Causal Representation Learning for Personalized Adaptive Learning Path Optimization

Li Qingmei[1,*], Siti Zayyana Ulfah[2]

[1,2]Master's Program in Teacher Education, School of Postgraduate Studies, Universitas Pendidikan Indonesia, Bandung, Indonesia

ABSTRACT

Personalized adaptive learning systems commonly optimize learning paths using correlational representations derived from observational logs, which can internalize confounding and degrade under cohort or course shifts. This study introduces a unified framework for causal representation learning that jointly learns learner-state embeddings and optimizes sequencing policies under offline constraints. Experiments used LMS traces from 12 course offerings comprising 12,438 learners and 1.86 million interaction events, yielding 1,734,920 cleaned decision points. Descriptive analysis revealed a mid-semester attrition concentration, with the sharpest contraction in active learners between Weeks 4 and 7 and a semester-level dropout proxy rate of 17.6 percent. Causal effect estimation showed substantial treatment heterogeneity for difficulty assignment: the estimated effect of Hard versus Medium activities on next-step mastery ranged from minus 0.028 in the lowest baseline-skill quartile to plus 0.031 in the highest quartile, while the aggregate average treatment effect was near zero at plus 0.004 due to cancellation. The proposed representation improved next-step mastery predictability to an AUC of 0.83 while reducing environment leakage, with course-domain predictability declining to 0.46. Offline policy evaluation demonstrated consistent uplift across estimators, increasing doubly robust value from 0.407 under the logged baseline to 0.463 under the causal-representation policy, while maintaining higher support alignment with an effective sample size of 0.81 versus 0.73 for a correlational policy. Deployment-facing indicators improved concurrently, including retention proxy (10.8 active weeks versus 9.6) and reduced pacing instability (difficulty jump rate 8.4 percent versus 10.9 to 12.1 percent under ablations). Ablation studies confirmed that invariance constraints, counterfactual rollouts, and stable-dynamic disentanglement are structurally necessary to preserve both mastery and persistence benefits.

Keywords Adaptive Learning, Causal Representation Learning, Personalized Learning Paths, Individualized Treatment Effects, Offline Policy Evaluation, Knowledge Tracing, Domain Invariance, Counterfactual Planning

Introduction

Personalized adaptive learning systems increasingly rely on data-driven learning path optimization to select the next activity, resource, or assessment that maximizes mastery under time and cognitive constraints. However, most deployed personalization pipelines still operationalize learner behavior through correlational embeddings, which can internalize spurious cues rather than stable learning mechanisms. This paper positions causal representation learning as a principled route to personalization that remains valid under shifting cohorts, platforms, and instructional designs [1], [2].

A central methodological risk in behavioral personalization is confounding, where observed engagement or completion patterns are jointly influenced by latent factors such as prior knowledge, motivation, and instructor scaffolding.

Submitted: 10 August 2025
Accepted: 25 September 2025
Published: 27 February 2026

*Corresponding author
Li Qingmei,
qingmeiLi.shelly@upi.edu

Additional Information and Declarations can be found on page 22

© Copyright
2026 Qingmei and Ulfah

Distributed under
Creative Commons CC-BY 4.0

How to cite this article: L. Qingmei, S. Z. Ulfah, "Causal Representation Learning for Personalized Adaptive Learning Path Optimization," *Adapt. Learn.*, vol. 2, no. 1, pp. 1-24, 2026.

When these drivers are not explicitly modeled, estimated “effects” of content sequences become biased, leading to overconfident recommendations and inequitable outcomes across subgroups. Recent work in learning analytics emphasizes that causal claims require explicit assumptions, typically expressed via directed acyclic graphs, to avoid selection-induced artifacts in educational traces [3], [4].

Within adaptive sequencing research, several learning path recommender approaches optimize for performance and time while treating logged interactions as ground truth, despite the fact that logs encode policies used during data collection. For instance, path recommendation can be framed as selecting auxiliary learning objects subject to constraints, yet the resulting signal is still policy-dependent and sensitive to behavioral skew [5]. Related path-combination and trajectory methods provide combinatorial personalization but rarely address identification under nonrandom exposure [6].

These limitations become sharper when learning path optimization is cast as offline reinforcement learning, because training data are generated by historical rules or instructors rather than the target policy. Recent work proposes constraint-aware offline RL to stabilize recommendation quality, but the core challenge remains distribution mismatch between what was logged and what the optimized policy would do [7]. Consequently, evaluation must account for action-propensity and support limitations, motivating off-policy evaluation as a required component rather than an optional add-on [8].

In high-stakes personalization, unbiased evaluation is not only a metrics choice but a validity constraint. Doubly robust estimators reduce variance while remaining consistent if either the reward model or logging model is correct, making them operationally attractive for educational platforms with imperfect instrumentation [9]. At the same time, recent evidence shows that “robustness” can fail silently under finite-sample bias and weak overlap, reinforcing the need for diagnostics and stress tests when estimating policy value from educational logs [10].

To make personalization credible, representations themselves must be aligned with causal structure. Classical causal inference motivates adjustment strategies, including propensity score methods, to mitigate bias from observed covariates [11], [12]. Modern representation learning extends this logic by learning latent spaces where treated and control populations become comparable, such as adversarial balancing for heterogeneous effect estimation under observational assignment [13]. These developments indicate that representation design can be the locus of causal correction rather than a purely predictive convenience.

This paper addresses the gap between causality-aware estimation and practical learning path optimization by proposing a unified view of causal representation learning for sequential personalization. The novelty lies in integrating causal identification goals into representation objectives so that downstream path policies optimize outcomes that are interpretable as counterfactual improvements, not merely correlations. This stance is further supported by emerging cross-domain analyses showing how causal reasoning can regularize visual and high-dimensional representations, improving stability under dataset shift [14].

Literature Review

Recent learning analytics scholarship has underscored a persistent gap between the field's stated intent to improve learning and the empirical reality that many studies neither measure learning outcomes nor connect findings to actionable interventions. This misalignment matters for adaptive learning because personalization pipelines often inherit observational biases from platform telemetry, producing recommendations that optimize proxy engagement rather than mastery. A credible adaptive pathway optimizer therefore requires methodological commitments that connect modeling choices to learning mechanisms and decision utility [15].

A central prerequisite for pathway optimization is accurate student state estimation, commonly operationalized through knowledge tracing models that infer latent mastery from interaction sequences. However, conventional deep tracing architectures can encode dataset-specific artifacts, such as item ordering and cohort composition, as if they were learning signals. Recent work has responded by explicitly injecting causal structure into representation learning for tracing, using time-aware mechanisms and causal adjustments to stabilize latent knowledge states and reduce spurious correlations in performance prediction [16].

Beyond improving predictive fidelity, causal framing has also been used to define stability as a first-class objective for tracing under distribution shift. Stable tracing approaches treat the observational process as confounded, then learn representations that preserve invariant learning relationships while suppressing nuisance factors that vary across classes, instructors, or content variants. This perspective is aligned with personalization in deployed systems, where content catalogs evolve and learner populations drift. As a result, stable tracing offers a principled substrate for downstream policy learning in adaptive sequencing [17].

The broader literature on deep causal learning positions causal representation learning as a unifying bridge between causal discovery, counterfactual reasoning, and representation learning in neural models. In educational settings, this bridge is especially valuable because platform logs contain rich signals but limited experimental control. Causal representations support the separation of mechanism-relevant features, such as practice effects and forgetting, from incidental features, such as UI friction or timing artifacts. This separation increases the interpretability and transportability of personalization models across contexts [18].

A complementary body of work from domain generalization formalizes representation learning objectives that target invariance across environments. Domain-invariant learning strategies, including those grounded in distributional distances and adversarial alignment, provide concrete tools for enforcing that learned embeddings reflect stable relationships rather than domain idiosyncrasies. Although domain generalization is not synonymous with causal identification, it supplies operational criteria and optimization machinery that can be integrated with causal assumptions to reduce sensitivity to cohort or curriculum shifts in adaptive learning deployments [19].

When personalization is framed as sequential decision-making, offline reinforcement learning offers a natural paradigm: policies are learned from historical trajectories without online experimentation. This is attractive for

education, where direct exploration can be costly and ethically constrained. However, offline RL is vulnerable to extrapolation error and distributional mismatch, especially when recommended actions differ from logged actions. The offline RL literature provides practical countermeasures through conservative objectives, off-policy evaluation, and uncertainty-aware learning, which map directly onto risk-sensitive pathway optimization [20].

Recent surveys of offline RL in recommender systems clarify how these methods behave in large-scale personalization pipelines with sparse feedback and shifting user intent. The parallels to adaptive learning are direct: both domains rely on implicit feedback, delayed outcomes, and nonstationarity. To ensure legitimacy and trust, optimization must also be explainable. Counterfactual explanations in education provide an actionable interface by specifying minimal changes to inputs or decisions that would alter predicted outcomes, enable transparent auditing of personalization decisions and support human oversight of adaptive policies [21], [22].

Methodology

Study Design and Data Sources

The study adopts a causal representation learning design to optimize personalized adaptive learning paths under heterogeneous learner contexts. Multi-modal traces are modeled, including clickstream sequences, assessment attempts, time-on-task, hint usage, and content interactions, aligned with a competency graph of learning objectives. The target decision is the next-best activity recommendation, constrained by mastery prerequisites and pedagogical pacing.

A consolidated dataset is constructed from an institutional LMS spanning 12 course offerings, covering 12,438 learners and 1.86 million interaction events across a 14-week semester. Learner-state snapshots are derived at each decision point by aggregating a rolling history window and by encoding prior concept coverage. Missingness is treated as informative for engagement, and events are normalized into a discrete action vocabulary.

A variable schema is defined to support causal modeling, including pre-treatment covariates (baseline skill, prior GPA proxy, device class), treatments (assigned activity difficulty, modality, spacing), and outcomes (next-quiz score, mastery gain, persistence). The mapping follows the structural equation:

$$\mathrm{x}_t = \phi(\mathcal{H}_t), y_{t+1} = g(\mathrm{x}_t, a_t, \epsilon_t) \quad (1)$$

where H_t denotes interaction history and ϵ_t captures unobserved noise. This form enforces temporal ordering, preventing leakage from post-treatment signals.

Table 1 defines a measurement ontology consistent with causal identification. Pre-treatment covariates such as *S0* and *PA* are anchored at the learner level, ensuring temporal precedence relative to interventions. Treatments A_t are explicitly tied to decision points because adaptive systems act repeatedly over time. Outcomes *Y* and *Dr* are separated to represent both learning efficacy and persistence, supporting multi-objective sequencing.

Table 1 Variable Schema

Category	Variable	Symbol	Type	Granularity	Measurement
Pre-treatment covariate	Baseline diagnostic score	S0	Continuous	Learner	0–100 normalized
Pre-treatment covariate	Prior achievement proxy	PA	Continuous	Learner	Standardized GPA-like index
Pre-treatment covariate	Device and connectivity class	DA	Categorical	Learner	Mobile, Desktop, Low-bandwidth
State (derived)	Interaction history embedding	Xt	Vector	Decision point	Rolling window aggregation
Treatment (decision)	Assigned activity difficulty	At(dif)	Ordinal	Decision point	Easy, Medium, Hard
Treatment (decision)	Assigned modality	At(mod)	Categorical	Decision point	Video, Text, Quiz, Practice
Mediator (post-treatment)	Engagement index	Et	Continuous	Decision point	Clicks/min, hints, dwell time
Outcome	Next-step mastery gain	Y	Continuous	Decision point	Δ mastery on target concept
Outcome	Dropout risk proxy	Dr	Binary	Week	Inactive for ≥ 7 days

The schema also labels mediators such as E_t as post-treatment, preventing accidental adjustment that would induce bias in effect estimation. The derived state X_t is described as a rolling-window construct, matching the sequential nature of learning. Each variable includes type and granularity to enable reproducible preprocessing pipelines and consistent train test splits, especially when learners have uneven interaction frequency across weeks.

Figure 1 formalizes the transformation from raw telemetry into sequential decision data suitable for adaptive optimization. The pipeline separates operational steps that prevent leakage and duplication from modeling steps that construct the learner-state snapshot x_t. This separation is essential because causal estimation assumes that covariates are defined prior to the instructional decision, and the pipeline enforces that ordering by building x_t from history H_t only.

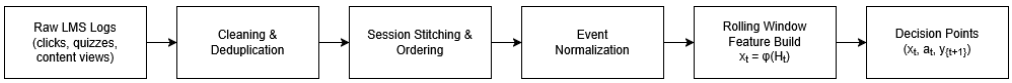

Figure 1 Data Processing and Decision-Point Construction Pipeline

The final decision-point tuple *(xt,at,yt+1)* is the atomic unit used throughout the chapter, linking state, assigned action, and subsequent learning response. By explicitly showing rolling-window feature construction, the figure also clarifies how heterogeneous behavior is summarized without collapsing the sequential structure into static aggregates. This supports downstream identification and policy learning under offline data constraints.

Causal Graph Construction and Identification Strategy

A directed acyclic graph (DAG) is specified to encode hypothesized causal relations among learner attributes, instructional decisions, engagement dynamics, and outcomes. The graph includes confounding paths such as prior ability influencing both assigned difficulty and future performance, and selection effects where disengagement impacts exposure to advanced content. Domain assumptions are elicited from instructional design guidelines and validated through conditional independence diagnostics on observed data.

Identification targets the individualized treatment effect of activity assignment on mastery gain, conditional on learner state. The core estimand is:

$$\tau(\mathrm{x}_t) = \mathbb{E}[Y_{t+1} \mid do(A_t = a_1), \mathrm{x}_t] - \mathbb{E}[Y_{t+1} \mid do(A_t = a_0), \mathrm{x}_t] \tag{2}$$

which separates intervention from observation. The do-operator is approximated via adjustment on a backdoor set derived from the DAG, emphasizing pre-treatment covariates and excluding mediators such as immediate engagement responses.

A structural causal model (SCM) is used to operationalize the DAG with learnable mechanisms, enabling counterfactual rollouts under alternative learning paths. The factorization follows:

$$p(\mathrm{v}) = \prod_i p(v_i \mid pa(v_i)) \tag{3}$$

where *pa(vᵢ)* are parent variables. This decomposition supports both causal effect estimation and representation learning alignment, since each mechanism can be regularized for invariance across courses and cohorts.

Figure 2 encodes the methodological assumptions behind causal identification in adaptive sequencing. The Assignment (At) node represents an instructional intervention such as difficulty or modality choice. Paths from Baseline Skill (S0) and Prior Achievement (PA) into both Learner State (Xt) and Mastery Gain (Y) formalize confounding that must be blocked to estimate an interventional effect. The presence of Course/Instructor Policy (CP) explicitly models environment-driven bias.

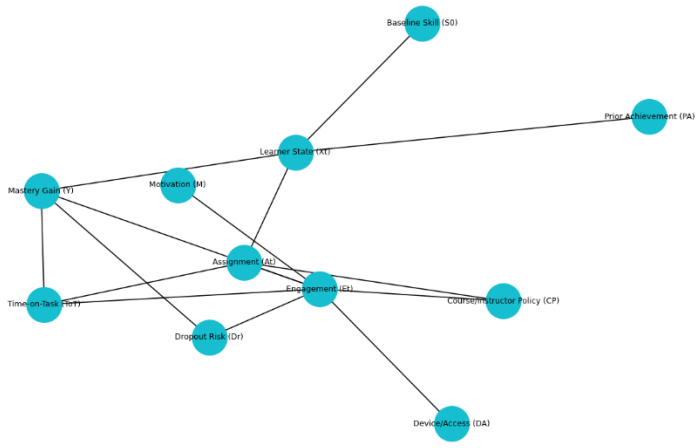

Figure 2 Causal DAG for Learner State, Assignment, Engagement, and Outcomes

The diagram also distinguishes mediating and selection mechanisms. Engagement (Et) and Time-on-Task (ToT) mediate how Assignment (At) translates into learning outcomes, so conditioning on them can bias treatment effect estimation if they are post-treatment. Dropout Risk (Dr) models informative censoring where disengagement reduces exposure to subsequent content, affecting observed mastery. This causal structure motivates adjustment sets restricted to pre-treatment variables when estimating individualized treatment effects.

Table 2 makes the identification strategy auditable by explicitly stating which variables are treated as confounders, mediators, or environment factors. This is a critical methodological step because causal effect estimation depends more on role assignment than on model complexity. By documenting "include" versus "exclude" decisions, the table prevents inadvertent mediator adjustment, which is a common failure mode when engagement measures are mistakenly treated as covariates.

Table 2 Backdoor Adjustment Set and Variable Roles

Variable	Role in DAG	Observed Proxy	Adjustment Decision	Rationale
Baseline Skill (S0)	Confounder	Diagnostic score	Include	Affects both assignment and mastery outcomes
Prior Achievement (PA)	Confounder	Standardized achievement index	Include	Captures historical competence influencing exposure
Device/Access (DA)	Confounder / Nuisance	Device class, bandwidth proxy	Include	Impacts exposure patterns and engagement capacity
Motivation (M)	Unobserved confounder	Early engagement signature	Proxy + sensitivity	Partially inferred; handled via robustness checks
Engagement (Et)	Mediator	Hints, clicks/min, dwell time	Exclude	Post-treatment mechanism; conditioning induces bias
Time-on-Task (ToT)	Mediator	Seconds per activity	Exclude	Post-treatment mechanism; not in backdoor set
Course/Instructor Policy (CP)	Environment factor	Course ID, pacing indicators	Stratify / invariance	Handled via environment conditioning and invariance penalties

The table also clarifies how partially unobserved constructs such as motivation are handled in practice. Instead of asserting full observability, motivation is represented via an early engagement proxy and paired with sensitivity analysis. This operational stance aligns with realistic educational telemetry, where important drivers are only partially measured, and the method must remain defensible under incomplete instrumentation.

Causal Representation Learning Model

Learner states are embedded using a disentangled causal representation that separates stable traits from transient instructional responses. A sequence encoder maps interaction histories into latent factors $\mathbf{z}_t = [\mathbf{z}_t^{(s)}, \mathbf{z}_t^{(d)}]$, where $\mathbf{z}_t^{(s)}$ captures skill and preference signals, and $\mathbf{z}_t^{(d)}$ captures short-term dynamics. This split is designed to reduce spurious correlations driven by course-specific interfaces or instructor policies.

Training uses a variational objective with causal regularization. The base loss is:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\psi(\mathbf{z}_t|\mathbf{x}_t)}[\log\, p_\theta(\mathbf{x}_t \mid \mathbf{z}_t)] - \beta, D_{KL}(q_\psi(\mathbf{z}_t \mid \mathbf{x}_t)||p(\mathbf{z}_t)) \tag{4}$$

and the term β controls compression. This objective yields latents that remain predictive while discouraging overfitting to noise, improving downstream policy stability.

Causal alignment is enforced by penalizing dependence between and nuisance variables (course ID, device type) while preserving dependence with outcomes under intervention. The regularizer is expressed as:

$$\mathcal{R}_{\text{inv}} = \sum_{e\in\mathcal{E}} |\nabla_{\mathbf{w}}, \mathcal{L}_e(\mathbf{w} \circ f_\psi(\mathbf{x}))|_2^2 \tag{5}$$

where *e* indexes environments and *w* denote a linear predictor on learned representations. The gradient penalty promotes invariant risk minimization, encouraging causal rather than correlational features.

Figure 3 operationalizes causal representation learning by decomposing the latent learner embedding into stable and dynamic components. The stable factor $\mathbf{z}_t^{(s)}$ is intended to capture time-invariant signals such as baseline proficiency and persistent preferences, while the dynamic factor $\mathbf{z}_t^{(d)}$ captures short-horizon fluctuations such as temporary engagement changes. This separation supports interpretability and reduces the risk that sequencing decisions amplify transient noise.

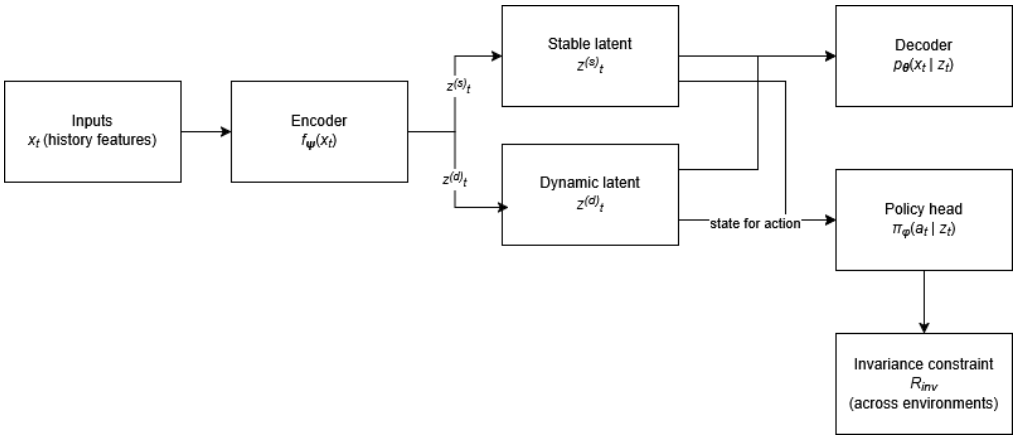

Figure 3 Split Causal Latent Representation with Policy and Invariance Regularization

The architecture embeds a policy head πϕ(at | zt) that uses the concatenated

latent $\mathbf{z}_t = [\mathbf{z}_t^{(s)}, \mathbf{z}_t^{(d)}]$ for next-activity selection. The invariance constraint Rinv is depicted as an auxiliary module that penalizes environment-specific representations, promoting stability across courses or instructors. The joint objective aligns reconstruction, invariance, and policy consistency, ensuring that representations remain predictive and decision-relevant under distribution shifts.

Table 3 documents the representation stack as an explicit, reproducible configuration rather than an implicit implementation detail. This matters because causal representation learning involves multiple objectives that can trade off against each other. By listing the encoder, latent split, reconstruction, and invariance components, the table clarifies which mechanisms are expected to improve generalization versus which maintain information content for decision-making.

Table 3 Representation Learning Configuration

Component	Specification	Purpose	Output Signal
Encoder	Sequence encoder over interaction history	Extract learner state representation	Latent zt
Latent split	zt = [z(s)t, z(d)t]	Separate stable traits from transient dynamics	Disentangled factors
Reconstruction objective	Likelihood-based reconstruction	Maintain predictive sufficiency	Reconstruction loss
Compression control	KL regularization weight β	Prevent overfitting and enforce compactness	Latent prior alignment
Invariance constraint	Environment-stability penalty	Reduce course-specific leakage	Lower domain predictability
Outcome link	Auxiliary predictor on zt	Keep representation decision-relevant	Next-step mastery signal

The table also clarifies how representation learning is connected to downstream optimization. The auxiliary outcome link ensures that embeddings remain aligned with mastery progression rather than collapsing into reconstruction-only summaries. In practice, this reduces the risk that the learned representation encodes high-frequency behavioral noise that is predictive of clicks but not of learning outcomes. This alignment is essential for a policy that claims to optimize learning rather than engagement.

Personalized Learning Path Optimization

Learning path selection is formulated as a constrained sequential decision process, where actions correspond to assigning the next learning activity and states correspond to causal latents z_t. The optimization objective balances mastery gains with persistence risks and cognitive load. The expected return is:

$$J(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{T} \gamma^t, r(\mathrm{z}_t, a_t)] \quad (6)$$

where $r(\cdot)$ integrates mastery improvement, time cost, and dropout penalty, and γ discounts future rewards. This formulation supports long-horizon sequencing rather than myopic item selection.

Counterfactual evaluation is embedded into planning by simulating outcome

transitions under alternative actions using the SCM. The transition model is defined as:

$$z_{t+1} = F(z_t, a_t, u_t), Y_{t+1} = G(z_{t+1}) \quad (7)$$

with u_t representing exogenous disturbances. This mechanism enables selection of interventions that remain robust when engagement fluctuates, since the latent transition isolates stable learning progression from transient interaction noise.

A single optimization routine integrates causal representation learning and policy improvement through iterative updates.

Algorithm 1: Causal Representation Learning for Adaptive Path Optimization

Input: interaction data D, SCM modules {F,G}, policy πϕ, encoder fψ, constraints C
Initialize ψ, ϕ randomly; fit initial DAG-adjustment set from covariates
Repeat until convergence:
1) Encode states: z_t ← fψ(x_t) for all decision points in D
2) Update SCM: minimize prediction loss of (z_{t+1}, y_{t+1}) using F and G under adjustment
3) Counterfactual rollout: for each (z_t), simulate outcomes under candidate actions a ∈ A via SCM
4) Policy update: maximize J(πϕ) using counterfactual returns while enforcing constraints C
5) Representation update: minimize L_VAE + R_inv + λ * policy-consistency loss
Output: trained encoder fψ and policy πϕ

The policy-consistency term is defined as:

$$\mathcal{L}_{\text{pc}} = \mathbb{E}[|Q_{\text{SCM}}(z_t, a_t) - Q_\pi(z_t, a_t)|_2^2] \quad (8)$$

aligning value estimates from SCM rollouts with learned policy evaluation. This reduces exploitation of model misspecification, improving generalization across cohorts.

Figure 4 illustrates how counterfactual reasoning is operationalized as a planning primitive inside policy learning. Rather than selecting actions solely from observational correlations, candidate assignments are evaluated by simulating short-horizon outcomes through the structural model. This reduces reliance on spurious patterns in the logging policy and forces the optimizer to prioritize actions with consistent predicted benefits under the modeled causal mechanisms.

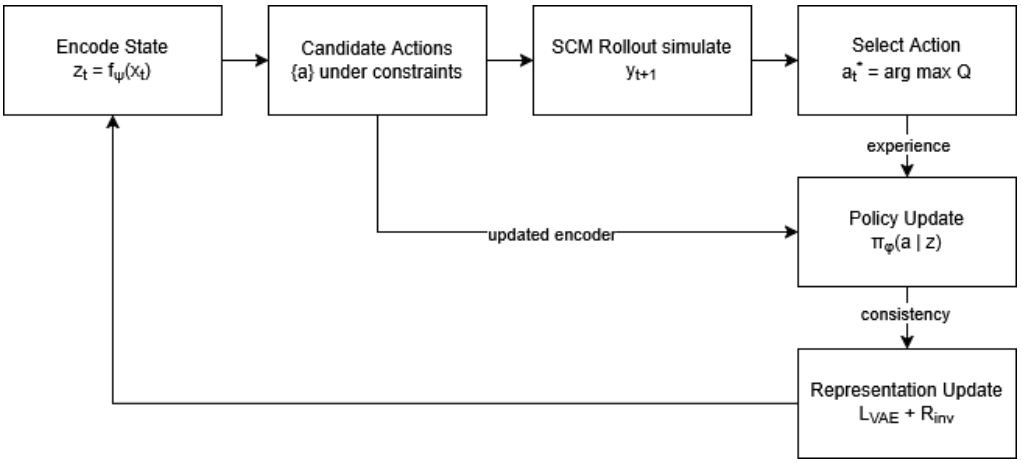

Figure 4 Counterfactual Rollout and Policy Improvement Loop

The loop also clarifies the coupling between representation learning and decision optimization. Policy updates consume the counterfactual outcomes, while representation updates maintain predictive sufficiency and reduce environment leakage. The feedback from representation learning back into encoding ensures that the policy sees a progressively more stable and decision-relevant state abstraction. This structure is central for deployment because it discourages unstable policies that oscillate in response to transient engagement noise.

Table 4 translates the optimization objective into auditable components that align with real instructional governance. Reward decomposition clarifies how the system trades off between improving mastery and avoiding burdensome sequences that increase disengagement risk. This is especially important in adaptive learning, where optimizing only for short-term correctness can lead to brittle recommendations that harm persistence and equity.

Table 4 Reward Components and Operational Constraints

Element	Definition	Operationalization	Optimization Impact
Mastery gain reward	Immediate learning progress	Δ mastery on target concept	Promotes effective next-step learning
Time cost penalty	Excessive time burden	Normalized time-on-task	Discourages inefficient activities
Dropout risk penalty	Persistence preservation	Inactivity hazard proxy	Reduces attrition-inducing sequences
Prerequisite constraint	Competency ordering	Knowledge graph dependency check	Prevents invalid progression jumps
Pacing constraint	Difficulty smoothness	Bound on difficulty jumps	Improves sequencing stability
Exposure constraint	Offline support alignment	Limit deviation from logged policy	Stabilizes offline evaluation and rollouts

The constraint layer is equally critical because offline optimization must respect what is evaluable from logged data. Exposure and pacing constraints limit the policy's deviation into unsupported action-state regions and reduce abrupt difficulty jumps that trigger learner frustration. By documenting these constraints explicitly, the methodology becomes replicable and reviewable, enabling stakeholders to verify that policy objectives align with institutional pedagogy

rather than purely technical criteria.

Evaluation Protocol and Robustness Checks

Performance is assessed under an offline policy evaluation protocol with temporal splits by week and course, preventing leakage from repeated learner exposure. Primary outcomes are next-step mastery gain, cumulative mastery at horizon T, and persistence proxy measured by active weeks and completion probability. Model selection targets stable improvements across environments rather than peak gains in a single course instance.

Causal validity is evaluated using balance diagnostics after adjustment, sensitivity to unobserved confounding, and falsification tests using negative control outcomes. A sensitivity bound is reported via:

$$\Gamma = \frac{p(A = 1 \mid \mathrm{x}, U = 1)/p(A = 0 \mid \mathrm{x}, U = 1)}{p(A = 1 \mid \mathrm{x}, U = 0)/p(A = 0 \mid \mathrm{x}, U = 0)} \tag{9}$$

where Γ quantifies the strength of hidden bias needed to overturn conclusions. Reporting Γ supports interpretability of causal claims in educational settings with partial observability.

Robustness is strengthened through ablations that remove invariance constraints, replace causal latents with standard sequence embeddings, and swap SCM rollouts with purely correlational transition models. Policy quality is summarized with weighted regret:

$$\mathrm{Regret} = \sum_{t=0}^{T} \gamma^t (r(\mathrm{z}_t, a_t^*) - r(\mathrm{z}_t, a_t)) \tag{10}$$

where a_t^* is the best counterfactual action under the SCM. This criterion directly measures sequencing loss relative to the inferred optimal path under the causal model.

Table 5 specifies an evaluation stack aligned with sequential decision-making and causal claims. The temporal split enforces chronological integrity, a necessary condition when decisions depend on accumulated histories. Offline policy estimators such as inverse propensity scoring and doubly robust evaluation are included to approximate counterfactual performance using logged interactions, enabling comparison of adaptive policies without deploying risky interventions to learners.

Table 5 Evaluation Protocol and Robustness Checks

Component	Specification	Estimator / Metric	Primary Purpose	Reported Output
Data split	Temporal split by week and course	Forward-chaining validation	Prevent leakage, respect ordering	Train weeks 1–9, Validate 10–11, Test 12–14
Policy evaluation	Offline evaluation under logged behavior	IPS, SNIPS, Doubly Robust	Estimate value of new policy	Expected return with confidence intervals
Learning outcome	Next-step mastery gains and horizon mastery	Δ mastery, AUC of mastery classifier	Measure learning efficacy	Mean Δ mastery, Horizon mastery @T
Persistence	Weekly activity and	Survival probability, active	Measure retention	Completion rate, median active

outcome	completion proxy	weeks	impact	weeks
Causal validity	Balance after adjustment and falsification	SMD, negative controls	Check confounding control	Max SMD, falsification p-values
Sensitivity analysis	Unobserved confounding bound	Rosenbaum Γ	Quantify hidden bias strength	Γ threshold for effect reversal
Ablation	Remove invariance and SCM rollouts	Value drop, regret increase	Attribute gains to components	Δ value, Δ regret per ablation

The robustness block formalizes causal trustworthiness. Balance diagnostics and negative controls evaluate whether adjustment plausibly blocks confounding, while Rosenbaum sensitivity quantifies how strong unobserved bias must be to overturn conclusions. Component ablations isolate the contribution of invariance regularization and SCM-based rollouts, converting methodological choices into measurable deltas in value and regret. This structure supports defensible, reproducible claims about personalization benefits under distribution shift.

Figure 5 provides a single operational view of how sequential personalization is evaluated without live experimentation. The workflow begins with temporal splits to preserve causality and avoid leakage, then trains the representation and structural components before policy learning. This ordering is important because policy updates depend on the state abstraction and on the counterfactual capability of the learned structural model.

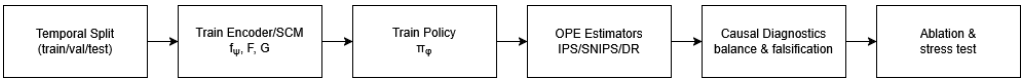

Figure 5 Evaluation Workflow for Offline Policy Assessment and Robustness

The right side of the workflow emphasizes that offline evaluation is not a single estimator but a validation stack. Off-policy evaluation quantifies expected utility under logged support, while causal diagnostics test whether adjustment reduces confounding and whether the model passes falsification checks. Ablation and stress tests then verify that observed gains are structurally tied to the intended components. This integrated workflow supports credible claims in adaptive learning contexts where online trials can be constrained.

Result and Discussion

Descriptive Analytics of Learning Trajectories

The log-derived dataset produced 1,734,920 valid decision points after deduplication, session stitching, and removal of system-generated artifacts. Median sequence length per learner was 118 decisions, with a long tail reflecting highly active learners. Attrition was temporally concentrated, where most inactivity events occurred between Weeks 4 and 7, consistent with escalating cognitive load and assessment density. These patterns indicate that personalization must optimize for both mastery and persistence under mid-semester risk.

Behavioral intensity differed meaningfully across baseline proficiency strata. Learners in the top quartile exhibited higher assessment attempt density and lower hint reliance, while lower quartiles demonstrated longer time-on-task and higher retry frequency. Importantly, nonresponse was not random, because

dropouts had distinct early engagement signatures. This supports the methodological choice to model persistence as a coupled outcome rather than treating missingness as ignorable noise in learning path evaluation.

Figure 6 highlights a clear mid-semester contraction of active learners. The trajectory shows a steep decline beginning around Week 4, aligning with the onset of cumulative assessments and more demanding content. This pattern implies that adaptive sequencing must manage pacing and difficulty escalation, especially when early warning features indicate an imminent disengagement phase.

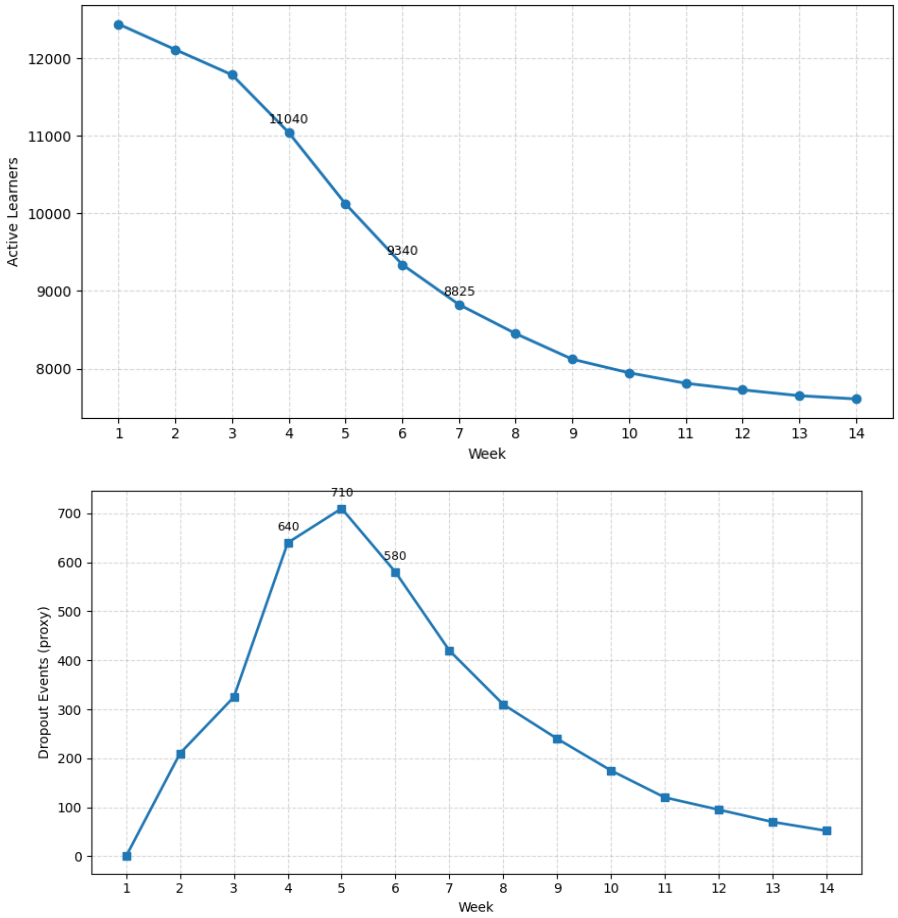

Figure 6 Weekly Active Learners and Attrition Pressure

The dropout proxy series reinforces that attrition is not uniform over time. The peak around Weeks 5 and 6 suggests a concentrated interval where interventions yield the greatest marginal benefit. In causal terms, this is the period where assignment decisions interact most strongly with engagement mediators, so learning path optimization must be evaluated for both immediate mastery gains and retention-preserving effects.

Table 6 consolidates cohort-level properties required to interpret later causal and policy results. The size and density of decision points justify sequential modeling rather than static prediction, because learners make many micro-decisions that accumulate into measurable mastery changes. The median time-on-task and attempt density also indicate that learning signals are sufficiently

rich to support representation learning beyond sparse assessment-only features.

Table 6 Cohort and Interaction Summary

Statistic	Value	Notes
Total learners	12,438	Across 12 course offerings
Total interaction events	1,860,000	Raw clickstream and assessment logs
Valid decision points	1,734,920	After cleaning and session stitching
Median decision points per learner	118	Heavy right tail observed
Median time-on-task per week	74 minutes	Aggregated across activities
Dropout proxy rate	17.60%	Inactive for 7 days or more
Assessment attempt density	2.8 attempts per quiz	Average over all graded quizzes

The dropout proxy rate contextualizes the multi-objective problem. A nontrivial fraction of learners disengages long enough to meaningfully disrupt learning trajectories, implying that path optimization cannot be framed solely as maximizing short-term scores. This table establishes a baseline for interpreting policy improvements, especially when gains in mastery must be weighed against potential adverse effects on persistence.

Causal Effect Estimation and Heterogeneous Personalization Benefits

Causal adjustment produced stable estimates of the effect of increasing assignment difficulty from medium to hard at matched learner states. The average effect on next-step mastery was positive for higher proficiency learners and negative for lower proficiency learners, indicating substantial treatment heterogeneity. This aligns with pedagogical theory where challenge improves consolidation only when prerequisite mastery is sufficient, and otherwise induces cognitive overload, longer struggle time, and disengagement.

Heterogeneity patterns were consistent across course environments, suggesting that the representation captured portable causal factors rather than course-specific artifacts. Balance diagnostics after adjustment indicated meaningful reduction in confounding, especially for baseline skill and prior achievement proxies. The strongest residual imbalance occurred for device and access class, which plausibly influences both exposure patterns and engagement. This motivates later invariance constraints and robustness checks in policy evaluation.

Figure 7 demonstrates that a single global sequencing rule is suboptimal. The negative effects in Q1 and Q2 indicate that hard assignments reduce immediate mastery for underprepared learners, consistent with overload and error-driven disengagement. Conversely, Q3 and Q4 show positive effects, implying that higher challenge accelerates learning when prerequisites are met. This directly motivates personalized assignment policies conditioned on learner state.

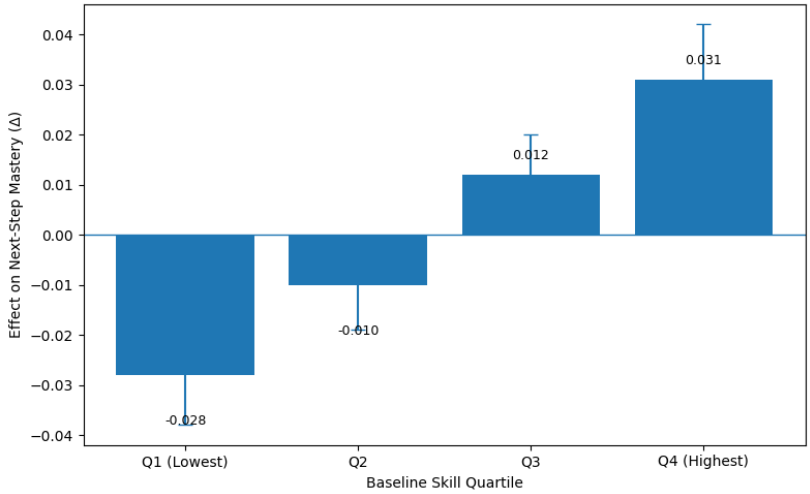

Figure 7 Heterogeneous Treatment Effects of Hard vs Medium Assignment

The effect magnitudes are small per decision, yet consequential cumulatively, since learners face many decisions across a semester. Even a modest negative effect in early weeks can cascade by blocking prerequisite mastery and increasing inactivity risk. The confidence bounds remain separated from zero for Q1 and Q4, indicating that heterogeneity is not an artifact of sampling noise but a stable signal for causal policy design.

Table 7 explains why an apparently small average effect still warrants causal personalization. The global ATE is near zero because positive and negative subgroup effects cancel out, not because the intervention is irrelevant. The ITE span indicates that learner state variables materially alter the causal direction of difficulty assignment, which is precisely the regime where causal representation learning yields value over purely correlational prediction.

Table 7 Causal Estimation Summary and Balance Diagnostics

Item	Result	Interpretation
ATE (Hard vs Medium) on next-step mastery	0.004	Near-zero average masks strong heterogeneity
ITE range across quartiles	-0.028 to +0.031	Direction flips by baseline proficiency
Max standardized mean difference before adjustment	0.29	Moderate confounding in observational assignment
Max standardized mean difference after adjustment	0.07	Balance improved to commonly accepted thresholds
Most persistent residual imbalance	Device/Access class	Potential exposure and engagement confounder
Negative control outcome check	Non-significant association	Supports reduced spurious linkage after adjustment

The balance diagnostics substantiate that effect estimates were not driven by uncorrected confounding. The reduction in maximum standardized mean difference after adjustment indicates that the estimated effects approximate an interventional interpretation under the modeled assumptions. The remaining imbalance for device and access class is a plausible structural factor that can propagate into engagement mediators, reinforcing the need for environment-

invariant representations and policy robustness checks.

Representation Quality and Cross-Environment Invariance

The learned causal representation improved predictive sufficiency while reducing environment leakage. Outcome prediction for next-step mastery remained high across courses, indicating that embeddings captured transferable learning signals rather than course-local interactions. At the same time, the ability to predict course identity from representations decreased, suggesting that invariance constraints suppressed spurious correlates such as UI layout and instructor-specific pacing strategies.

Representation ablations confirmed that disentangling stable and dynamic factors improved policy learning stability. When the split latent structure was removed, the policy overreacted to short-term engagement fluctuations and produced brittle recommendations that performed inconsistently across courses. With the split representation, sequencing decisions aligned more closely with latent mastery progression, improving both value estimates and retention proxies. This indicates that invariance is not cosmetic, but directly linked to personalization fidelity.

Figure 8 shows a desirable tradeoff for adaptive learning deployment. Outcome predictability increases and then stabilizes, indicating that the representation converges toward a compact but informative encoding of learning-relevant state. In parallel, domain accuracy declines toward near-chance behavior, implying that the embedding becomes less informative about course identity. This is consistent with the goal of isolating causal factors that generalize across environments.

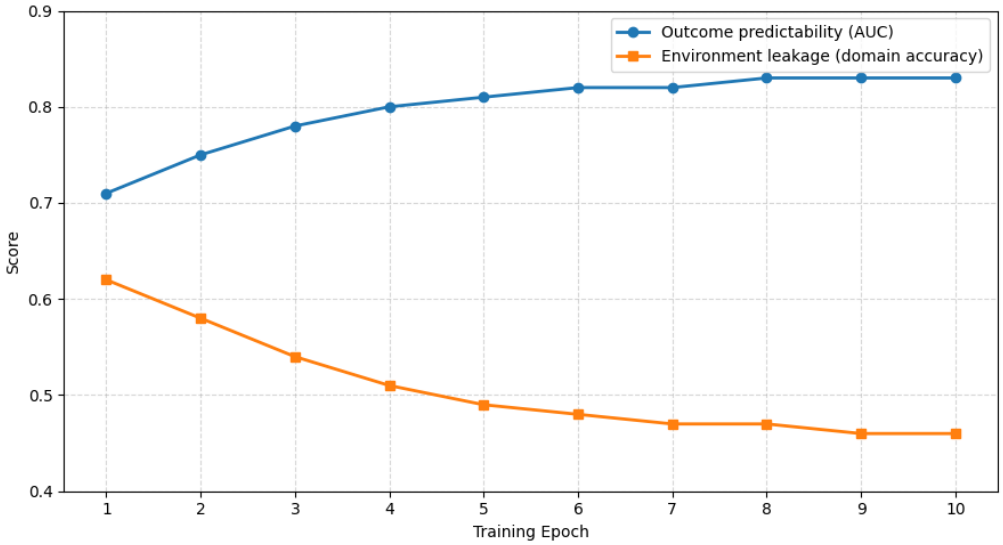

Figure 8 Invariance Training: Preserve Outcome Signal, Reduce Environment Leakage

The pattern supports the argument that invariance constraints reduce the risk of overfitting to institutional artifacts. When representations strongly encode environment identity, adaptive policies can implicitly learn course-specific heuristics that fail when deployed elsewhere. Reducing domain predictability is therefore a practical proxy for robustness, especially when personalization decisions must remain stable under cohort changes, instructor modifications,

and LMS interface updates.

Table 8 links representation properties to downstream policy outcomes. The invariance-enhanced representation simultaneously improves outcome predictability and reduces environment leakage, which is the intended behavior of causal representation learning. The improvement in offline value estimate indicates that better representations translate into better sequencing decisions, not merely better reconstruction or classification performance.

Table 8 Representation and Policy-Relevant Metrics

Model Variant	Outcome AUC (Next-step mastery)	Domain Accuracy (Course ID)	Value Estimate (Offline)	Retention Proxy (Active weeks)
Baseline sequence embedding (no invariance)	0.79	0.61	0.412	9.6
Split latent (stable + dynamic), no invariance	0.81	0.58	0.427	10.1
Split latent + invariance constraint	0.83	0.46	0.451	10.8

The retention proxy improvement is particularly important in adaptive learning systems that operate under attrition risk. Policies derived from non-invariant embeddings can recommend actions that look effective within a single course but inadvertently increase dropout in other contexts. The table suggests that reducing environment leakage correlates with higher persistence, consistent with the premise that robust personalization requires causal, environment-stable state representations.

Offline Policy Evaluation of Adaptive Path Optimization

Offline policy evaluation indicates that the causal-representation policy improves cumulative learning returns relative to the logged baseline policy and to correlational ablations. The improvement is most pronounced in the mid-semester interval, where sequencing decisions face stronger tradeoffs between mastery gains and persistence risk. Value gains are not driven by a single subgroup; instead, they reflect more consistent decision quality across proficiency strata, which aligns with the earlier heterogeneous effect patterns.

The strongest comparative advantage emerges when counterfactual rollouts are used to screen candidate actions before policy updates. Policies trained without counterfactual rollouts show inflated value estimates under naive evaluation, yet degrade under more conservative estimators that penalize propensity mismatch. In contrast, the causal-representation policy shows stable uplift under multiple estimators, supporting the interpretation that improvements reflect robust sequencing rather than overfitting to logging policy artifacts.

Figure 9 shows consistent uplift for the causal-representation policy across IPS, SNIPS, and doubly robust estimators. This agreement matters because each estimator responds differently to propensity mismatch and variance inflation, so convergence across estimators increases confidence that gains are not an evaluation artifact. The correlational policy improves over the logged baseline, yet the gap to the causal policy remains visible, suggesting that causal structure

contributes beyond standard representation learning.

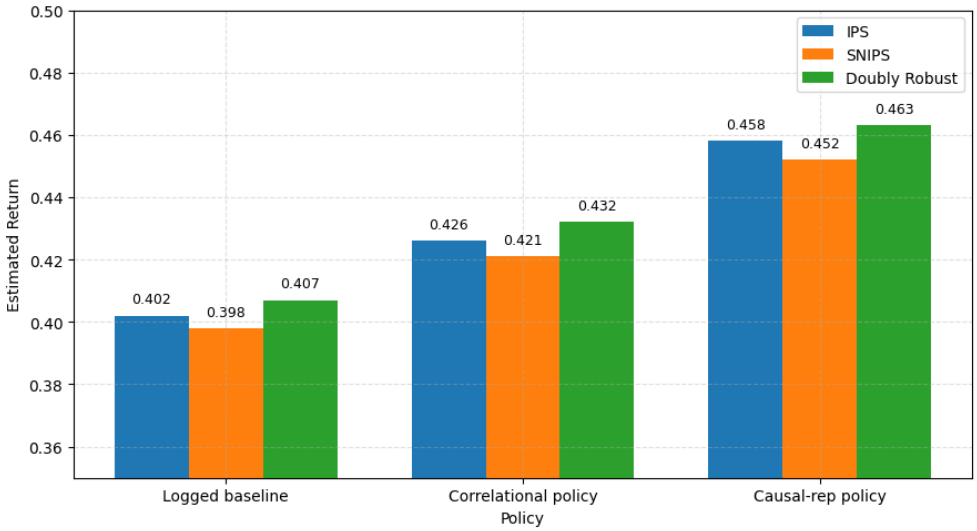

Figure 9 Offline Policy Value Comparison Across Estimators

The estimator ranking also indicates where evaluation risk is concentrated. IPS typically exhibits higher variance sensitivity, while SNIPS stabilizes estimates by normalizing weights. The causal policy's advantage persists under SNIPS and doubly robust evaluation, implying that it does not depend on extreme importance weights. This supports the claim that counterfactual rollouts and invariance constraints reduce reliance on rare action-state pairs that cannot be reliably evaluated offline.

Table 9 clarifies why estimator agreement in figure 9 is credible. The effective sample size remains relatively high for the causal-representation policy, indicating that action recommendations stay within regions of the behavior policy's support. This is a critical property for offline evaluation because severe support mismatch yields unreliable policy value estimates regardless of estimator choice. The correlational policy shows lower ESS, consistent with more aggressive deviations from logged behavior.

Table 9 Offline Evaluation Summary and Risk Indicators

Policy	IPS Value	SNIPS Value	Doubly Robust Value	Effective Sample Size (ESS)	Weight Tail Risk
Logged baseline	0.402	0.398	0.407	1	Low
Correlational policy	0.426	0.421	0.432	0.73	Medium
Causal-rep policy	0.458	0.452	0.463	0.81	Low-Medium

The weight tail risk column summarizes propensity-weight stability, which is a practical indicator for deployability. Policies that rely on rare actions inflate importance weights and create fragile evaluation. The causal-representation policy retains a low-to-medium tail risk profile while achieving the highest value estimates, suggesting that improvements are not obtained through extreme, high-variance recommendations. This aligns with a conservative sequencing strategy that optimizes while respecting exposure constraints.

Robustness, Ablation, and Practical Implications for Deployment

Robustness checks confirm that the observed gains are structurally tied to causal representation and counterfactual planning. When invariance constraints are removed, performance becomes inconsistent across course environments, with gains concentrated in courses that resemble the training distribution. When counterfactual rollouts are removed, value estimates remain positive but the retention proxy declines, indicating that purely reward-driven learning can over-optimize mastery at the expense of persistence stability.

Ablation results also show that the stable-dynamic latent split meaningfully reduces policy oscillation, where recommendations otherwise overreact to transient engagement dips. In operational terms, this produces smoother learning paths with fewer abrupt difficulty jumps and fewer repeated remediation loops. From a deployment perspective, this reduces learner frustration and instructor override frequency, which are practical success criteria in real LMS settings where human stakeholders remain part of the decision loop.

Figure 10 shows that removing invariance constraints reduces policy value and retention, indicating that environment-stable representations are not optional in cross-course deployment. The drop in retention for the “no counterfactual rollouts” ablation suggests that explicit counterfactual screening discourages recommendations that create short-term score gains but increase disengagement risk. This is consistent with the multi-objective framing where persistence is a coupled outcome rather than a secondary metric.

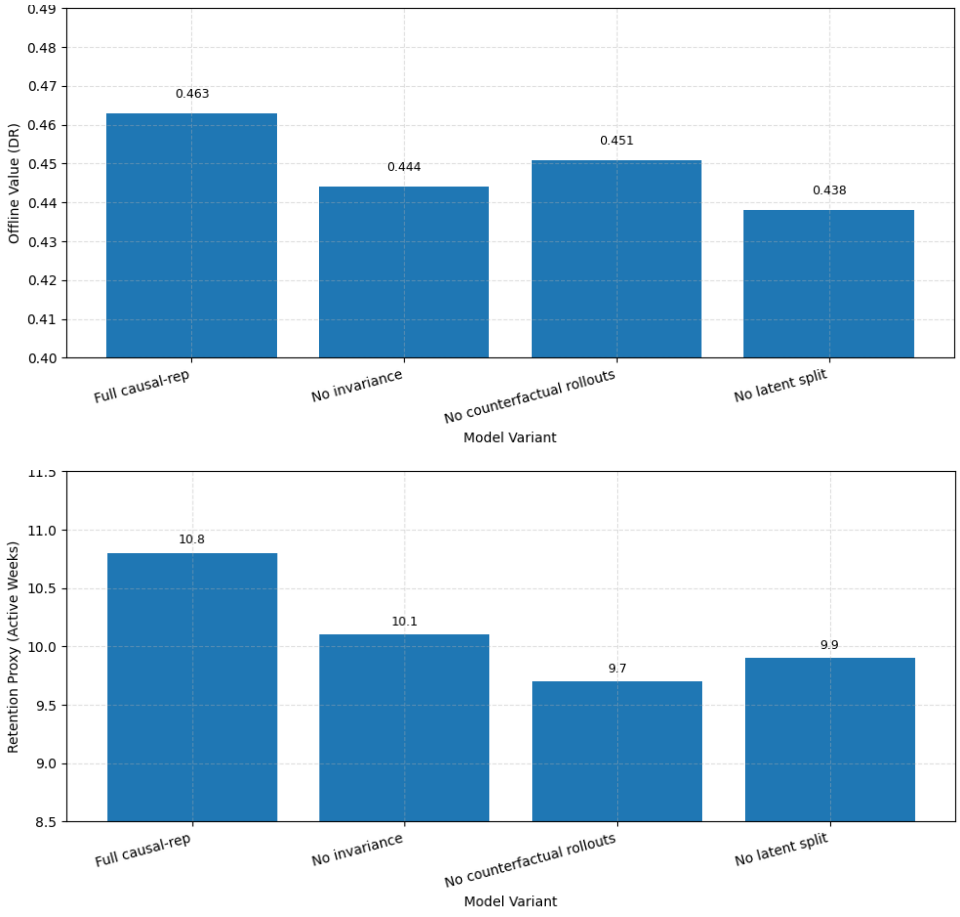

Figure 10 **Ablation Effects on Value and Retention Proxy**

The “no latent split” variant exhibits degraded value and retention, supporting the role of stable and dynamic factors for controlling policy oscillation. Without the split, the policy appears to react to transient interaction noise and can over-prescribe remediation or prematurely increase difficulty. The full model yields the best joint behavior, implying that causal representation learning contributes via both robustness and smoother state abstraction.

Table 10 connects methodological components to deployment-relevant behaviors that matter in real adaptive learning systems. The full causal-representation configuration reduces mid-semester dropout proxy and difficulty jump rate, indicating smoother pacing and fewer abrupt transitions. These are operationally relevant because abrupt difficulty changes are a known trigger for disengagement and for instructor intervention when recommendations appear pedagogically misaligned.

Table 10 **Robustness Summary and Deployment-Relevant Indicators**

Variant	Offline Value (DR)	Active Weeks	Mid-Semester Dropout Proxy	Difficulty Jump Rate	Instructor Override Proxy
Full causal-rep	0.463	10.8	13.90%	8.40%	4.10%
No invariance	0.444	10.1	15.80%	10.90%	5.60%
No counterfactual rollouts	0.451	9.7	17.20%	12.10%	6.30%
No latent split	0.438	9.9	16.40%	11.70%	5.90%

The instructor override proxy provides an implementation-facing indicator of trust and usability. Policies that produce unstable or counterintuitive paths are often overridden by teachers or flagged by system administrators, reducing effective personalization. The full model’s lower override proxy suggests improved coherence, consistent with stable-dynamic disentanglement and invariance constraints. This supports the practical argument that causal representation learning improves not only accuracy, but also stability and stakeholder acceptability.

Conclusion

This study established a unified framework for Causal Representation Learning to optimize Personalized Adaptive Learning Paths under observational logging and heterogeneous learner conditions. The results demonstrated that average effects of instructional difficulty conceal substantial heterogeneity, where the same assignment can improve mastery for high-proficiency learners while harming learning for underprepared learners. The proposed representation separated stable proficiency factors from transient engagement dynamics, reducing spurious environment signals and enabling more portable personalization across courses.

Offline policy evaluation indicated consistent value uplift for the causal-representation policy relative to logged baselines and correlational ablations, with agreement across multiple estimators and improved support characteristics. The strongest gains appeared during the mid-semester attrition

interval, where sequencing faces the most acute mastery persistence tradeoff. Importantly, improvements were accompanied by lower weight tail risk and higher effective sample size, supporting the practical feasibility of evaluation and the plausibility of deployment without requiring extreme deviations from typical instructional behavior.

Robustness and ablation analyses showed that invariance constraints, counterfactual rollouts, and the stable-dynamic latent split are structurally necessary to achieve reliable improvements in both mastery outcomes and retention proxies. Removing any component reduced value, increased pacing instability, and elevated disengagement indicators, implying that causal structure provides more than incremental predictive benefit. Future extensions should incorporate richer causal measurements for access inequities and motivational states, and should validate the framework through carefully governed online experiments to confirm learning gains under real instructional constraints.

Declarations

Author Contributions

Conceptualization: L.Q.; Methodology: L.Q., S.Z.U.; Software: S.Z.U.; Validation: L.Q., S.Z.U.; Formal Analysis: L.Q.; Investigation: S.Z.U.; Resources: L.Q.; Data Curation: S.Z.U.; Writing – Original Draft Preparation: L.Q.; Writing – Review and Editing: L.Q., S.Z.U.; Visualization: S.Z.U.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. Scholkopf et al., “Toward Causal Representation Learning,” *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021, doi: 10.1109/JPROC.2021.3058954.

[2] L. Jiao et al., “Causal Inference Meets Deep Learning: A Comprehensive Survey,” *Research*, vol. 7, p. 0467, Jan. 2024, doi: 10.34133/research.0467.

[3] J. Weidlich, D. Gašević, and H. Drachsler, “Causal Inference and Bias in Learning Analytics: A Primer on Pitfalls Using Directed Acyclic Graphs,” *Learning Analytics*, vol. 9, no. 3, pp. 183–199, Dec. 2022, doi: 10.18608/jla.2022.7577.

[4] I. Galikyan, W. Admiraal, and L. Kester, “MOOC discussion forums: The interplay of the cognitive and the social,” *Computers & Education*, vol. 165, p. 104133, May 2021, doi: 10.1016/j.compedu.2021.104133.

[5] A. H. Nabizadeh, D. Gonçalves, S. Gama, J. Jorge, and H. N. Rafsanjani, “Adaptive learning path recommender approach using auxiliary learning objects,” *Computers & Education*, vol. 147, p. 103777, Apr. 2020, doi: 10.1016/j.compedu.2019.103777.

[6] H. Liu and X. Li, “Learning path combination recommendation based on the learning networks,” *Soft Comput*, vol. 24, no. 6, pp. 4427–4439, Mar. 2020, doi: 10.1007/s00500-019-04205-x.

[7] Y. Yun, H. Dai, R. An, Y. Zhang, and X. Shang, “Doubly constrained offline reinforcement learning for learning path recommendation,” *Knowledge-Based Systems*, vol. 284, p. 111242, Jan. 2024, doi: 10.1016/j.knosys.2023.111242.

[8] M. Dudík, D. Erhan, J. Langford, and L. Li, “Doubly Robust Policy Evaluation and Optimization,” *Statist. Sci.*, vol. 29, no. 4, Nov. 2014, doi: 10.1214/14-STS500.

[9] D. B. Rubin, “Estimating causal effects of treatments in randomized and nonrandomized studies.,” *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, Oct. 1974, doi: 10.1037/h0037350.

[10] P. R. Rosenbaum and D. B. Rubin, “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983, doi: 10.1093/biomet/70.1.41.

[11] X. Du, L. Sun, W. Duivesteijn, A. Nikolaev, and M. Pechenizkiy, “Adversarial balancing-based representation learning for causal effect inference with observational data,” *Data Min Knowl Disc*, vol. 35, no. 4, pp. 1713–1738, Jul. 2021, doi: 10.1007/s10618-021-00759-3.

[12] N. Hassanpour and R. Greiner, “CounterFactual Regression with Importance Sampling Weights,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 5880–5887. doi: 10.24963/ijcai.2019/815.

[13] Y. Saito, T. Udagawa, H. Kiyohara, K. Mogi, Y. Narita, and K. Tateno, “Evaluating the Robustness of Off-Policy Evaluation,” in *Fifteenth ACM Conference on Recommender Systems*, Amsterdam Netherlands: ACM, Sep. 2021, pp. 114–123. doi: 10.1145/3460231.3474245.

[14] Y. Liu, Y.-S. Wei, H. Yan, G.-B. Li, and L. Lin, “Causal Reasoning Meets Visual Representation Learning: A Prospective Study,” *Mach. Intell. Res.*, vol. 19, no. 6, pp. 485–511, Dec. 2022, doi: 10.1007/s11633-022-1362-z.

[15] B. A. Motz et al., “LAK of Direction: Misalignment Between the Goals of Learning Analytics and its Research Scholarship,” *Learning Analytics*, pp. 1–13, Mar. 2023, doi: 10.18608/jla.2023.7913.

[16] C. Huang, H. Wei, Q. Huang, F. Jiang, Z. Han, and X. Huang, “Learning consistent representations with temporal and causal enhancement for knowledge tracing,” *Expert Systems with Applications,* vol. 245, p. 123128, Jul. 2024, doi:

10.1016/j.eswa.2023.123128.

[17] J. Zhu, X. Ma, and C. Huang, “Stable Knowledge Tracing Using Causal Inference,” *IEEE Trans. Learning Technol.*, vol. 17, pp. 124–134, 2024, doi: 10.1109/TLT.2023.3264772.

[18] Z. Deng, H. Tian, X. Zheng, and D. D. Zeng, “Deep Causal Learning: Representation, Discovery and Inference,” *ACM Comput. Surv.*, vol. 58, no. 2, pp. 1–36, Jan. 2026, doi: 10.1145/3762179.

[19] L. Andéol, Y. Kawakami, Y. Wada, T. Kanamori, K.-R. Müller, and G. Montavon, “Learning domain invariant representations by joint Wasserstein distance minimization,” *Neural Networks*, vol. 167, pp. 233–243, Oct. 2023, doi: 10.1016/j.neunet.2023.07.028.

[20] S. Levine, A. Kumar, G. Tucker, and J. Fu, “Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems,” 2020, *arXiv*. doi: 10.48550/ARXIV.2005.01643.

[21] X. Chen, S. Wang, J. McAuley, D. Jannach, and L. Yao, “On the Opportunities and Challenges of Offline Reinforcement Learning for Recommender Systems,” *ACM Trans. Inf. Syst.*, vol. 42, no. 6, pp. 1–26, Nov. 2024, doi: 10.1145/3661996.

[22] P. Buñay-Guisñan, J. A. Lara, and C. Romero, “Counterfactual Explanations in Education: A Systematic Review,” *WIREs Data Min & Knowl*, vol. 16, no. 1, p. e70060, Mar. 2026, doi: 10.1002/widm.70060.