TRACE: Trajectory Recovery for Continuous Mechanism Evolution in Causal Representation Learning

Shicheng Fan[1] Kun Zhang[2,3] Lu Cheng[1]

[1]University of Illinois Chicago [2]Carnegie Mellon University [3]MBZUAI

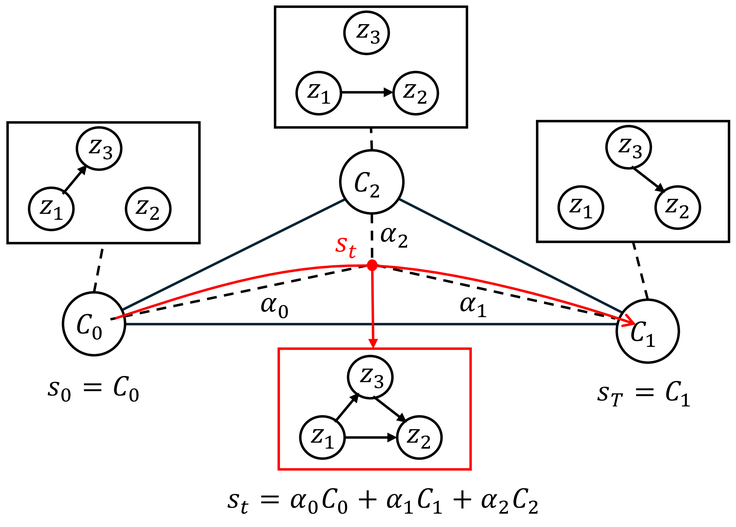
LinkedIn WeChat Paper & Code

ICML International Conference On Machine Learning

The Problem: Mechanisms Evolve *Continuously*

Temporal causal representation learning (CRL) assumes causal mechanisms switch **instantaneously between discrete domains**. Real systems evolve **smoothly**: a vehicle's dynamics shift gradually through a turning maneuver, and human gait reorganizes continuously from walking to running.

Our view. A transitional mechanism is a *convex combination* of finitely many *atomic mechanisms* $\{C_0, \ldots, C_{K-1}\}$; the state s_t is a point that traces a **continuous trajectory through a simplex** of canonical causal graphs.



$$s_t = \alpha_0 C_0 + \alpha_1 C_1 + \alpha_2 C_2$$

Figure 1. Continuous mechanism transitions as a trajectory through the simplex of atomic mechanisms. Each vertex is a canonical causal graph over z_1, z_2, z_3; any interior point $s_t = \sum_k \alpha_k(t) C_k$ is a mixture. Discrete-switching models can only jump between vertices.

Formulation

Latent dynamics $\mathbf{z}_t = f(\mathbf{z}_{t-L:t-1}; \theta_t) + \boldsymbol{\epsilon}_t$, $\mathbf{x}_t = g(\mathbf{z}_t)$. Rather than switching θ_t *discretely* (prior CRL), we let it evolve *continuously*: $\theta_t = \sum_k \alpha_k(t)\theta^{(k)}$, $\boldsymbol{\alpha}(t) \in \Delta^{K-1}$; with transition matrices $W^{(k)}$, the effective $W(t) = \sum_k \alpha_k(t) W^{(k)}$. **Asm. 3.1 (distinguishable):** $\{W^{(k)} - W^{(0)}\}$ linearly independent $\Rightarrow K \leq d+1$.

Goal. Jointly identify (i) the latent causal variables and (ii) the *continuous mixing trajectory* $\boldsymbol{\alpha}(t)$, including intermediate mechanism states **never observed during training**.

Key Contributions

1. **Problem.** Formalize CRL with *continuously* evolving mechanisms as a trajectory through a simplex of atomic mechanisms.
2. **Method. TRACE**: a Mixture-of-Experts where each expert learns one atomic mechanism, enabling test-time recovery.
3. **Theory.** Latents *and* mixing trajectory $\boldsymbol{\alpha}(t)$ are jointly identifiable, with finite-sample bounds.
4. **Results.** Up to **0.99** weight correlation, **3–4**× better than discrete baselines.

References

[1] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[2] W. Yao, G. Chen, and K. Zhang, "Temporally disentangled representation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[3] X. Song, W. Yao, Y. Fan, X. Dong, G. Chen, J. C. Niebles, E. Xing, and K. Zhang, "Temporally disentangled representation learning under unknown nonstationarity," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8092–8113, 2023.

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[5] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2017.

[6] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, "Variational autoencoders and nonlinear ICA: A unifying framework," in *International conference on artificial intelligence and statistics*, pp. 2207–2217, PMLR, 2020.

[7] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," in *International Conference on Learning Representations*, 2017.

Method: TRACE (two stages)

Stage 1: MoE representation learning

A shared encoder maps observations to latents; one expert per domain learns one atomic mechanism (normalizing-flow inverse transition), routed by one-hot gating during training.

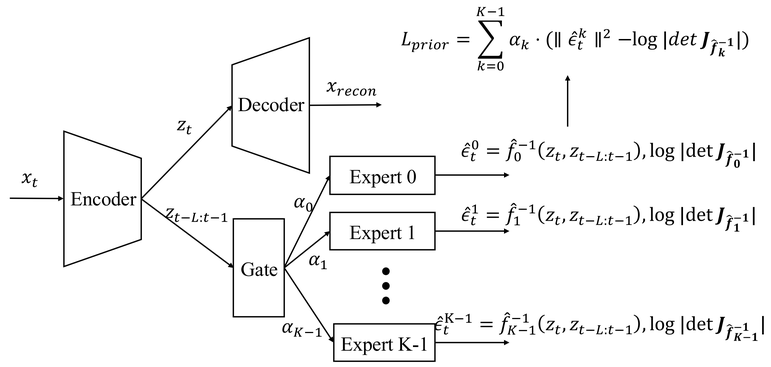


Figure 2. Stage 1 architecture: a shared encoder with domain-specific experts; one-hot gating routes each pure-domain sample to its expert, so each expert specializes in one atomic mechanism.

Stage 2: mechanism trajectory inference

Recover the mixing coefficients by *least-squares projection onto the simplex* (basis $\hat{B}$ of baseline shifts $\delta\hat{\boldsymbol{\mu}}^{(k)}$), then temporally average, with no labels or intermediate-state ground truth needed at test time:

$$\hat{\boldsymbol{\alpha}}(t) = \operatorname*{Proj}_{\Delta^{K-1}} \left(\hat{B}^{\dagger}(\mathbf{z}_t - \hat{\boldsymbol{\mu}}^{(0)}) \right).$$

Identifiability Theory

Thm 4.1: Latent variables

Under sufficient variability, any learned representation satisfies $\hat{z}_i = h_i(z_{\pi(i)})$ for a permutation π and strictly monotonic h_i (recovery up to permutation & component-wise transform).

Thm 4.2: Pointwise trajectory recovery

The least-squares estimator obeys

$$\|\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}(t)\| \leq \tfrac{1}{\sigma_{\min}} \big(\|\hat{\boldsymbol{\epsilon}}_t\| + \delta_{\text{approx}} \big),$$

controlled by the basis conditioning $\sigma_{\min}(\hat{B})$.

Thm 4.3: Smooth trajectory recovery

For bounded total variation $\mathrm{TV}(\boldsymbol{\alpha}^*) \leq V$, the smoothness-regularized estimator attains the minimax-optimal rate $\frac{1}{T}\sum_t \mathbb{E}\|\hat{\boldsymbol{\alpha}}_t^{\text{smooth}} - \boldsymbol{\alpha}_t^*\|^2 = O(T^{-2/3})$. *Capacity*: $K \leq d+1$ is **sufficient, not necessary**; recovery holds well beyond it when active mechanisms are few.

Does It Hold on Real Data?

Reconstruction alone is unidentifiable, so we make the bound *checkable, not asymptotic*: every quantity is estimable from data, and the effective signal-to-noise ratio

$$\mathrm{SNR}_{\text{eff}} = \frac{\sigma_{\min}(\hat{B})}{\bar{\epsilon} + \delta_{\text{approx}}}$$

predicts recovery *before* deployment. The basis conditioning $\sigma_{\min}(\hat{B})$ is a **pre-deployment diagnostic** of mechanism separability (mechanism-impurity stress test):

Impurity ε	Weight Corr	$\sigma_{\min}(\hat{B})$
0%	0.996	0.168
20%	0.981	0.126
50%	0.936	0.049

Recovery degrades exactly as the $O(1/\sigma_{\min})$ bound predicts. With no ground-truth latents, real data is validated via Pearson correlation, invariant to the monotonic ambiguity the theory allows.

Real-World: Vehicle Turning (UAVDT)

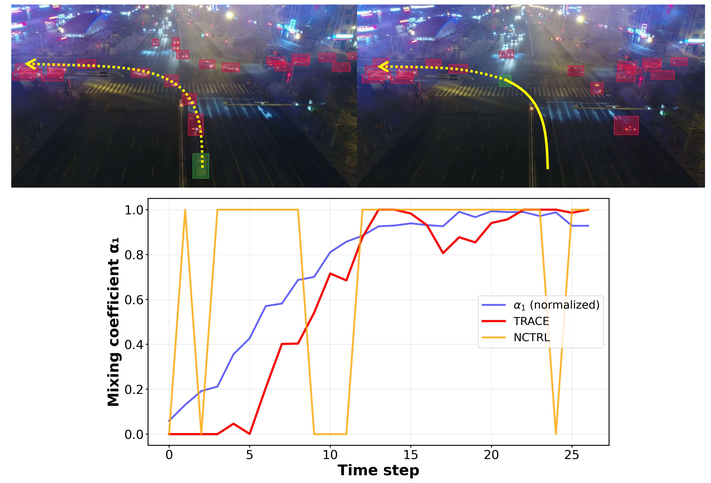


Figure 3. A vehicle executing a left turn (yellow = tracked trajectory). TRACE's α_1 rises *smoothly* from 0 to 1 (Corr = **0.96**), while NCTRL oscillates erratically between discrete states (Corr = 0.24).

Real-World: Human Gait (CMU MoCap)

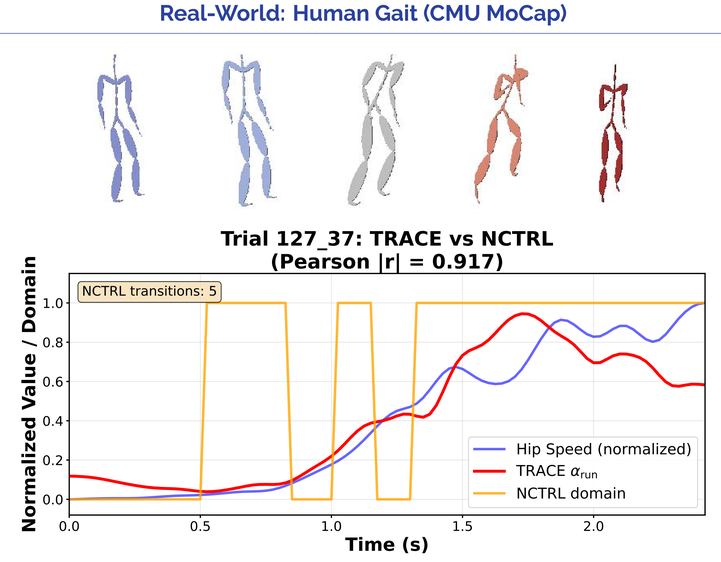


Figure 4. Walk→run gait (skeletons: blue → red over time). TRACE's α_{run} tracks the hip-speed proxy *smoothly* (Corr = 0.917); NCTRL fires **5 spurious** transitions (CartPole pixels: MCC 0.97).

Sample Efficiency

Temporal smoothing (Thm 4.3) keeps trajectory recovery robust as training data shrinks (d=8, K_{total}=5, length-50 trajectories):

Train traj./domain	Latent MCC	Weight Corr (K=3)
40,000	0.963	0.986
8,000	0.886	0.968
4,000	0.752	**0.960**
400	0.622	0.688

With 10× **less** data (4,000/domain), per-step MCC falls to 0.75 yet the recovered trajectory holds at Corr **0.96**.