IDENTIFIABILITY OF LATENT STATE DYNAMICS UNDER ADAPTIVE MULTI-ENVIRONMENT SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent state models underpin many machine learning systems, but in practice data is rarely collected i.i.d. and is instead generated by an adaptive mechanism that depends on the latent state itself. This introduces a form of hidden confounding that standard causal representation learning (CRL) methods do not address. We propose the Latent Adaptive Selection Model (LASM), a structural causal framework for modeling latent dynamics under adaptive data collection. Unlike prior CRL settings with exogenous interventions, LASM captures endogenous selection, where policies both depend on and distort the latent state. Our main contributions are: (i) a geometric characterization of non-identifiable directions induced by adaptive policies via the Policy Selection Gradient Matrix; (ii) a sharp identifiability result showing that, under a logged-policy anchor and a policy diversity condition, the latent transition matrix is exactly recoverable across environments; (iii) a tight lower bound $E^\star = \lceil d/(k-1) \rceil$ on the number of environments required for identifiability; and (iv) the Multi-environment Adaptive Invariant Latent Estimator algorithm, which enforces latent frame anchoring via a logged-policy consistency objective and promotes diversity through a log-determinant regularizer. Experiments on synthetic data verify the predicted identifiability threshold, while real-world experiments on ASSISTments demonstrate that recovered skill-transition structures are highly reproducible across schools ($\rho = 0.91$) only under multi-environment pooling, and collapse in single-environment settings.

1 INTRODUCTION

In intelligent systems, inferring a latent state that is never directly observed is important, as we can only access it through noisy observations generated under state-dependent data collection (Xu et al., 2025). This arises in clinical decision-making, recommender systems, and intelligent tutoring (Liang et al., 2026), where actions, e.g., tests, recommendations, exercises, are chosen based on an estimate of the latent state.

A major problem is that the observation process is adaptive: the action A_t depends on the latent state Z_t, making the data distribution a policy-distorted view of the underlying dynamics. Consequently, observed correlations may reflect the collection mechanism rather than the true latent causal structure, and parts of the latent space may be under-observed.

Existing causal representation learning (CRL) approaches achieve identifiability by leveraging exogenous environmental variation, where an external variable injects independent changes into the latent process (Hyvärinen et al., 2019; Khemakhem et al., 2020; Ahuja et al., 2022; Brehmer et al., 2022; Lachapelle et al., 2023). In contrast, adaptive data collection introduces an endogenous selection mechanism: actions come from the latent state and simultaneously influence observations. Rather than injecting variation, the policy removes variation along state-dependent directions. As a result, standard multi-environment CRL arguments do not work here.

Even with a known policy, identifiability remains challenging. A single adaptive environment entangles causal transitions with policy-induced correlations: along directions where the policy is insensitive, observational effects of selection are indistinguishable from true latent dynamics. Thus, disentangling causality from selection bias is impossible from one environment alone.

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.

Our key insight is that multiple adaptive policies resolve this problem by providing complementary views of the latent space. Each logged policy τ^e defines a Policy Selection Gradient Matrix (PSGM) Φ^e that constrains only a subspace of the latent variables. Identification is possible when these subspaces jointly span the latent space, i.e., $\bigcap_e \ker(\Phi^e) = \{0\}$. This yields an intrinsic requirement on the number of environments, $E^\star = \lceil d/(k-1) \rceil$, which we show is tight. Our contributions:

- **Modeling adaptive latent dynamics.** We propose the Latent Adaptive Selection Model (LASM), a structural causal model for latent Markov systems under state-dependent observation policies.
- **Fundamental limits of single-environment learning.** We characterize the exact confounded directions induced by adaptive policies via the PSGM and show that a single environment is insufficient for identifiability.
- **Multi-environment identifiability and tightness.** We prove that latent dynamics are exactly identifiable under a policy diversity condition across environments, and establish a tight lower bound $E^\star = \lceil d/(k-1) \rceil$.
- **Algorithms and alternative identification routes.** We propose a consistent Multi-environment Adaptive Invariant Latent Estimator (MAILE).

2 Related Work

Causal representation learning. CRL seeks to recover latent causal structure from high-dimensional observations (Schölkopf et al., 2021; Mameche et al., 2025). A major line of work achieves identifiability by introducing exogenous sources of variation, such as auxiliary variables in nonlinear ICA (Hyvärinen et al., 2019; Khemakhem et al., 2020), or multi-environment shifts (Ahuja et al., 2022; Brehmer et al., 2022; Lachapelle et al., 2023; Ahuja et al., 2024). These approaches assume that environments or auxiliary variables inject independent variation into the latent space, which is then exploited for identifiability. In contrast, our setting is endogenous: adaptive policies depend on the latent state and filter variation, yielding a distinct identifiability structure and a sharp environment threshold.

Identifiability of latent variable models. Classical identifiability results for latent variable models rely on structural assumptions such as non-Gaussianity linear ICA (Comon, 1994) or sparsity constraints (Xu et al., 2024). In linear ICA, independent non-Gaussian noise identifies latent factors up to a signed permutation (Eriksson & Koivunen, 2004). Chen et al. (2022) relax the non-Gaussianity requirement to only a subset of latent variables, broadening the applicability of ICA-style identification while still relying on distributional rather than multi-environment structure. Le et al. (2025) instead recover a structural causal model from observational data given a known causal ordering, using an invertible flow parameterization to obtain a non-iterative estimator. Our setting is complementary: identifiability arises from policy diversity across adaptive environments, yielding exact recovery together with a sharp lower bound on the number of environments.

Offline reinforcement learning and logged policies. Offline reinforcement learning studies learning or evaluating policies from logged data, typically using inverse propensity scoring or doubly robust estimators (Dudík et al., 2014; Wang et al., 2017). These methods correct for distribution shift to estimate quantities such as policy value or counterfactual rewards, and assume the latent state is either observed or not required. In contrast, we use the logged policy as a structural anchor for latent state identification rather than for reweighting or policy evaluation.

3 Problem Formulation

3.1 The Latent Adaptive Selection Model

Definition 1 (Latent Adaptive Selection Model (LASM))**.** A LASM is a tuple $\mathcal{M} = (\mathcal{Z}, \mathcal{A}, \mathcal{X}, f, g, \mu_0, \tau)$ where $\mathcal{Z} \subseteq \mathbb{R}^d$ is the latent state space, $\mathcal{A} = \{1, \dots, k\}$ is the finite action space of size k, $\mathcal{X} \subseteq \mathbb{R}^p$ is the observation space ($p \geq d$), $f : \mathcal{Z} \times \mathcal{A} \times \mathcal{E}_f \to \mathcal{Z}$ is the latent transition function, $g : \mathcal{Z} \times \mathcal{A} \times \mathcal{E}_g \to \mathcal{X}$ is the emission function, $\mu_0 \in \Delta(\mathcal{Z})$ is the initial state distribution, $\tau : \mathcal{Z} \to \Delta(\mathcal{A})$ is the adaptive selection policy, shown in Figure 1.

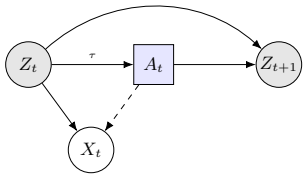


Figure 1: Causal graph of LASM. Gray: latent; white: observed; blue: adaptive action. The adaptive policy $Z_t \to A_t$ and transition $A_t \to Z_{t+1}$ induce confounding in latent dynamics, while the action-dependent emission edge $A_t \to X_t$ confounds measurement additionally.

The data-generating process at each time step t is:

$$A_t \sim \tau(Z_t), \tag{1}$$

$$X_t = g(Z_t, A_t, \epsilon_t^g), \tag{2}$$

$$Z_{t+1} = f(Z_t, A_t, \epsilon_t^f), \tag{3}$$

with $Z_1 \sim \mu_0$ and all noise terms i.i.d. and mutually independent.

The adaptive confounding structure. Eq. 1 is the defining feature: A_t is chosen as a function of the latent state Z_t, the same state whose dynamics we seek to recover. This couples the action to Z_t through two paths. Because A_t both depends on Z_t (through τ) and drives the next state Z_{t+1} (through f, Eq. 3), the policy opens a backdoor $Z_{t+1} \leftarrow Z_t \to A_t \to Z_{t+1}$ that confounds the latent transition (our identification target). When the emission is action dependent it additionally creates the measurement backdoor $X_t \leftarrow Z_t \to A_t \to X_t$, so $P(X_t \mid A_t)$ departs from $P(X_t \mid \mathrm{do}(A_t))$. Both would vanish if Z_t were observed, but it is seen only through the emission 2, and only up to a latent frame; as we show, the policy is invariant to exactly the frame directions it cannot detect, which is the precise source of non-identifiability.

3.2 Multi-Environment Setting

We consider E environments $\{e\}_{e=1}^E$ that share a common latent dynamical system (f, μ_0, g) but differ in their behavioral policies τ^e. Such environments include, e.g., hospitals, recommender systems, or tutoring platforms, where the same underlying causal process is observed under different intervention or decision rules.

Assumption 1 (Shared Markov Latent System)**.** All environments share a Markov latent process with initial distribution μ_0, transition kernel $Z_{t+1} \sim f(\cdot \mid Z_t, A_t)$, and emission model $X_t \sim g(\cdot \mid Z_t, A_t)$.

Assumption 2 (Linear-Gaussian Latent Dynamics)**.** The transition is linear-Gaussian:

$$Z_{t+1} = BZ_t + C_{A_t} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_f),\ \Sigma_f \succ 0, \tag{4}$$

where $B \in \mathbb{R}^{d\times d}$ and $C = [c_1, \ldots, c_k] \in \mathbb{R}^{d\times k}$.

Assumption 3 (Collective Observability)**.** Let $\mathcal{A}^+$ denote actions with positive probability. The stacked emission map $z \mapsto (g(z, a, \cdot))_{a\in\mathcal{A}^+}$ is injective in z, with Jacobian of full rank d on the support of Z (hence $p|\mathcal{A}^+| \geq d$).

Assumption 4 (Known Log-Linear Policies)**.** Each environment policy is a known softmax:

$$\tau^e(a \mid z) = \frac{\exp(w_a^{e\top} z + \beta_a^e)}{\sum_{a'} \exp(w_{a'}^{e\top} z + \beta_{a'}^e)}, \tag{5}$$

where $\{w_a^e, \beta_a^e\}$ are fully observed (logged).

Remark 1. The known policy assumption reflects logged decision systems in modern digital platforms, where the deployment rule is recorded alongside data. It also resolves the usual latent-variable identifiability ambiguity up to invertible reparameterization, since the policy couples observed actions to latent states in a fixed coordinate system. While we adopt a log-linear form for clarity and sharp identifiability results, our analysis depends only on the score geometry of τ^e, and extends to broader policy classes satisfying analogous spanning conditions.

3.3 IDENTIFIABILITY TARGET

Because any latent model is invariant under an invertible reparametrisation of $\mathcal{Z}$, the matrix B is only well defined relative to a choice of latent coordinates. We adopt the coordinates in which the logged policies take their known form (Assumption 4): the policy plays the role that known interventions or auxiliary variables play in CRL, supplying the external convention that fixes the frame. Our target is therefore B in the policy anchored frame. We stress that this does not trivialise the problem—τ^e is a known function of an unobserved state, the emission g is unknown and nonlinear, and each policy pins only the slice of the frame it responds to (§4), so spanning the full frame is an inherently multi-environment task.

Let μ_Z^e denote the marginal (stationary) distribution of Z_t under the LASM with policy τ^e.

Definition 2 (Identifiability up to policy symmetry)**.** Two LASMs $\mathcal{M}, \widetilde{\mathcal{M}}$ that share the same logged policies $\{\tau^e\}$ are observationally equivalent if they generate the same joint distribution $P^e(A_{1:T}, X_{1:T})$ for all e and all $T \geq 1$. We say B is identifiable up to the policy symmetry group

$$\mathcal{G} := \big\{M \in GL(d) : \tau^e(\cdot \mid Mz) = \tau^e(\cdot \mid z)\ \forall e, \forall z \in \mathcal{Z}\big\}, \tag{6}$$

if every $\widetilde{\mathcal{M}}$ observationally equivalent to $\mathcal{M}$ satisfies $\widetilde{B} = MBM^{-1}$ for some $M \in \mathcal{G}$. $\mathcal{G}$ is the stabiliser of the (known) policies under reparametrisation of $\mathcal{Z}$; when $\mathcal{G} = \{I\}$, B is exactly identified.

Remark 2. $\mathcal{G}$ is always at least a discrete group reflecting trivial relabellings; for log-linear policies the diversity condition forces $\mathcal{G} = \{I\}$. We work with conjugation MBM^{-1} rather than the $PBP^\top$ used in permutation-only settings, because a frame change $z \mapsto M^{-1}z$ acts on B by similarity, and $P^\top = P^{-1}$ only for orthogonal P.

4 NON-IDENTIFIABILITY UNDER A SINGLE ENVIRONMENT

4.1 POLICY SELECTION GRADIENT MATRIX

The central obstacle in adaptive data collection is that the policy does not respond equally to every latent direction. Some perturbations of the latent state substantially change the action probabilities, whereas others leave the policy almost unchanged. Only the former directions are observable through the logged policy and can therefore contribute to fixing the latent coordinate frame. This motivates a geometric measure of the policy's local sensitivity.

For an environment e, consider an infinitesimal perturbation $z \mapsto z + \varepsilon v$. A first-order Taylor expansion gives

$$\log \tau^e(a|z + \varepsilon v) = \log \tau^e(a|z) + \varepsilon \phi_a^e(z)^\top v + o(\varepsilon),$$

where

$$\phi_a^e(z) := \nabla_z \log \tau^e(a|z)$$

is the score of the policy with respect to the latent state. Hence, $\phi_a^e(z)^\top v$ measures the first-order sensitivity of the policy along direction v.

Definition 3 (Policy Selection Gradient Matrix)**.** The Policy Selection Gradient Matrix (PSGM) of environment e is $\Phi^e = \mathbb{E}_{Z\sim\mu_Z^e}\mathbb{E}_{A\sim\tau^e(\cdot|Z)}\left[\phi_A^e(Z)\phi_A^e(Z)^\top\right]$.

The PSGM measures the average first-order sensitivity of the logged policy to perturbations of the latent state. In particular, $v^\top\Phi^e v = \mathbb{E}[(\phi_A^e(Z)^\top v)^2]$, so directions with larger values are more strongly reflected in the observed actions, while directions in $\ker(\Phi^e)$ are locally invisible to policy.

Lemma 1. *For every environment e,*

$$\Phi^e \succeq 0.$$

Proof. For any $v \in \mathbb{R}^d$,

$$v^\top\Phi^e v = \mathbb{E}\left[(\phi_A^e(Z)^\top v)^2\right] \geq 0,$$

which proves positive semidefiniteness. □

For the log-linear policy,

$$\phi_a^e(z) = w_a^e - \sum_{a'} \tau^e(a'|z) w_{a'}^e,$$

so

$$\mathrm{range}(\Phi^e) = \mathrm{span}\{w_a^e - w_{a'}^e\},$$

and therefore

$$\mathrm{rank}(\Phi^e) \le k - 1.$$

$\ker(\Phi^e)$ contains exactly the latent directions that cannot be distinguished by the logged policy.

The kernel of the PSGM consists precisely of latent directions along which the logged policy is locally invariant, since

$$v^\top \Phi^e v = \mathbb{E}\left[(\phi_A^e(Z)^\top v)^2\right].$$

Thus,

$$v \in \ker(\Phi^e) \quad \Longleftrightarrow \quad \phi_A^e(Z)^\top v = 0 \text{ a.s.},$$

meaning that infinitesimal perturbations along v leave the policy unchanged to first order.

Definition 4. The confounding subspace of environment e is $\mathcal{C}^e = \ker(\Phi^e)$.

Perturbations within $\mathcal{C}^e$ cannot be distinguished from the logged policy and therefore remain unconstrained when recovering the latent coordinate frame.

4.2 Characterization of Single-Environment Non-identifiability

The confounding subspace characterizes exactly the latent directions along which the adaptive policy cannot distinguish different coordinate systems. Consequently, transformations confined to this subspace produce observationally equivalent latent models.

Proposition 1 (Single-environment non-identifiability)**.** *Suppose $\mathcal{C}^e = \ker(\Phi^e)$ is nontrivial. Then for every nonzero $v \in \mathcal{C}^e$ there exists a one-parameter family of invertible latent transformations $\{M_\lambda\}_\lambda$ such that*

$$\widetilde{B} = M_\lambda B M_\lambda^{-1},$$

while

$$P(A_{1:T}, X_{1:T}) = P_\lambda(A_{1:T}, X_{1:T})$$

for every finite horizon T. Hence the transition matrix B is not identifiable from a single adaptive environment. Moreover, every infinitesimal non-identifiable transformation is generated by a direction in $\ker(\Phi^e)$.

Proof sketch. Let $v \in \ker(\Phi^e)$ and define the infinitesimal coordinate transformation $M_\lambda = I + \lambda v u^\top$. The transformed latent process $\widetilde{Z} = M_\lambda Z$ induces the standard similarity transformation $\widetilde{B} = M_\lambda B M_\lambda^{-1}$ together with the corresponding transformed emission. The key observation is that $v \in \ker(\Phi^e)$ implies

$$(w_a^e - w_{a'}^e)^\top v = 0 \quad \forall a, a',$$

so

$$\tau^e(a \mid M_\lambda z) = \tau^e(a \mid z).$$

Hence the logged policy is invariant under the coordinate transformation. The transformed model therefore generates exactly the same joint distribution over actions and observations, while generally satisfying

$$M_\lambda B M_\lambda^{-1} \neq B.$$

Thus B is not identifiable. The full algebraic derivation is given in Appendix A. □

5 IDENTIFIABILITY FROM MULTIPLE ENVIRONMENTS

5.1 RESOLVING CONFOUNDING THROUGH POLICY DIVERSITY

Section 4 showed that a single adaptive environment leaves the latent directions in $\mathcal{C}^e = \ker(\Phi^e)$ unidentified. These directions correspond to transformations of the latent coordinate system that preserve both the logged policy and the observational distribution. The key question is therefore whether multiple environments can jointly eliminate this ambiguity. Suppose observations are collected under E environments with different adaptive policies. Each environment reveals only the latent directions lying in $\mathrm{range}(\Phi^e)$, while leaving the orthogonal directions $\ker(\Phi^e)$ confounded. Pooling environments progressively shrinks the set of directions that remain invisible to every policy. Consequently, the only remaining ambiguity is the intersection

$$\bigcap_{e=1}^{E} \ker(\Phi^e).$$

If this intersection is trivial, every latent direction is observed by at least one policy, and the latent coordinate frame becomes uniquely determined. This motivates the following diversity condition.

Definition 5 (Policy Diversity)**.** A collection of environments satisfies the policy diversity if $\bigcap_{e=1}^{E} \ker(\Phi^e) = \{0\}$. Equivalently, $\mathrm{rank}\left(\sum_{e=1}^{E} \Phi^e\right) = d$.

The equivalence follows because every Φ^e is positive semidefinite, implying

$$\ker\left(\sum_{e=1}^{E} \Phi^e\right) = \bigcap_{e=1}^{E} \ker(\Phi^e).$$

Unlike the confounding subspace of a single environment, which reflects the limitations of one adaptive policy, the policy diversity condition characterizes the collective information supplied by multiple policies. It requires only that every latent direction be observed by at least one environment; no individual environment needs to identify the entire latent space.

5.2 IDENTIFIABILITY VIA POLICY DIVERSITY

The policy diversity condition completely removes the single-environment ambiguity characterized in Proposition 1.

Theorem 1 (Multi-environment Identifiability)**.** *Under Assumptions 1–4, if the policy diversity condition holds, then the latent transition matrix B is exactly identifiable from the joint collection of multi-environment observations. Equivalently, the policy symmetry group reduces to $\mathcal{G} = \{I\}$, and every observationally equivalent LASM satisfies $\widetilde{B} = B$.*

Proof. See Appendix A for more details. □

5.3 MINIMUM NUMBER OF ENVIRONMENTS

The policy diversity condition immediately yields a lower bound on the number of environments required for identifiability.

Theorem 2 (Minimum Number of Environments)**.** *Suppose the action space has cardinality $k \geq 2$. Then at least $E^\star = \lceil d/(k-1) \rceil$ environments are necessary for identifiability. Moreover, this bound is tight.*

Proof. See Appendix A for more details. □

6 THE MAILE ALGORITHM

MAILE estimates $(B, \{c_a\}_{a=1}^{k}, g)$ from logged data using the known policies $\{\tau^e\}$ by minimising:

$$\mathcal{L}(\theta) = \sum_{e=1}^{E} \mathcal{L}^e_{\text{pred}}(\theta) + \beta\, \mathcal{L}_{\text{div}}(\theta). \tag{7}$$

The transition invariance (shared B across environments) is not a penalty term but a structural constraint: a single matrix B is shared across all environments, so the cross-environment M-step for B already enforces this constraint exactly.

Prediction loss. For each environment e:

$$\mathcal{L}^e_{\text{pred}} = \frac{1}{T_e}\sum_{t=1}^{T_e} \ell\Big(X^e_t,\, g(\hat{Z}^e_t, A^e_t)\Big) + \ell_\pi\Big(A^e_t,\, \tau^e(\cdot \mid \hat{Z}^e_t)\Big)\,, \tag{8}$$

where $\hat{Z}^e_t$ is the recovered latent representation (the top-d principal subspace of the pooled observations;), ℓ is the emission reconstruction loss, and ℓ_π is the negative log-likelihood of the observed action under the logged policy τ^e. This second term forces the recovered latent frame to be consistent with the known policy, which is what anchors the frame (absent it, $\hat{B}$ is determined only up to conjugation).

Diversity penalty. A finite-sample conditioner for the diversity condition:

$$\mathcal{L}_{\text{div}} = -\log\det\left(\sum_{e=1}^{E} \hat{\Phi}^e + \delta I\right), \tag{9}$$

where $\hat{\Phi}^e = \frac{1}{T_e}\sum_t \phi^e_{A^e_t}(\hat{Z}^e_t)\,\phi^e_{A^e_t}(\hat{Z}^e_t)^\top$ is the empirical PSGM, and $\delta > 0$ is a regularisation constant. The diversity condition is a property of the given environments, so $\mathcal{L}_{\text{div}}$ does not change whether identifiability holds; rather, maximising $\log\det(\sum_e \hat{\Phi}^e + \delta I)$ keeps the recovered latent coordinates aligned with the well-observed directions, improving the conditioning of the frame-anchoring step when the summed PSGM is near-singular. Setting $\beta = 0$ recovers a pure logged-policy estimator.

Algorithm 1 MAILE (linear-Gaussian closed-form solver)

Require: Data $\{(A^e_t, X^e_t)\}_{e,t}$; logged policies $\{\tau^e\}$ with weights $\{w^e_a\}$; conditioner $\delta > 0$.
Ensure: Estimated transition $\hat{B}$.

1. **Emission reconstruction.** Project the pooled $\{X^e_t\}$ onto their top-d principal subspace to obtain a representation $\hat{Z} = MZ$, recovering g up to the unknown frame M (minimises $\sum_e \ell$).
2. **Shared dynamics.** Solve the cross-environment pooled OLS $(\hat{B}^{\text{frame}}, \{c_a\}) \leftarrow \arg\min \sum_{e,t} \|\hat{Z}^e_{t+1} - B\hat{Z}^e_t - c_{A^e_t}\|_2^2$, yielding $\hat{B}^{\text{frame}} = MBM^{-1}$.
3. **Frame anchoring (logged-policy consistency).** Fit a softmax to $(A^e, \hat{Z}^e)$ to get $\hat{W}^e = W^e M^{-1}$, then solve the ridge-regularised least squares $\hat{M} \leftarrow \arg\min_M \sum_e \|\hat{W}^e M - W^e\|_F^2 + \delta\|M\|_F^2$ (minimises $\sum_e \ell_\pi$; the conditioner δ plays the role of $\mathcal{L}_{\text{div}}$).
4. **return** $\hat{B} \leftarrow \hat{M}^{-1}\,\hat{B}^{\text{frame}}\,\hat{M}$ in the anchored frame.

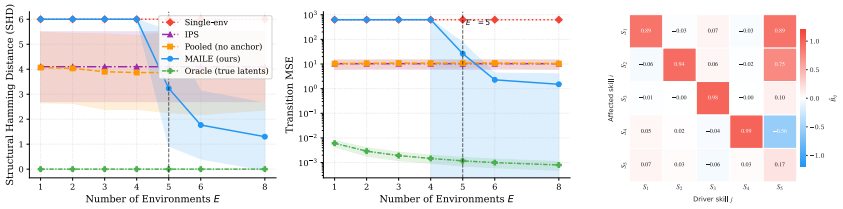
7 Experiments

We evaluate MAILE on both synthetic and real-world adaptive data to validate (i) the predicted identifiability threshold, and (ii) the role of multi-environment policy diversity.

7.1 Synthetic Validation

We generate a latent linear-Gaussian system with $d = 5$, $p = 16$, binary actions ($k = 2$), sparse transitions, injective emissions, and independently sampled logged policies. We vary the number of environments $E \in \{1, 2, 3, 4, 5, 6, 8\}$, generate 2000 transitions per environment, and compare MAILE against Single-env, IPS, Pooled (no anchor), and Oracle using SHD and transition MSE.

Verifying the Identifiability Threshold. Figure 2(a) evaluates recovery as the number of environments increases. Consistent with Theorem 2, MAILE exhibits a sharp phase transition at the predicted threshold $E^\star = 5$. When $E < E^\star$, recovery fails because the policy diversity condition is violated. At $E = E^\star$, estimation becomes feasible but remains sensitive to finite sample variability, leading to increased variance across random seeds. Once $E > E^\star$, both SHD and transition



a) Structure Recovery vs. Number of Environments (d = 5)

b) Recovered transition matrix

Figure 2: (a) Structure recovery versus the number of environments. The dashed line indicates the theoretical threshold $E^\star = \lceil d/(k-1) \rceil$. (b) Skill-transition matrix $\hat{B}$ recovered by MAILE from ASSISTments 2009, pooling nine schools (Elo ability observation, action = practised skill, per-school estimated policy; $d = 5$; $S1 = Eq.Solve \leq 2$, $S2 = Frac/Dec/\%\ Conv.$, $S3 = Add/SubInt.$, $S4 = Add/SubFrac.$, $S5 = Eq.Solve > 2$).

Table 1: Structure recovery on the synthetic benchmark.

Method	**Envs**	**SHD** $\downarrow$	**Trans. MSE** $\downarrow$
Single-env	1	6.0 ± 0.0	627.8 ± 17.1
IPS	1	4.1 ± 1.4	10.3 ± 4.2
Pooled (no anchor)	$E^\star$	3.9 ± 1.6	10.9 ± 4.0
MAILE	$E^\star$	3.2 ± 2.3	26.3 ± 48.7
MAILE	$E^\star{+}1$	1.8 ± 1.4	2.3 ± 2.0
MAILE	$E^\star{+}3$	$\mathbf{1.3 \pm 1.3}$	$\mathbf{1.5 \pm 2.5}$
Oracle	$E^\star$	0.0 ± 0.0	0.001 ± 0.001

MSE improve rapidly and approach the oracle estimator. In contrast, Single-env, IPS, and Pooled (no anchor) show little benefit from additional environments, demonstrating that multi-environment pooling alone is insufficient without logged-policy frame anchoring. Across $d \in \{3, 5, 8, 10\}$, the empirical threshold exactly matches $E^\star = \lceil d/(k-1) \rceil$.

Comparison with Baselines. Table 1 compares MAILE against competing estimators at and beyond the identifiability threshold. Single-environment training completely fails, while IPS substantially reduces prediction bias but cannot recover the latent transition matrix because the latent coordinate frame remains unidentified. Pooled (no anchor) also performs poorly despite using all environments, confirming that environment pooling alone does not resolve the latent ambiguity. In contrast, MAILE consistently improves once the policy diversity condition is satisfied, approaching oracle performance as the number of environments increases.

Overall, the synthetic experiments validate both theoretical claims of the paper: identifiability emerges precisely at the predicted environment threshold, and exploiting policy diversity through logged-policy anchoring substantially improves latent dynamics recovery over existing baselines.

7.2 Real-world Validation

We evaluate MAILE on the ASSISTments 2009 dataset (Feng et al., 2009), which contains student–skill interaction logs collected across multiple schools. Following our multi-environment formulation, each school is treated as a distinct environment with its own adaptive assignment policy, while sharing the underlying latent skill dynamics. Because the true latent transition matrix is unavailable in real educational data, we evaluate MAILE from two complementary perspectives: (i) a real-grounded semi-synthetic benchmark with known dynamics, and (ii) reproducibility of the recovered latent dynamics on the original dataset.

Semi-synthetic Validation. We first construct a semi-synthetic benchmark by preserving the real school-level interaction patterns while generating latent transitions from a known ground-truth model. The logged policies are estimated separately for each school, thereby retaining the adaptive data collection characteristics of the original dataset.

Table 2: Real-grounded semi-synthetic recovery on ASSISTments-derived structure and policies (10 seeds). Mean transition MSE below vs. at the threshold; the empirical $E^\star$ matches the theoretical $\lceil d/(k-1) \rceil$ in both the $k = 2$ and $k = 3$ regimes.

d	k	$E^\star_{\text{th}}$	$E^\star_{\text{emp}}$	$\text{MSE}_{<E^\star}$	$\text{MSE}_{=E^\star}$
5	2	5	5	~ 635	4.8
6	3	3	3	~ 905	19.7

Table 3: Reproducibility of the recovered transition structure on real ASSISTments 2009 logs: Pearson correlation of the off-diagonal entries of $\hat{B}$ between disjoint leave-one-out subsets of schools (multi-env) vs. between single schools. Multi-environment pooling is reproducible at $d = 5$; single environments are not.

d	E	Multi-env ρ	Single-env ρ	Diag. std (s/p)
5	9	0.91 ± 0.08	-0.04 ± 0.29	4.72/0.31
8	9	0.24 ± 0.59	0.02 ± 0.13	4.62/0.17

Table 2 shows that MAILE exhibits the same phase transition predicted by Theorem 2. In both binary ($k = 2$) and multi-action ($k = 3$) settings, transition recovery is poor below the theoretical threshold $E^\star = \lceil d/(k-1) \rceil$, but improves dramatically once the threshold is reached. Moreover, the empirical threshold exactly matches the theoretical prediction, demonstrating that the identifiability result extends beyond purely synthetic simulations.

Real-data Reproducibility. Since no ground-truth latent dynamics are available for real students, we evaluate identifiability through reproducibility. Specifically, we estimate the latent transition matrix independently from disjoint leave-one-out subsets of schools and compute the Pearson correlation between the recovered off-diagonal entries.

Table 3 shows that multi-environment pooling produces highly reproducible transition structures ($\rho \approx 0.9$ for $d = 5$), whereas estimates obtained from individual schools are essentially uncorrelated. This demonstrates that policy diversity is essential for recovering stable latent dynamics under adaptive data collection, rather than merely improving predictive performance.

Figure 2(b) visualizes one representative transition matrix recovered by MAILE from the pooled multi-school data. The strong diagonal entries indicate persistent skill mastery, while the off-diagonal structure reveals consistent cross-skill influences that are repeatedly recovered across independent subsets of schools. The figure provides a qualitative illustration of the reproducibility quantified in Table 3.

Overall, the real-world experiments complement the synthetic results by demonstrating that the theoretical identifiability conditions translate into stable latent dynamics recovery under realistic adaptive data collection. Together, Table 2–3 and Figure 2(b) show that policy diversity is not only sufficient in theory, but also essential in practice for recovering reproducible latent transition structures.

8 Conclusion

We introduced the Latent Adaptive Selection Model (LASM) as the first structural causal framework jointly modelling latent state dynamics and adaptive observation mechanisms. Our main theoretical results are: (i) the kernel of the PSGM precisely characterises the confounded directions under single-environment adaptive collection; (ii) with logged policies, multi-environment data from policy-diverse environments identifies the latent transition exactly; and (iii) the minimum number of required environments is $\lceil d/(k-1) \rceil$, a tight bound. These results apply across any domain where latent states are inferred from adaptively collected data. Our future works include extending exact identification to nonlinear transitions, testing the logged-policy recovery on a domain with recorded propensities (e.g. recommendation logs), and adopting richer emission models to scale LASM beyond moderate d.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for manuscript language polishing and assistance in identifying relevant related work. We have not used generative AI tools for generating research results, conducting experiments, analyzing data, or developing scientific claims; the remaining disclosure tasks are not applicable. We have reviewed all AI-assisted work, verified the accuracy of the relevant literature, and revised the AI-assisted text as necessary. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15516–15528, 2022.

Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *International Conference on Artificial Intelligence and Statistics*, pp. 865–873. PMLR, 2024.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38319–38331, 2022.

Zhengming Chen, Feng Xie, Jie Qiao, Zhifeng Hao, Kun Zhang, and Ruichu Cai. Identification of linear latent variable model with arbitrary distribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6350–6357, 2022.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Jan Eriksson and Visa Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19(3): 243–266, 2009.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868, 2019.

Ilyes Khemakhem, Diederik P Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217, 2020.

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, pp. 18171–18206. PMLR, 2023.

Minh Khoa Le, Kien Do, and Truyen Tran. Learning structural causal models from ordering: identifiable flow models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17831–17839, 2025.

Jiahao Liang, Yutian Xiao, Haoran Yang, Zhiwen Yu, Jia-Nan Liu, and Kaixiang Yang. C-hypod: Causal hyperbolic representation learning with prototype orthogonal disentanglement for graph out-of-distribution recommendation. In *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pp. 6541–6550, 2026.

Sarah Mameche, Lénaïg Cornanguer, Urmi Ninad, and Jilles Vreeken. Spacetime: Causal discovery from non-stationary time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19405–19413, 2025.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597, 2017.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55389–55433, 2024.

Hangtong Xu, Yuanbo Xu, Chaozhuo Li, and Fuzhen Zhuang. Causal structure representation learning of unobserved confounders in latent space for recommendation. *ACM Trans. Inf. Syst.*, 43(6): 146:1–146:29, 2025.

A Appendix

A.1 Proof for Propostion 1

Proof. Assume that $\ker(\Phi^e) \neq \{0\}$ and choose any nonzero $v \in \ker(\Phi^e)$. Let $u \in \mathbb{R}^d$ be arbitrary and define the rank-one perturbation

$$M_\lambda = I + \lambda v u^\top,$$

where $|\lambda|$ is sufficiently small so that M_λ is invertible. We construct an alternative latent model by the change of coordinates $\tilde{z} = M_\lambda z$.

The transformed parameters are

$$\tilde{B} = M_\lambda B M_\lambda^{-1}, \qquad \tilde{c}_a = M_\lambda c_a,$$

and the emission function is

$$\tilde{g}(\tilde{z}) = g(M_\lambda^{-1}\tilde{z}).$$

Consequently,

$$\tilde{z}_{t+1} = \tilde{B}\tilde{z}_t + \tilde{c}_{a_t} + \tilde{\epsilon}_t,$$

where $\tilde{\epsilon}_t = M_\lambda \epsilon_t$, so the latent dynamics are well defined under the transformed coordinates.

It remains to verify that the policy is unchanged. For the multinomial logistic policy,

$$\tau(a|z) = \frac{\exp(w_a^\top z + b_a)}{\sum_j \exp(w_j^\top z + b_j)},$$

the decision boundaries depend only on the differences $\{w_a - w_b\}_{a,b}$. Because

$$v \in \ker(\Phi^e) \quad \Longrightarrow \quad (w_a - w_b)^\top v = 0, \qquad \forall a, b.$$

Hence

$$\begin{aligned}(w_a - w_b)^\top M_\lambda z &= (w_a - w_b)^\top (I + \lambda v u^\top) z \\ &= (w_a - w_b)^\top z + \lambda (w_a - w_b)^\top v \, u^\top z \\ &= (w_a - w_b)^\top z.\end{aligned}$$

Therefore,

$$\tau(a|M_\lambda z) = \tau(a|z),$$

so the conditional action distribution is preserved.

Moreover,

$$\tilde{g}(\tilde{z}) = g(M_\lambda^{-1}\tilde{z}) = g(z),$$

implying that the observation model is unchanged. Therefore, for every trajectory, the transformed model induces exactly the same joint distribution over observations and actions as the original model.

Finally,

$$\tilde{B} = M_\lambda B M_\lambda^{-1}.$$

For any $\lambda \neq 0$, this equals B only in the nongeneric case that B commutes with $vu^\top$. Hence, generically,

$$\tilde{B} \neq B,$$

while the induced observational distribution is identical. Thus infinitely many transition matrices are observationally equivalent, and B is not identifiable whenever $\ker(\Phi^e) \neq \{0\}$. □

A.2 Proof for Theorem 1

Proof. Suppose two LASMs generate identical joint distributions over actions and observations in every environment.

By Assumption 3, the emission model is injective, so the two latent representations are related by an affine change of coordinates

$$\widetilde{Z} = MZ + b,$$

for some invertible matrix M and vector b.

Since both models employ the same logged policies, their conditional action distributions must coincide,

$$\tau^e(a|Mz + b) = \tau^e(a|z), \qquad \forall e.$$

For the log-linear policy,

$$\tau^e(a|z) = \frac{\exp((w_a^e)^\top z + \beta_a^e)}{\sum_{a'} \exp((w_{a'}^e)^\top z + \beta_{a'}^e)},$$

this equality implies

$$(w_a^e - w_{a'}^e)^\top (M - I)z = -(w_a^e - w_{a'}^e)^\top b$$

for every z and every pair of actions.

Matching the coefficients of z gives

$$(M - I)^\top (w_a^e - w_{a'}^e) = 0, \qquad \forall a, a', e.$$

Using Remark 3,

$$\mathrm{range}(\Phi^e) = \mathrm{span}\{w_a^e - w_{a'}^e\},$$

we conclude that

$$(M - I)^\top \mathrm{range}(\Phi^e) = 0$$

for every environment.

Since the policy diversity condition implies

$$\sum_{e=1}^{E} \mathrm{range}(\Phi^e) = \mathbb{R}^d,$$

the only linear transformation satisfying the above constraints is

$$M = I.$$

Substituting $M = I$ back into the policy equality gives

$$(w_a^e - w_{a'}^e)^\top b = 0, \qquad \forall a, a', e.$$

Again, policy diversity implies

$$b = 0.$$

Hence

$$\widetilde{Z} = Z.$$

Finally,

$$\widetilde{B} = MBM^{-1} = B,$$

so the latent transition matrix is uniquely determined.

□

A.3 Proof for Theorem 2

Proof. For the log-linear policy, Remark 3 shows that

$$\operatorname{rank}(\Phi^e) \le k-1.$$

Therefore,

$$\operatorname{rank}\left(\sum_{e=1}^{E} \Phi^e\right) \le \sum_{e=1}^{E} \operatorname{rank}(\Phi^e) \le E(k-1).$$

If

$$E(k-1) < d,$$

then

$$\operatorname{rank}\left(\sum_{e} \Phi^e\right) < d,$$

so the policy diversity condition cannot hold. Consequently,

$$E \ge \left\lceil \frac{d}{k-1} \right\rceil$$

is necessary.

Conversely, suppose

$$E = \left\lceil \frac{d}{k-1} \right\rceil.$$

For generic policy parameters, each environment contributes an independent $(k-1)$-dimensional visible subspace. Because

$$E(k-1) \ge d,$$

these subspaces span $\mathbb{R}^d$ with probability one, implying

$$\operatorname{rank}\left(\sum_{e} \Phi^e\right) = d.$$

The policy diversity condition therefore holds almost surely, and Theorem 1 establishes exact identifiability.

Hence

$$E^\star = \left\lceil \frac{d}{k-1} \right\rceil$$

is both necessary and sufficient.

□