection in direction j, and let $q_{c,j}(p)$ be the corresponding true target-transition quantile at probability p. The reported error is

$$\left(\frac{1}{128 \cdot J_{\mathrm{dir}} \cdot 512} \sum_{c=1}^{128} \sum_{j=1}^{J_{\mathrm{dir}}} \sum_{k=1}^{512} \left[\widehat{q}_{c,j,k} - q_{c,j}\left(\frac{k - 1/2}{512} \right) \right]^2 \right)^{1/2} . \tag{61}$$

Evaluation uses fresh samples independent of training, with current-state pairs, projection directions, and source draws shared across methods. For every method, we evaluate conditional-transition

recovery at the supplied $(\mathbf{x}, \mathbf{y})$; marginal flow ignores these two context inputs. Projected Gaussian quantiles are analytic; projected mixture quantiles are computed by 55 bisection steps on their exact cumulative distribution functions. Thus the metric does not require an independently sampled target reference. The oracle error is constant across training sizes because its parameters and evaluation draws are fixed.

D.1 Gaussian transition experiments

The processes have transitions $P_X(\mathbf{x}, \cdot) = N(\mathbf{A}_X\mathbf{x}, \mathbf{Q}_X)$ and $P_Y(\mathbf{y}, \cdot) = N(\mathbf{A}_Y\mathbf{y}, \mathbf{Q}_Y)$. For each seed, all four matrices have independent random orthogonal eigenbases. Source transition eigenvalue magnitudes are uniform on $[0.72, 0.94]$ and target magnitudes on $[0.15, 0.58]$, with independent random signs. Noise covariance eigenvalues λ satisfy $\log\lambda \sim \mathcal{U}(\log 0.25, \log 1.75)$. The true-kernel oracle is

$$\Phi^{\mathrm{or}}_{\mathbf{x},\mathbf{y}}(\mathbf{z}) = \mathbf{A}_Y\mathbf{y} + \mathbf{Q}_Y^{1/2}\mathbf{Q}_X^{-1/2}(\mathbf{z} - \mathbf{A}_X\mathbf{x}). \tag{62}$$

It maps $N(\mathbf{A}_X\mathbf{x}, \mathbf{Q}_X)$ to $N(\mathbf{A}_Y\mathbf{y}, \mathbf{Q}_Y)$.

D.2 Nonlinear mixture benchmark

For independent random orthogonal matrices $\mathbf{R}_X, \mathbf{R}_Y$, define

$$\mathbf{f}(\mathbf{x}) = 0.85\mathbf{R}_X \tanh(\mathbf{R}_X^\top\mathbf{x}), \qquad \mathbf{g}(\mathbf{y}) = \mathbf{R}_Y\left[0.55\tanh(\mathbf{R}_Y^\top\mathbf{y}) + 0.2\sin(2\mathbf{R}_Y^\top\mathbf{y})\right],$$

where the nonlinearities act coordinatewise. Let $\mathbf{u} = d^{-1/2}\mathbf{R}_Y(1, \dots, 1)^\top$, so $\|\mathbf{u}\| = 1$. The source transition is $P_X(\mathbf{x}, \cdot) = N(\mathbf{f}(\mathbf{x}), 0.6^2\mathbf{I})$, and

$$P_Y(\mathbf{y}, \cdot) = \tfrac{1}{2}N(\mathbf{g}(\mathbf{y}) - 0.8\sqrt{d}\,\mathbf{u}, 0.25^2\mathbf{I}) + \tfrac{1}{2}N(\mathbf{g}(\mathbf{y}) + 0.8\sqrt{d}\,\mathbf{u}, 0.25^2\mathbf{I}).$$

There is one mixture sign per noise vector, giving two multivariate modes. The rotations couple the observed coordinates in both conditional means.

Let F_{N} be the standard normal cumulative distribution function and H_d that of $\frac{1}{2}N(-0.8\sqrt{d}, 0.25^2) + \frac{1}{2}N(0.8\sqrt{d}, 0.25^2)$. For a supplied source next state $\mathbf{z}$, set $\mathbf{w} = (\mathbf{z} - \mathbf{f}(\mathbf{x}))/0.6$ and $s = \mathbf{u}^\top\mathbf{w}$. The oracle is

$$\Phi^{\mathrm{or}}_{\mathbf{x},\mathbf{y}}(\mathbf{z}) = \mathbf{g}(\mathbf{y}) + 0.25(\mathbf{w} - s\mathbf{u}) + \mathbf{u}H_d^{-1}(F_{\mathrm{N}}(s)). \tag{63}$$

Under the source transition, $s \sim N(0,1)$ independently of $\mathbf{w} - s\mathbf{u} \sim N(0, \mathbf{I} - \mathbf{u}\mathbf{u}^\top)$. The quantile map gives the required mixture along $\mathbf{u}$, while the orthogonal component supplies independent Gaussian noise, yielding exactly $P_Y(\mathbf{y}, \cdot)$. For a unit evaluation direction $\mathbf{a}$, the projected target law is a mixture with means $\mathbf{a}^\top\mathbf{g}(\mathbf{y}) \pm 0.8\sqrt{d}\,\mathbf{a}^\top\mathbf{u}$ and variance 0.25^2 in each component; its quantiles enter (61).

D.3 Results at dimension 100

We repeat both sample-size experiments at $d = 100$ with width 256 and all other training and evaluation settings unchanged. At $m = 8192$, the mean next-state errors (standard errors in parentheses) are

Benchmark	Conditional flow	Without $\mathbf{x}_i$	Marginal flow	Oracle
Gaussian	0.733 (0.015)	1.043 (0.024)	1.032 (0.016)	0.070 (0.001)
Nonlinear mixture	0.565 (0.027)	0.666 (0.021)	0.829 (0.009)	0.129 (0.006)

At this sample size, the full conditional flow has lower mean error than both learned baselines in both benchmarks. Its improvement over the source-state ablation is smaller for nonlinear transitions. Errors remain above the oracle, and the nonlinear ablation performs better at some smaller sample sizes under the fixed training budget.

E TEN-SUBJECT EEG STUDY

E.1 PROTOCOL

We evaluate all ten THINGS-EEG2 subjects using the released ENIGMA implementation (Kneeland et al., 2026; Alljoined, 2025) under the matched subject-adaptation protocol in Section 5. ENIGMA uses subject-specific alignment and fine-tuning on calibration data to adapt to differences between subjects (Kneeland et al., 2026, Sections 3.3 and 4.3). Other EEG retrieval methods address these differences through label-free calibration or coordinate alignment (Huang & Zhu, 2026; Cui et al., 2026). Each recording contains 250 time samples from 63 channels over one second following image presentation. ENIGMA maps recordings to image embeddings and ranks candidate images by embedding similarity. Split seed 20260904 selects 4,096 calibration and 1,000 validation images. Evaluation averages 80 EEG repetitions per image in the official 200-image gallery. Baseline and flow use identical splits and seeds. This 63-channel adaptation protocol differs from ENIGMA's published 32-channel joint-training experiment (Kneeland et al., 2026, Sections 3.3–4.1).

For adapted subject s, ENIGMA is pretrained on reference subjects $\mathcal{S}_s$: subjects 5–10 for $s \in \{1,2,3,4\}$; subjects $\{1,2,3,4,7,8,9,10\}$ for $s \in \{5,6\}$; and subjects 1–6 for $s \in \{7,8,9,10\}$. The baseline and both flow widths share this pretrained model, whose encoder and projection network remain frozen. We average seeds 0, 1, and 42 within each subject. Confidence intervals are percentiles 2.5 and 97.5 over 100,000 paired bootstrap resamples of ten subjects. For Figure 2c, each resample recomputes the mean CFM and baseline accuracies $\bar{a}$ and $\bar{b}$ and their relative lift $100(\bar{a}/\bar{b}-1)\%$; for the absolute gains below, it computes their difference. All configurations use the final training update; no test value selects a checkpoint, width, or context length.

E.2 ADAPTER CONFIGURATIONS

Inputs and outputs. Let $\mathbf{E}_s(u) \in \mathbb{R}^{63\times 250}$ be subject s's EEG epoch for image u, $F_{\boldsymbol{\theta}}$ the flow map with learned parameters $\boldsymbol{\theta}$, and D the frozen ENIGMA encoder and projection network. The flow precedes the subject-specific temporal map $\mathbf{W}_s$; its endpoints and predicted embedding are

$$\mathbf{A}(u) = \mathbf{E}_s(u), \qquad \mathbf{B}_q(u) = \mathbf{E}_q(u) \quad (q \in \mathcal{S}_s), \qquad D\left(\mathbf{W}_s\left(F_{\boldsymbol{\theta}}(\mathbf{A}(u))\right)\right). \tag{64}$$

The map $\mathbf{W}_s$ has a 250×250 weight matrix and 250-dimensional bias shared across channels, initialized by averaging the reference subjects' pretrained maps. We train one flow and one map for each adapted subject, seed, width, and context length. The flow outputs a 63×250 epoch; D maps the adapted epoch to an image embedding.

Past-state conditioning. For any epoch $\mathbf{E}$ with columns $\mathbf{E}_j \in \mathbb{R}^{63}$, define the ordered history vector $\mathbf{h}_k(\mathbf{E}, j) = (\mathbf{E}_{j-1}, \ldots, \mathbf{E}_{j-32})$, with each slot replaced by zero if its lag exceeds k or its time index is at most zero. Thus $k \in \{0, 1, 4, 16, 32\}$ specifies the number of previous samples available, while the input dimension stays fixed. The velocity $\mathbf{v}_{\boldsymbol{\theta}}(t, \mathbf{z}; \mathbf{h}_k(\mathbf{A}, j), \mathbf{h}_k(\mathbf{B}_q, j)) \in \mathbb{R}^{63}$ receives flow time t, the current interpolation $\mathbf{z}$, and both histories. Its implementation concatenates one zero 63-channel slot, the source history, $\mathbf{z}$, the target history, and t, giving 4,159 input coordinates. The current source endpoint is used to form the interpolation but is not separately supplied to the velocity. A network shared across all EEG samples has two hidden GELU layers, each of width 64 for our main model or 14 for the smaller-model comparison, and a 63-dimensional linear output; all layers have biases. There is no normalization, dropout, skip, or recurrent layer. Zero output-layer initialization makes the initial flow the identity.

One source and multiple reference targets. For a batch of $N = 64$ distinct calibration images u_b, write $\mathbf{A}_b = \mathbf{A}(u_b)$ and $\mathbf{B}_{bq} = \mathbf{B}_q(u_b)$, with columns $\mathbf{a}_{bj}$ and $\mathbf{b}_{bqj}$. Pair each source epoch separately with every $q \in \mathcal{S}_s$. Draw independent $t_{bq} \sim \mathcal{U}(0,1)$, shared across the 250 samples of that pair, and set $\mathbf{z}_{bqj} = (1-t_{bq})\mathbf{a}_{bj} + t_{bq}\mathbf{b}_{bqj}$ and $\mathbf{r}_{bqj} = \mathbf{b}_{bqj} - \mathbf{a}_{bj}$.

Let $\widehat{L}_{\mathrm{ENIGMA}}$ denote the supervised embedding and contrastive objective described in the optimization paragraph below. We add the flow-matching loss $\widehat{L}_{\mathrm{FM}}$ to obtain

$$\widehat{L}_{\mathrm{EEG}} = \widehat{L}_{\mathrm{ENIGMA}} + 0.02\widehat{L}_{\mathrm{FM}},$$
$$\widehat{L}_{\mathrm{FM}} = \frac{1}{N|\mathcal{S}_s|\,63 \cdot 250} \sum_{b=1}^{N} \sum_{q \in \mathcal{S}_s} \sum_{j=1}^{250} \left\| \mathbf{v}_{\boldsymbol{\theta}} \left(t_{bq}, \mathbf{z}_{bqj}; \mathbf{h}_k(\mathbf{A}_b, j), \mathbf{h}_k(\mathbf{B}_{bq}, j) \right) - \mathbf{r}_{bqj} \right\|^2 . \tag{65}$$

Reference subjects receive equal weight, with one loss term for each source–reference pair. The same velocity network handles all references, without a reference-subject identity input. Training conditions on the observed reference history $\mathbf{h}_k(\mathbf{B}_{bq}, j)$, held fixed as flow time varies.

Sequential generation. For a source epoch $\mathbf{A}$ with columns $\mathbf{a}_j$, generate columns $\widehat{\mathbf{b}}_j$ of $\widehat{\mathbf{B}} = F_{\boldsymbol{\theta}}(\mathbf{A})$ in order $j = 1, \ldots, 250$ by approximately solving

$$\mathbf{z}_j(0) = \mathbf{a}_j, \qquad \frac{\mathrm{d}\mathbf{z}_j(t)}{\mathrm{d}t} = \mathbf{v}_{\boldsymbol{\theta}} \left(t, \mathbf{z}_j(t); \mathbf{h}_k(\mathbf{A}, j), \mathbf{h}_k(\widehat{\mathbf{B}}, j) \right), \qquad \widehat{\mathbf{b}}_j = \mathbf{z}_j(1). \tag{66}$$

Two explicit-midpoint steps of size $1/2$ integrate each local ODE, holding both histories fixed. Target history contains only previously generated samples; at $j = 1$, both histories are zero. Thus the initial sample uses a learned marginal flow, and generation requires no reference recording or reference-subject choice. The supervised loss $\widehat{L}_{\mathrm{ENIGMA}}$ differentiates through the full sequential generation and $\mathbf{W}_s$; D remains frozen. For $k = 1$ and $j > 1$, the velocity inputs and source initialization follow our theoretical transition construction. Here training pairs recordings of the same image, and generation uses a learned flow for the initial sample. The theory assumes independent source and target training trajectories and stationary target initialization. We do not assume that the EEG data satisfy its stationary Markov or mixing hypotheses.

Optimization. We retain ENIGMA's mean squared error over image-embedding coordinates and its symmetric image and text contrastive losses, weighted 1, 0.5, and 0.05, respectively (Alljoined, 2025). Contrastive scores use normalized target embeddings, unnormalized predictions, and the fixed pretrained score multiplier; the squared loss uses unnormalized image embeddings. The baseline trains only $\mathbf{W}_s$; our method jointly trains $\mathbf{W}_s$ and the flow. Both use 300 AdamW updates, peak learning rate $3 \cdot 10^{-4}$, weight decay 0.01 excluding biases and normalization parameters, gradient clipping at norm 1, and a one-cycle schedule with 10% warmup and cosine decay. Each update draws a fresh image batch and fresh flow times.

E.3 Additional results

Table 1 reports the $k = 32$ comparison with the flow before $\mathbf{W}_s$. Our main width-64 model achieves 49.05% mean Top-5 accuracy, versus 43.87% for the baseline: a gain of 5.18 percentage points (95% interval $[4.08, 6.08]$), or 11.82% relative lift ($[9.25, 14.06]$%). The smaller width-14 model reaches 46.73%, a gain of 2.87 points ($[2.13, 3.53]$), or 6.53% relative lift ($[4.81, 8.22]$%). Both improve on the baseline for all ten subjects after averaging seeds. Figure 2c reports every context length under the same protocol. Longer history improves retrieval at both widths without increasing model size within either sweep.

Table 1: Retrieval accuracy (%) at $k = 32$, averaging seeds within subjects and then ten subjects. Both CFM networks precede $\mathbf{W}_s$.

Adapter	Top-1	Top-5	Top-10
$\mathbf{W}_s$ only	17.78	43.87	57.40
CFM, width 64	20.55	49.05	62.40
CFM, width 14	18.60	46.73	59.82

E.4 Flow after the temporal map

We also place the flow after $\mathbf{W}_s$, giving the predicted embedding $D(F_{\boldsymbol{\theta}}(\mathbf{W}_s(\mathbf{E}_s(u))))$. The subjects, reference sets, data splits, seeds, architectures, optimization settings, and final checkpoint rule

are unchanged. The source and reference endpoints are now $\mathbf{A}(u) = \mathbf{W}_s(\mathbf{E}_s(u))$ and $\mathbf{B}_q(u) = \mathbf{W}_q(\mathbf{E}_q(u))$. Each reference map $\mathbf{W}_q$ is frozen at its pretrained value; $\mathbf{W}_s$ and $F_{\boldsymbol{\theta}}$ train jointly, with gradients through $\mathbf{W}_s$ in both the retrieval and flow-matching losses. The velocity inputs, loss (65), and sequential generation (66) use these transformed endpoints. Thus k counts preceding samples in the transformed sequence. Since the temporal maps mix all 250 raw time samples, this history need not correspond to preceding measurements in raw EEG time.

Figure 3 reports all five context lengths for both widths. Relative Top-5 lift increases from 3.91% at $k = 0$ to 11.63% at $k = 32$ for width 64, and from 1.44% to 7.26% for width 14. At $k = 32$, the width-64 model achieves 48.97% Top-5 accuracy, a gain of 5.10 percentage points over the 43.87% baseline (95% interval $[4.05, 6.28]$), or 11.63% relative lift ($[8.83, 15.00]\%$). Width 14 reaches 47.05%, a gain of 3.18 points ($[2.10, 4.33]$), or 7.26% relative lift ($[4.61, 10.37]\%$). Both improve on the baseline for all ten subjects after averaging seeds. These results show that the benefit of additional context also appears when the flow acts on temporally transformed EEG. Table 2 gives the corresponding Top-1, Top-5, and Top-10 accuracies.

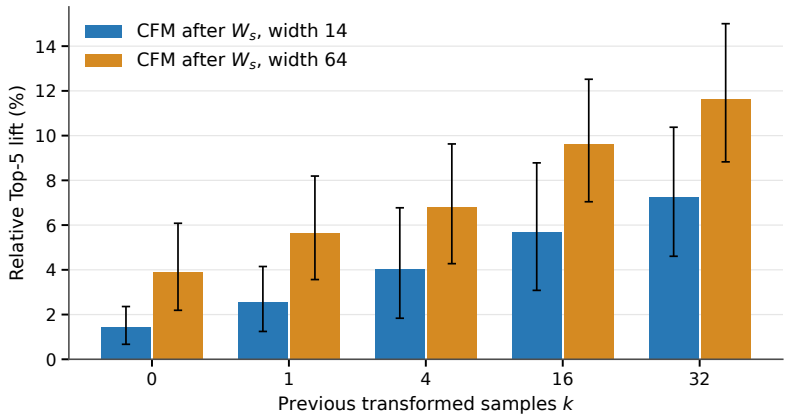


Figure 3: Additional EEG comparison with both flow networks after $\mathbf{W}_s$. Relative Top-5 retrieval accuracy lift over the matched $\mathbf{W}_s$-only baseline is plotted against context length k in transformed EEG coordinates. Accuracies average three seeds within each of ten subjects; error bars give 95% paired subject-bootstrap percentile intervals.

Table 2: Retrieval accuracy (%) at $k = 32$ with CFM after $\mathbf{W}_s$, averaging three seeds within subjects and then ten subjects.

Adapter	Top-1	Top-5	Top-10
$\mathbf{W}_s$ only	17.78	43.87	57.40
CFM after $\mathbf{W}_s$, width 64	20.65	48.97	62.23
CFM after $\mathbf{W}_s$, width 14	19.38	47.05	60.32

F Further comparisons with prior work

F.1 Sequential transport and generation

Time-series adaptation methods align observations, time axes, or representations across domains (Painblanc et al., 2023; Mishra et al., 2026). Flow models generate entire functions or correlated sequences (Kerrigan et al., 2024b; Wei & Ma, 2025); Schrödinger-bridge methods learn dynamics between endpoint laws or complete time-series laws (Tong et al., 2024; Hamdouche et al., 2026). COT-GAN uses causal transport as a training objective for sequential data (Xu et al., 2020). We instead transform a prescribed fresh source trajectory by recursively applying a learned conditional transition map.

Causal transport requires $\mathrm{Law}(\mathbf{y}_{0:i} \mid \mathbf{x}_{0:n}) = \mathrm{Law}(\mathbf{y}_{0:i} \mid \mathbf{x}_{0:i})$; bicausal transport imposes the reverse identity as well (Backhoff et al., 2017; Moulos, 2021). For Markov chains, Moulos (2021) compute bicausal transport costs by value iteration, and O'Connor et al. (2022) optimize stationary transition couplings for long-run average cost. O'Connor et al. (2024) give finite-sample cost bounds for estimated stationary couplings of finite-valued processes under mixing assumptions. Our statistical bound controls the squared prediction error of a continuous conditional velocity; (49) converts it into accuracy of generated states.

F.2 Statistical error and dependence

Benton et al. (2024, Theorem 1) and Gao et al. (2024, Proposition 6.2) convert population velocity error into Wasserstein error. For stochastic diffusion flow matching, Silveri et al. (2024, Theorems 2–3) bound Kullback–Leibler error; Bansal et al. (2026, Theorems 1–2) analyze rectified-flow discretization. We combine velocity estimation from dependent transitions with conditional Wasserstein bounds and recursive generation. Our estimation proof uses the coupling and fourth-moment approach to dependent square-loss regression, as in Roy et al. (2021, Theorem 3.1).

For convex or realizable predictor classes, Ziemann et al. (2024, Theorem 3.1 and discussion following (5)) show that, under suitable moment assumptions, the leading square-loss error term need not contain a mixing-time factor. Dependence enters the minimum sample size and higher-order terms. Our lower bound is consistent with this distinction: it establishes a mixing-dependent sample threshold.

F.3 Distribution mismatch and rollout stability

In offline policy evaluation, Zanette & Wainwright (2022, Section 4.3, Lemma 2, (16b)) compare squared Bellman-residual norms under the target-policy occupancy law d^π and the data law μ. If d^π has density at most M relative to μ, then every f with $\mathbb{E}_{(s,a)\sim\mu} f(s,a)^2 < \infty$ satisfies

$$\mathbb{E}_{(s,a)\sim d^\pi} f(s,a)^2 \le M \mathbb{E}_{(s,a)\sim\mu} f(s,a)^2,$$

where (s,a) is a state–action pair. Their application takes f to be a Bellman residual. Duan et al. (2020, Section 1) similarly restrict distribution mismatch to a function class through a restricted χ^2 divergence. The concentrability condition (41) compares our conditional velocity error under the law $\Gamma_i^{\mathbf{v}}$ of source and generated state pairs and the training law $\pi_X \otimes \pi_Y$, with the conditional risk $\ell_{\mathbf{v}}$ from (40) playing the role of the squared Bellman residual.

Rudolf & Schweizer (2018, Section 3.1, Theorem 3.1) propagate one-transition discrepancies through contracting Markov dynamics in first-order Wasserstein distance. Proposition 5 gives a conditional W_2 coupling that retains the prescribed source transitions and controls both state and joint-trajectory error.