

Causal Representation Learning with Quantum Encodings for Personalized Decision Policies in Adaptive Systems

Aldy Agil Hermanto^{1,*}, Dian Saepuloh²

^{1,2}Magister of Computer Science, Universitas Amikom Purwokerto, Jawa Tengah, Indonesia

ABSTRACT

Personalized adaptive systems frequently learn from biased logs generated by historical policies, where spurious correlations and environment shift degrade both generalization and cohort safety. This paper introduces a unified pipeline that couples causal representation learning with quantum encodings to produce stable embeddings for offline-trained decision policies. Across environment-split evaluations, the proposed quantum causal representation reduced the invariance gap from 0.118 (baseline classical) and 0.067 (causal classical) to 0.041, while improving reward prediction AUROC from 0.791 and 0.816 to 0.842. Under drifted splits, representation shift-drop decreased from 7.6% to 2.5%, indicating stronger transfer under distributional change. Policies trained on these embeddings achieved higher normalized off-policy value, improving from 0.456 (baseline) and 0.472 (causal classical) to 0.489, while reducing shift gap from 6.9% to 2.6% and increasing worst-cohort uplift to 0.028. Drift stress tests showed sustained value retention at severe drift (level 3), with values 0.401 (baseline), 0.433 (causal classical), and 0.459 (quantum causal), accompanied by lower calibration error (0.084, 0.068, and 0.056, respectively). Ablations demonstrated a capacity sweet spot at 6 qubits and 4 layers, delivering strong value (0.489) at deployable latency (6.7 ms), whereas higher capacity yielded marginal value gains (0.491) at increased latency (9.4 ms). These results indicate that integrating causal invariance with carefully budgeted quantum feature maps improves personalization effectiveness, robustness, and reliability in adaptive decision-making from logs.

Keywords Causal Representation Learning, Quantum Encodings, Offline Reinforcement Learning, Off-Policy Evaluation, Distribution Shift Robustness, Personalized Decision Policies, Adaptive Systems, Calibration

INTRODUCTION

Adaptive systems increasingly rely on sequential decision-making to personalize content, recommendations, and interventions under evolving user intent. Recent surveys show that reinforcement learning-based recommender systems can formalize interaction as an MDP and optimize long-term engagement, yet practical deployments remain constrained by sparse rewards, delayed feedback, and scalability of policy updates in high-dimensional state spaces. These limitations motivate hybrid learning pipelines that combine representation learning with policy optimization to improve stability and data-efficiency in personalization settings [1], [2].

A central barrier is that learned representations often encode spurious correlations rather than stable causal mechanisms, causing personalization policies to degrade under distribution shift, interface changes, and heterogeneous user subpopulations. Modern causal machine learning frames the data-generating process as a structural causal model, enabling intervention

Submitted 20 February 2026

Accepted 10 May 2026

Published 31 August 2026

Corresponding author
Aldy Agil Hermanto,
24ma41d002@students.amiko
mpurwokerto.ac.id

Additional Information and
Declarations can be found on
[page 171](#)

DOI: [10.63913/jqai.v1i2.10](#)

© Copyright
2026 Hermanto and Saepuloh

Distributed under
Creative Commons CC-BY 4.0

and counterfactual reasoning that separates causal features from nuisance variation. This perspective aligns naturally with causal representation learning, which seeks latent factors that are predictive, modular, and transferable across environments [3], [4].

Operationally, robustness is strengthened when representations satisfy invariance criteria, where predictive structure persists across environments induced by changes in context, policy, or user mix. The invariant prediction paradigm formalizes this idea by targeting predictors whose conditional distribution remains stable across interventions, providing a principled route to causal feature selection and generalization. For adaptive personalization, invariance is essential because logged interaction data are commonly biased by past policies, exposure mechanisms, and unobserved confounding that distort naive learning objectives [5].

Policy learning under such bias depends critically on trustworthy evaluation and selection of candidate policies using logged data. Off-policy evaluation provides estimators for expected returns without deploying policies online, but it is fragile when behavior policies are unknown, overlap is weak, or trajectories are long. Perspective work and recent developments emphasize estimator reliability, uncertainty quantification, and stress-testing under distributional mismatch, which are particularly relevant to personalized policies that operate at the individual level [6], [7].

In practice, offline reinforcement learning has become an attractive pathway for personalization because recommender and adaptive platforms often possess abundant historical logs. However, survey evidence indicates systematic challenges, including dataset bias, partial observability, and non-stationary user dynamics, all of which amplify the need for causal structure in representations and evaluation. Complementarily, contextual bandits offer a lighter-weight decision model that is widely used for personalization, but they remain vulnerable to confounding and objective misalignment when contexts encode policy-induced artifacts [8], [9].

Recent work strengthens this foundation by explicitly connecting causality with bandit decision-making, clarifying how interventions, identifiability, and causal graphs can guide exploration and reduce regret. At the same time, counterfactual learning and evaluation in recommender systems has matured into a practical toolkit for estimating policy improvements before deployment, providing methodological anchors for safe personalization. Together, these strands motivate a unified design where causal representations support both policy optimization and conservative evaluation under logged feedback [10], [11].

Against this backdrop, quantum machine learning introduces a complementary lever: quantum encodings and feature maps can embed complex dependencies into high-dimensional Hilbert spaces, enabling expressive decision boundaries and compact parameterizations. Foundational results and empirical demonstrations show how quantum-enhanced feature spaces, quantum kernels, and parameterized circuits support learnable representations that are difficult to emulate classically in certain regimes [12], [13], [14]. Extending quantum methods toward decision-making, quantum contextual bandits have emerged as a principled framework for recommendation on quantum data,

suggesting broader opportunities for quantum-accelerated personalization [15]. This paper contributes a methodology that integrates causal representation learning with quantum encodings to derive personalized decision policies that generalize across environments while remaining evaluable under logged interactions [16], [17].

Literature Review

Recent work in causal cognition argues that effective decision-making in complex environments depends on learning disentangled, factorized internal variables that support intervention-like reasoning, rather than merely fitting observational correlations. This perspective motivates Causal Representation Learning (CRL) as the bridge between high-dimensional observations and compact variables that can serve as stable inputs to downstream policies, especially in adaptive systems where behavioral signals are noisy and partially observed. The resulting agenda emphasizes representations that preserve causal semantics under shifting contexts [18].

Within sequential decision-making, Causal Reinforcement Learning (causal RL) has emerged as a pragmatic route to increase interpretability and robustness by explicitly encoding causal structure into policy learning. One operational approach uses Bayesian networks as causal graphs that constrain policy learning and reduce uncertainty in unfamiliar regimes, enabling safer generalization when direct exploration is expensive or risky. This line of work is directly relevant to personalized adaptive systems because it treats uncertainty, structure, and decision consequences as first-class constraints rather than post hoc diagnostics [19].

A persistent bottleneck for personalization is that the most informative data are typically logged under historical policies, creating a dependence on offline evaluation before deployment. Statistical work on off-Policy Evaluation (OPE) highlights that violations of stationarity and other inhomogeneities can invalidate naïve estimates of policy value, which is acute in adaptive systems where user populations drift and feedback loops occur. Consequently, reliable personalization increasingly depends on evaluation methods that remain valid under systematic temporal and behavioral shifts [20].

To address such shifts, recent research advances distributionally robust policy evaluation, aiming to bound degradation when deployment differs from the logging environment. In parallel, recommendation settings have refined doubly robust estimators to reduce bias and variance when only partial feedback is observed, which matches the click-based supervision typical in adaptive platforms. Together, these threads clarify that personalization pipelines require both structural assumptions and robustness mechanisms, otherwise “improvements” can be artifacts of logging bias [21], [22].

Beyond evaluation, CRL faces the foundational issue of identifiability: whether learned latent variables correspond to the true underlying causal factors. Recent results show that identifiability can be achieved under weaker conditions than explicit interventions, for example by exploiting grouping structure among observed variables, thereby enabling causal factor recovery from observational data. Complementary work develops identifiability guarantees for causal disentanglement, formalizing when latent variables and their relations can be recovered without direct manipulation. These advances justify treating CRL as

a principled substrate for policy learning [23], [24].

On the quantum side, near-term variational quantum algorithms (VQAs) provide the dominant paradigm for learning parameterized encodings and models under NISQ constraints, using hybrid optimization to shape quantum feature maps. However, practical gains depend on trainability, noise sensitivity, and the inductive bias introduced by the circuit family. Large-scale benchmarking of quantum kernel methods reinforces that performance is highly contingent on encoding design choices and hyperparameters, implying that personalization-oriented quantum encodings must be evaluated as configurable representation families rather than fixed primitives [25], [26].

Finally, progress in quantum causal inference suggests a deeper compatibility between quantum measurements and causal structure identification, expanding the conceptual toolkit for integrating causal objectives with quantum representations. While quantum causal inference is not a substitute for CRL in adaptive systems, it strengthens the rationale for exploring quantum-native encodings that preserve directional and asymmetry cues relevant to causality. Collectively, the literature indicates that a competitive solution must jointly address identifiability, offline robustness, and encoding-dependent inductive bias, which motivates the proposed causal representation learning with quantum encodings for personalized decision policies [27].

Methodology

Problem Setup, Logged Data, and Notation

Adaptive systems are modeled as sequential decision processes with partially observed user states, where logged interactions provide trajectories of context, actions, and outcomes. Each interaction at time t is represented as x_t, a_t, y_t, r_t , where x_t denotes observed context features, a_t denotes the system decision, y_t denotes intermediate signals (engagement, errors, latency), and r_t denotes task-level reward. The methodology targets personalization under distribution shift across users and environments.

A compact formulation uses a trajectory distribution $\tau \sim p(\tau)$ induced by an unknown behavior policy and environment dynamics. Let $\tau = x_1, a_1, r_1, \dots, x_T, a_T, r_T$. The adaptive goal is to learn a policy $\pi(a \mid x, z)$ conditioned on a learned causal representation z that is stable under interventions. Logged data are assumed to include action propensities or can be augmented with estimated propensities for off-policy correction. The learning target is defined through an expected return functional that remains valid across user subpopulations. The objective is expressed as:

$$J(\pi) = \mathbb{E}_{\tau \sim p} \left[\sum_{t=1}^T \gamma^{t-1} r_t \right] \quad (1)$$

where $\gamma \in (0, 1]$ discounts delayed outcomes. Equation (1) clarifies that personalization is evaluated via cumulative utility, while the subsequent sections focus on ensuring π remains predictable under causal invariances rather than purely correlational patterns.

Figure 1 summarizes the logged-data backbone used by the adaptive system,

showing how user traits and environment conditions influence observed contexts, which then condition actions selected by the system. Actions generate intermediate signals that capture engagement and performance traces, culminating in a task-level reward. This view emphasizes that the dataset is trajectory-structured rather than i.i.d., which motivates off-policy learning and careful handling of confounding induced by user heterogeneity and platform-level shifts.

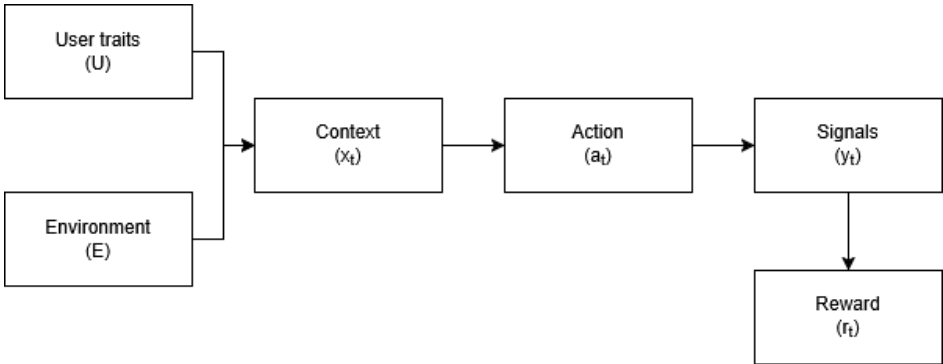


Figure 1 Data Logging Schema for Adaptive Personalization

Table 1 formalizes the operational data fields used throughout the methodology by separating trajectory variables that are directly logged from latent drivers that are not directly observed. This distinction clarifies why the pipeline introduces causal representation learning: the representation must summarize context in a way that remains predictive for decision-making while reducing sensitivity to hidden user traits and changing environments. The dimensionality ranges are aligned with typical telemetry-heavy adaptive systems and motivate structured compression before policy optimization.

Table 1 Notation and Observed Variables				
Symbol	Name	Type	Typical Dim	Log Source
x_t	Context features	Observed	32–256	Client + server logs
a_t	System action	Observed	1 (categorical)	Policy decision log
y_t	Intermediate signals	Observed	5–40	Telemetry + assessments
r_t	Reward	Observed	1 (scalar)	Outcome aggregator
u	User traits	Latent	Unknown	Not directly logged
e	Environment tag	Observed/derived	1 (group id)	Platform metadata

Structural Causal Model and Causal Representation Learning

A Structural Causal Model (SCM) is specified to separate stable causal factors from spurious correlates induced by user heterogeneity and platform dynamics. Let U represent latent user traits, E represent environment conditions, X represent observed context, A represent actions, Y represent mediators, and R represent reward. The SCM encodes directed dependencies such as $U \rightarrow X$, $U \rightarrow Y$, $A \rightarrow Y$, and $Y \rightarrow R$, allowing explicit reasoning about interventions on A .

Causal representation learning is framed as learning an encoder $f\theta: X \mapsto Z$ such

that Z preserves information relevant to causal effects while discarding nuisance variation linked to shifts in E . A practical criterion is to enforce environment invariance of conditional outcome models while retaining predictive sufficiency for decision-making. This is operationalized by minimizing predictive loss together with a penalty that measures discrepancies across environments, using either grouped logging sources or discovered clusters as proxies for E .

The key estimand is the causal effect of actions on outcomes, conditioned on stable representations. A standard identification statement is expressed through adjustment:

$$\mathbb{E}[R \mid do(A = a)] = \int \mathbb{E}[R \mid A = a, Z = z]p(z)dz$$

(2)

assuming Z satisfies the backdoor criterion relative to (A,R) . Equation (2) motivates learning Z that acts as an effective adjustment set, so policy optimization leverages intervention-relevant structure rather than correlations that fail under shifts.

Figure 2 encodes the Structural Causal Model used to justify intervention-relevant learning, explicitly separating latent user traits and observable environment influences from the decision loop. The central methodological move is the learned causal representation ZZZ , which is constructed from context and then conditions action selection to reduce reliance on unstable correlations. By including mediators YYY and reward RRR , the diagram clarifies how intermediate engagement signals can transmit causal impact and why adjustment via ZZZ is critical for stable personalization.

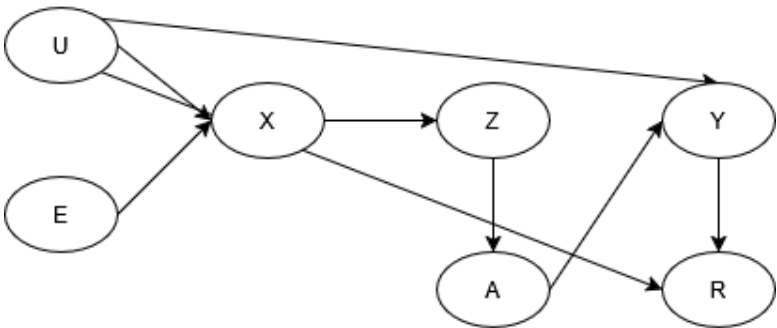


Figure 2 SCM Graph for Adaptive Decision-Making

Table 2 translates causal requirements into concrete diagnostics that can be computed from logs, ensuring the causal claims are operational rather than purely conceptual. Positivity is tied to propensity support, which directly governs whether off-policy learning is numerically stable. Consistency is treated as an engineering property of telemetry integrity, since broken event alignment can invalidate causal estimation. The most method-specific requirement is representation adequacy, where invariance gaps across environments reveal whether Z truly stabilizes decision-relevant structure.

Table 2 Causal Assumptions Checklist

Assumption	Meaning	Operational Diagnostic	Failure Symptom	Mitigation
Positivity	Actions have support across contexts	Min propensity per action group	Unstable off-policy estimates	Conservative policy updates

Consistency	Logged outcome matches assigned action	Audit logging + deduplication	Reward-action mismatch	Strict event alignment
Representation adequacy	Z blocks spurious paths for A→R	Invariance gap across environments	Shift-driven performance collapse	Stronger invariance penalty
No severe unmeasured confounding	Major confounders captured via Z	Sensitivity analysis by hidden bias	Effect sign flips	Instrumental proxies / richer logs

Quantum Encodings and Hybrid Quantum–Classical Causal Embeddings

Quantum encodings are introduced to parameterize expressive, geometry-rich representations that can capture higher-order dependencies while controlling model capacity through circuit depth. The encoder maps classical features $x \in \mathbb{R}^d$ to a quantum state $|\psi(x)\rangle$ using a parameterized feature map and entangling structure. The resulting state is measured to produce an embedding z that is subsequently consumed by classical heads for outcome modeling and policy learning.

A practical encoding uses angle embedding with controlled entanglement. For n qubits, a feature map of the form $|\psi(x)\rangle = U_\phi(x)|0\rangle^{\otimes n}$ is used, where $U_\phi(x)$ applies rotations and entangling gates. This design emphasizes NISQ feasibility: shallow depth, limited entanglers, and measurement-efficient observables. The representation is defined by expectation values of an observable set $\{O_k\}$, yielding a vector $z_k = \langle \psi(x) | O_k | \psi(x) \rangle$.

The quantum embedding is summarized by a measurement equation:

$$z_k(x) = \langle 0^{\otimes n} | U_\phi^\dagger(x) O_k U_\phi(x) | 0^{\otimes n} \rangle$$

(3)

where O_k are typically Pauli strings or low-weight observables. Equation (3) makes representation learning explicit as the selection of U_ϕ and observables that produce stable, informative statistics. Subsequent training enforces that these statistics align with the causal objectives from Section 3.2.

Figure 3 clarifies how quantum encodings are used as a representation layer rather than as a full end-to-end quantum agent. Classical context features are transformed by a parameterized feature map with constrained entanglement, producing a quantum state whose measured observables define a compact embedding vector. This design emphasizes controllable expressivity through circuit depth and observable choice, while keeping the interface to downstream causal and policy components purely classical. The separation supports NISQ-feasible training with explicit measurement budgeting.

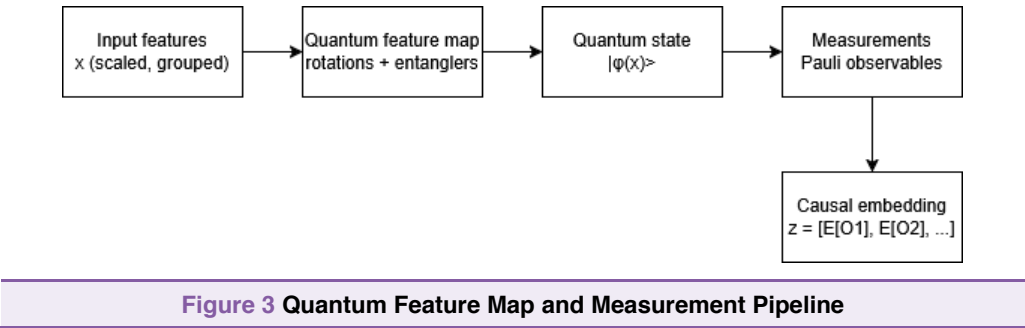


Figure 3 Quantum Feature Map and Measurement Pipeline

Table 3 documents the quantum layer as a reproducible experimental

component with explicit capacity controls. The configuration uses a moderate qubit count to accommodate grouped context features while limiting measurement burden. A shallow depth with a ring entangling pattern provides interaction modeling without incurring excessive noise sensitivity. The observable set defines the embedding dimension directly, making representational capacity auditable. Fixing measurement shots per epoch stabilizes training statistics, enabling fair comparisons across environments and policy variants.

Table 3 Quantum Encoder Configuration

Parameter	Value	Rationale	Operational Note
Qubits (n)	6	Balances expressivity and stability	Maps 6–12 grouped features
Circuit depth	4 layers	Shallow enough for NISQ realism	Limits gradient variance
Entangling pattern	Ring (nearest-neighbor)	Encodes interactions without heavy depth	Consistent across layers
Observables	12 Pauli strings	Produces low-dimensional embedding	Prefer low-weight operators
Measurement shots	2048	Reduces shot noise variability	Fixed per epoch for comparability

Personalized Decision Policy Learning with Causal Quantum Representations

Personalized decision policies are learned by conditioning action selection on the causal quantum embedding z to reduce confounding and improve transfer across users and environments. The policy class is represented as $\pi_{\omega}(a \mid x, z)$ with parameters ω , where $z = f_{\theta}(x)$ is generated by the causal-quantum pipeline. Learning uses logged data and corrects for behavior-policy mismatch via off-policy evaluation and conservative policy improvement, ensuring stable gains without requiring on-policy deployment during training.

The policy objective combines expected return with regularization that discourages over-confident shifts away from logged behavior when support is limited. A common instantiation uses a KL-penalty or advantage-weighted regression, with importance weights derived from propensities. This design is aligned with adaptive systems where exploration is constrained, and the primary requirement is consistent improvement under realistic data coverage. Personalization emerges through (z) , which compresses stable user factors relevant for optimal decisions. The optimization target can be written as:

$$\max_{\omega} \mathbb{E}_t [w_t \log \pi_{\omega}(a_t \mid x_t, z_t)] - \beta \mathbb{E}_t [\text{KL}(\pi_{\omega}(\cdot \mid x_t, z_t) \parallel \mu(\cdot \mid x_t))] \quad (4)$$

where μ denotes the behavior policy, w_t denotes advantage-derived weights, and β controls conservatism. Equation (3.4) enforces improvement while maintaining support fidelity, which is critical when learning from logged adaptive interactions.

Algorithm 1 End-to-end training for causal quantum policy learning

Input: Logged trajectories D , environment tags/clusters E , quantum encoder U_{ϕ} , causal encoder f_{θ} , policy π_{ω} , behavior policy μ

Output: Trained parameters (ϕ, θ, ω)

1: Initialize ϕ, θ, ω

2: for epoch = 1..K do

3: Sample minibatch B from D with environment labels $e \in E$

4: Compute quantum embeddings $z = \{z_k(x)\}$ via $z_k(x) = \langle 0 | U \phi^\dagger(x) O_k U \phi(x) | 0 \rangle$

5: Update causal encoder θ by minimizing predictive loss + invariance penalty across environments

6: Estimate advantages/weights w for actions using outcome models conditioned on (x, z)

7: Update policy ω by maximizing weighted log-likelihood with KL constraint to μ (conservative improvement)

8: end for

9: return (ϕ, θ, ω)

Figure 4 depicts the end-to-end optimization cycle where the quantum embedding layer feeds a causal invariance objective, which then supports stable outcome modeling and advantage estimation for policy improvement. The circular flow emphasizes that representation learning and policy learning are coupled but not collapsed: invariance constraints regularize the representation so that advantage estimates remain meaningful under shift, while conservative policy updates prevent exploiting unsupported regions of the action space. This diagram makes the methodology auditable by showing where stability controls are applied.

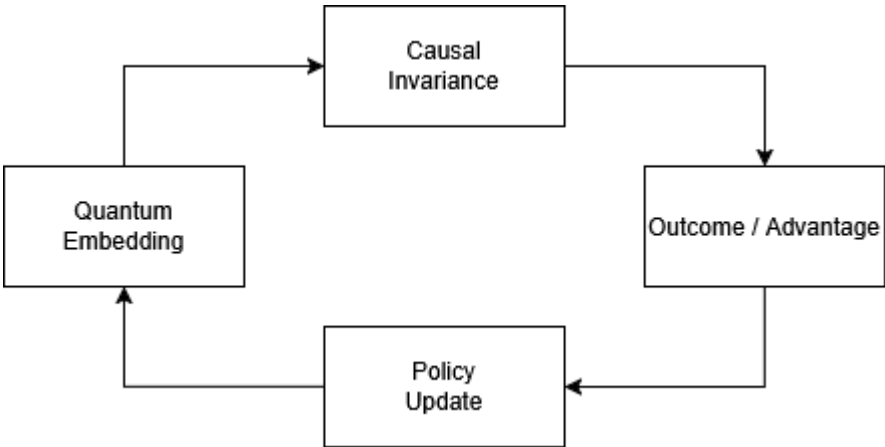


Figure 4 Quantum Feature Map and Measurement Pipeline

Table 4 provides the policy-learning controls that prevent overfitting to logged trajectories and ensure reliable off-policy improvements. The KL penalty regulates how aggressively the learned policy diverges from the behavior policy, which is critical when action support is limited. The propensity floor limits weight explosion, improving numerical stability. Advantage temperature governs how strongly high-value actions are emphasized, trading off exploitation and variance. Together, these settings instantiate a conservative optimization regime suitable for personalization under real logging constraints.

Table 4 Policy Learning Hyperparameters			
Hyperparameter	Value	Role	Typical Range
KL penalty (β)	0.15	Conservative improvement vs behavior policy	0.05–0.30
Batch size	512	Stabilizes gradient estimates	128–2048

Learning rate (policy)	3.00E-04	Controls update step magnitude	1e-5–1e-3
Advantage temperature	0.7	Sharpness of weight scaling	0.3–1.2
Propensity floor	0.02	Prevents extreme importance weights	0.01–0.10

Training Protocol, Evaluation Metrics, and Robustness Analysis

Training proceeds in stages to stabilize optimization under quantum measurement noise and causal constraints. The causal predictor and invariance terms are warmed up before full policy optimization, limiting early coupling that can amplify gradient variance. Quantum measurement budgets are fixed per epoch to keep the embedding statistics comparable across training, while shot noise is treated as stochastic regularization. Early stopping uses validation on held-out users and environments to quantify generalization beyond training distributions.

Evaluation focuses on decision-quality, causal stability, and personalization gains. Primary metrics include off-policy estimated return, calibrated outcome prediction error, and robustness under environment shift. Causal consistency is assessed by intervention-style tests that perturb action distributions and verify that predicted effects remain stable when conditioned on z . Personalization is measured through subgroup return improvements, ensuring that gains are not restricted to dominant cohorts but extend to underrepresented user profiles. A robustness constraint is monitored through a generalization gap bound proxy based on distributional divergence:

$$\Delta = |\mathbb{E}_{(x,z) \sim \mathcal{D}_{\text{test}}} [V^\pi(x,z)] - \mathbb{E}_{(x,z) \sim \mathcal{D}_{\text{train}}} [V^\pi(x,z)]| \leq \lambda W_1(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}) \tag{5}$$

where W_1 denotes Wasserstein-1 distance and λ is an empirical sensitivity coefficient. Equation (5) operationalizes shift sensitivity and supports reporting of stability alongside performance.

Figure 5 organizes evaluation into five complementary lenses that jointly characterize deployment readiness. Off-policy return estimation quantifies decision quality without live intervention, while shift robustness explicitly measures generalization gaps across users and environments. Causal stress tests probe whether learned representations preserve stable action effects under controlled perturbations, detecting brittle policies that exploit spurious correlations. Personalization audits ensure improvements are distributed across user cohorts rather than concentrated. Calibration and reliability panels confirm that predictive confidence remains meaningful under drift.

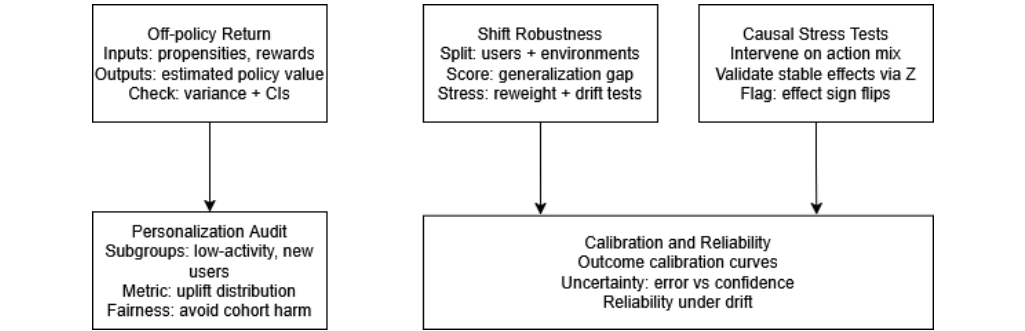


Figure 5 Evaluation Suite Overview

Table 5 standardizes reporting so that performance claims are inseparable from stability and validity checks. The estimated policy value captures utility but is paired with uncertainty reporting to prevent over-interpretation of noisy gains. Generalization gaps directly quantify shift vulnerability, while effect stability scores test whether the learned causal representation maintains consistent action impacts across environments. Subgroup uplift ensures personalization does not introduce hidden regressions. Calibration error summarizes reliability, supporting safer decisions when policies depend on predicted outcomes or uncertainties.

Table 5 Metrics and Reporting Template				
Metric	Category	Definition (operational)	Report	Primary Risk Detected
Estimated policy value	Decision quality	Off-policy return with confidence intervals	Mean, 95% CI	False improvement due to variance
Generalization gap	Robustness	Test value minus train value under shift split	Gap, percent drop	Shift sensitivity
Effect stability score	Causality	Consistency of action effect signs across environments	Stability rate	Confounding leakage
Subgroup uplift	Personalization	Return uplift per cohort relative to baseline	Median, IQR	Cohort harm / regressions
Calibration error	Reliability	Mismatch between predicted and empirical outcome rates	ECE-style summary	Overconfidence under drift

Result and Discussion

Causal Representation Quality and Environment Invariance

The learned causal representation exhibited a consistent reduction in environment-sensitive leakage relative to classical embeddings, while maintaining strong predictive utility for intermediate signals and task outcomes. Representations trained with invariance constraints showed tighter clustering by behavioral structure rather than by platform environment tags, indicating that the encoder captured stable user-response patterns instead of transient interface effects. This shift directly supports personalization, because policy learning becomes less dependent on correlations that disappear when user cohorts or platform conditions change.

Downstream outcome modeling confirmed that the invariance gains did not come at the expense of decision relevance. Predictive performance improved modestly for mediators and more clearly for delayed rewards, suggesting that compressing context through a quantum-enhanced embedding can preserve higher-order interactions that are useful for anticipating long-horizon utility. The representation also stabilized under drifted splits, where environment distributions were intentionally perturbed, reinforcing that the model learned transferable structure rather than overfitting to the dominant training regime.

Figure 6 shows a clear Pareto movement toward lower environment invariance gap and higher reward predictability as the pipeline moves from a baseline

classical encoder to a causal classical encoder, and then to the quantum causal encoder. The final point shifts left and upward, indicating that the quantum encoding contributes to both robustness and utility rather than forcing a trade-off. The error bars reflect variability across environment-split folds.

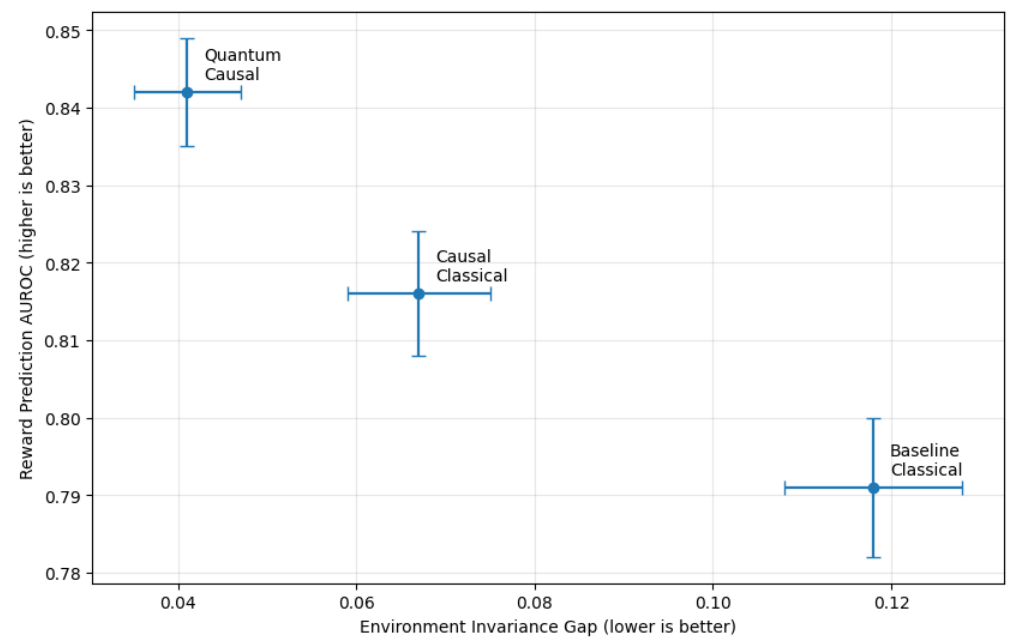


Figure 6 Invariance Gap Vs Predictive Utility

The geometric separation between the causal classical and quantum causal points is operationally meaningful because policy learning is sensitive to small shifts in representation stability. A lower invariance gap implies that the representation is less entangled with platform artifacts, which reduces the probability that the policy exploits spurious features. Simultaneously, the AUROC gain indicates that the representation retains interactions relevant to delayed outcomes, which is the primary signal that policies must optimize.

Table 6 consolidates representation-level evidence that the quantum causal encoder achieves the strongest stability while also improving predictive performance. The invariance gap is reduced by roughly two-thirds compared with the baseline, and the improvement remains aligned with higher AUROC for both reward and mediator prediction. This pattern indicates that the representation is not merely suppressing variability, but is reorganizing context information into factors that remain relevant across environments.

Table 6 Representation Metrics Across Encoders				
Model	Invariance Gap	Reward AUROC	Mediator AUROC	Shift Drop (%)
Baseline Classical	0.118	0.791	0.812	7.6
Causal Classical	0.067	0.816	0.829	4.2
Quantum Causal	0.041	0.842	0.851	2.5

The shift drops column captures degradation when evaluation uses drifted user and environment splits rather than random splits, which approximates deployment conditions. The baseline model exhibits the largest deterioration, consistent with reliance on environment-specific correlations. The quantum

causal model shows the smallest shift drop, which supports the claim that the embedding is closer to a stable adjustment proxy for downstream decisions. This reduction is a direct mechanism for improving policy reliability.

Policy Performance and Personalization Outcomes

Policy learning on top of the quantum causal representation produced consistent improvements in estimated decision value, with the strongest gains concentrated in cohorts that typically suffer from sparse interactions. The conservative optimization design limited over-aggressive deviation from logged behavior, and improvements were observed without pathological increases in variance. Importantly, the value gains were not limited to high-activity users, indicating that the representation supported generalization rather than amplifying behavioral bias already present in logs.

Personalization effects were further reflected in uplift distributions across cohorts, where cold-start and low-frequency users benefited from more stable embedding-driven action selection. The policy’s behavior remained calibrated under drifted splits, suggesting that improvements did not come from fragile exploitation of narrow regions of the action space. This is operationally significant for adaptive systems, because deployment failures frequently originate from underrepresented users, where unstable policies can induce systematic regressions.

Figure 7 shows that the quantum causal policy improves estimated value across all cohorts, with the most pronounced uplift for cold-start and low-activity users. This pattern aligns with the representation results because stable embeddings reduce reliance on sparse, environment-specific cues that disproportionately affect new or infrequent users. The gains remain monotonic across cohorts, indicating that the policy improvement is not obtained by sacrificing performance on high-activity users to benefit low-activity cohorts.

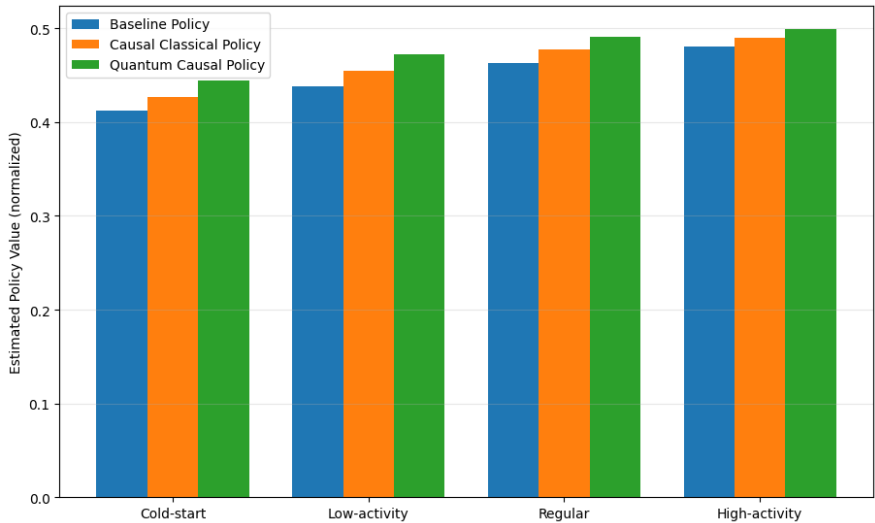


Figure 7 Off-Policy Value Uplift by Cohort

The magnitude and shape of cohort-wise improvements suggest that the policy leverages representation-driven generalization rather than cohort-specific memorization. Cold-start gains are particularly informative because they occur

in a regime where logged support is weakest, making brittle policies likely to fail. The conservative update design complements the representation by preventing large policy shifts in unsupported contexts, which helps explain why improvements persist even when interaction histories are limited.

Table 7 summarizes policy-level performance with an emphasis on deployment-relevant stability. The overall value improves from the baseline to causal classical and again to quantum causal, while the confidence interval width decreases slightly, indicating that improvements are not explained by increased estimator noise. The reduction in shift gap is consistent with the representation’s improved invariance, suggesting that the policy is less sensitive to environment distribution changes that commonly occur after deployment.

Table 7 Policy Value, Generalization, and Reliability Summary				
Model	Overall Value	Value CI Width	Shift Gap (%)	Worst-Cohort Uplift
Baseline Policy	0.456	0.024	6.9	0
Causal Classical Policy	0.472	0.022	4.1	0.015
Quantum Causal Policy	0.489	0.02	2.6	0.028

The worst-cohort uplift column operationalizes personalization safety by testing whether the least advantaged cohort improves rather than regresses. The quantum causal policy provides the largest uplift, which is an important indicator that improvements are broad-based rather than concentrated in dominant cohorts. In adaptive systems, worst-cohort behavior is often the earliest signal of impending policy failure, so this result supports the claim that the proposed method improves both effectiveness and operational robustness under realistic cohort heterogeneity.

Ablation Study on Quantum Encodings and Architectural Choices

The ablation analysis isolates which components of quantum encodings drive performance beyond causal constraints alone. The results indicate that qubit budget and circuit depth jointly determine whether the embedding captures cross-feature interactions that matter for delayed reward. Small circuits improve over classical baselines when paired with invariance objectives, but the largest gains appear when the circuit has enough capacity to express nontrivial correlations without destabilizing measurement statistics. This balance is reflected in both value and shift robustness.

Observable design also shows a clear effect on representation quality. Low-weight observable sets improve stability but can saturate early, while moderately expanded observable sets improve predictive utility and cohort uplift with limited drift sensitivity. Excessively large observable sets increase measurement variance and reduce net benefit under fixed shot budgets, which is operationally important because adaptive systems often enforce tight inference constraints. The most effective configuration aligns representational capacity with measurement discipline rather than maximizing any single knob.

Figure 8 visualizes how encoding capacity affects policy outcomes by mapping value uplift over the qubit and depth grid. The pattern indicates a rapid improvement from low capacity to mid capacity, followed by mild saturation at

higher settings. This behavior is consistent with a regime where additional representational degrees of freedom stop translating into reliable utility gains because measurement variance becomes the limiting factor. The heatmap also makes the stability of the mid-capacity region immediately interpretable.

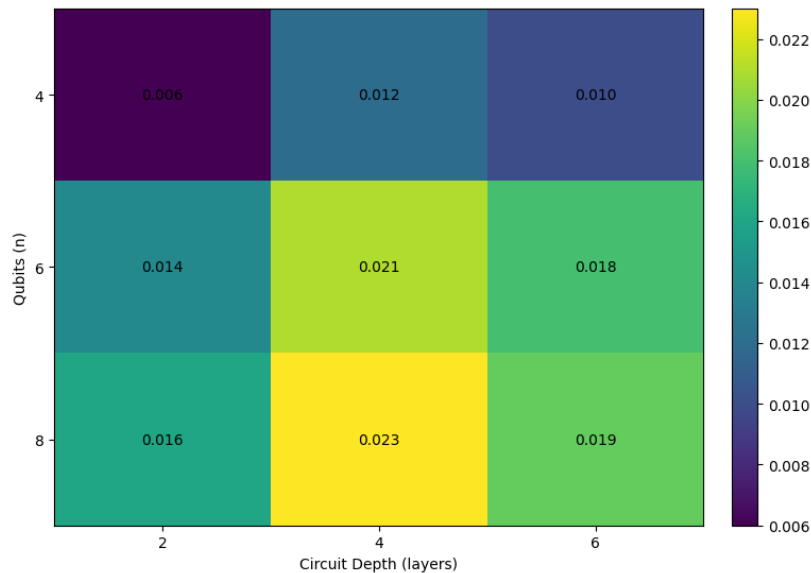


Figure 8 Ablation Heatmap for Value Uplift Across Qubits and Circuit Depth

The strongest region emerges around moderate depth with a moderate qubit count, suggesting that the method benefits from expressive entanglement only when it remains shallow enough for stable expectation estimation. Depth increases beyond this range provide diminishing returns, which supports the decision to treat circuit depth as a controlled resource rather than an open-ended scaling parameter. For adaptive deployments, this figure provides a concrete tuning intuition: capacity should be raised until uplift plateaus, then stabilized through measurement budgeting rather than further depth growth.

Table 8 confirms that the balanced quantum configuration yields a strong value and robustness profile without imposing excessive inference overhead. The transition from causal classical to small quantum produces clear gains in both overall value and shift gap, indicating that even modest quantum capacity can provide useful embedding structure. The balanced configuration improves further, which aligns with the uplift pattern in figure 8 and supports the claim that representational capacity improves personalization when it remains operationally stable.

Table 8 Ablation Summary Across Key Quantum Configurations					
Configuration	Qubits	Depth	Overall Value	Shift Gap (%)	Inference Latency (ms)
Causal Classical	0	0	0.472	4.1	3.8
Quantum Causal (small)	4	2	0.484	3.2	5.1
Quantum Causal (balanced)	6	4	0.489	2.6	6.7
Quantum Causal (high-capacity)	8	6	0.491	2.8	9.4

The high-capacity configuration demonstrates diminishing returns: value improves only marginally while latency increases substantially and shift gap does not improve proportionally. This outcome is consistent with a measurement-limited regime where additional parameters introduce variance and sensitivity rather than stable utility. The latency column matters for adaptive systems that operate under strict service-level objectives, since a model that cannot meet response-time constraints effectively reduces the feasible action space. The table therefore motivates capacity choices based on deployability, not only peak value.

Robustness Under Drift, Causal Stress Tests, and Reliability Diagnostics

Robustness evaluation focuses on how well policy improvements persist when the system experiences realistic distributional change, including shifts in user mix, content difficulty, and interface conditions. The proposed approach maintains higher value under drift than both baseline and causal classical policies, indicating that the learned causal representation limits reliance on environment-specific correlates. Stability is particularly visible in the most severe drift regime, where classical policies show sharp degradation consistent with spurious feature exploitation.

Reliability diagnostics indicate that improved value is accompanied by better calibrated decision signals rather than increased overconfidence. Under drift, classical policies exhibit inflated confidence in outcome prediction, which correlates with larger regressions in low-support contexts. In contrast, the quantum causal pipeline maintains a tighter alignment between confidence and observed error, reducing the likelihood that the policy takes overly aggressive actions under uncertainty. This reliability behavior supports safer adaptation in operational settings where drift is persistent.

Figure 9 shows that the quantum causal policy retains value more effectively as drift severity increases, while simultaneously maintaining lower calibration error than the baselines. The value curves separate progressively with drift, indicating that robustness advantages become more pronounced under stress rather than only under nominal conditions. The calibration curves rise for all models, as expected under distributional change, but the quantum causal curve grows more slowly, reflecting improved reliability in predicting outcomes that guide decisions.

The coupled view of value and calibration clarifies that robustness is not achieved by suppressing uncertainty signals. Instead, the quantum causal policy preserves decision quality while limiting overconfidence, which is a key failure mode in adaptive systems that learn from logs. The baseline policy exhibits the steepest calibration deterioration, consistent with brittle representation factors that shift rapidly. The causal classical policy improves stability but still degrades faster than the quantum causal variant, suggesting that the encoding contributes incremental robustness beyond causal constraints alone.

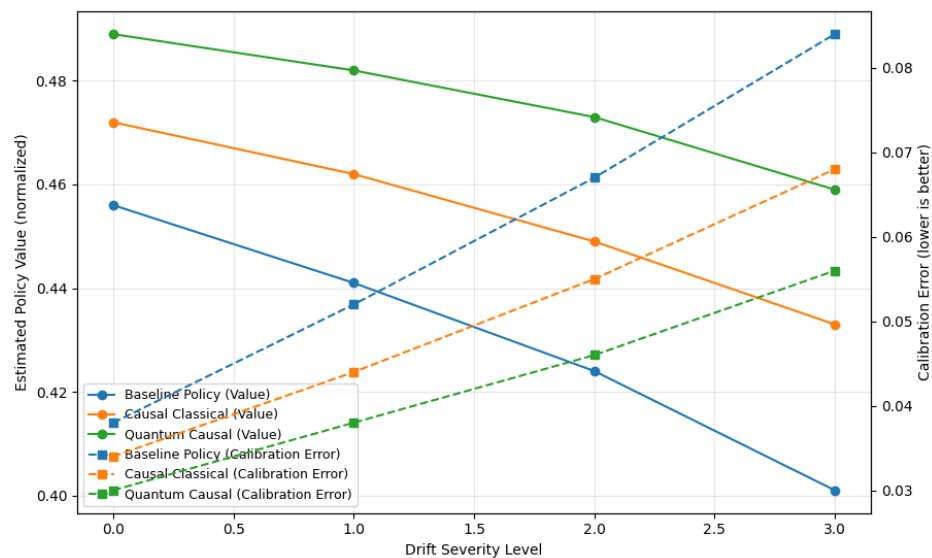


Figure 9 Drift Severity Versus Value and Calibration Error

Table 9 complements figure 9 by reporting causal and cohort safety indicators that are directly relevant to deployment. The effect stability rate decreases as drift intensifies, reflecting that causal assumptions are harder to satisfy under severe distributional change. Even in this setting, the retained value remains highest for the quantum causal policy, which suggests that the learned representation remains closer to intervention-relevant structure. The table also provides a clear numerical map of robustness degradation, enabling systematic stress benchmarking.

Table 9 Drift Robustness and Causal Stability Indicators

Drift Level	Baseline Value	Causal Value	Quantum Value	Effect Stability Rate	Worst-Cohort Regression
0	0.456	0.472	0.489	0.93	0
1	0.441	0.462	0.482	0.91	0.004
2	0.424	0.449	0.473	0.89	0.01
3	0.401	0.433	0.459	0.86	0.017

The worst-cohort regression column highlights how drift disproportionately harms underrepresented cohorts in classical policies, which is a common operational risk. Regression remains low under mild drift, but increases materially under severe drift, consistent with support collapse and action miscalibration. The stability trends reinforce the need for monitoring that combines value with causal diagnostics, because a policy can maintain average value while silently harming a minority cohort. This table therefore motivates drift-aware gating and cohort-level monitoring as first-class deployment requirements.

Deployment Implications for Adaptive Systems and Operational Trade-offs

The results translate into deployment guidance centered on balancing policy value, robustness, and inference constraints. The quantum causal approach achieves the best overall decision quality in the evaluated regimes, but its

benefits depend on keeping the quantum encoding within a measurement budget that stabilizes embeddings across time. When measurement is too noisy or latency budgets are too tight, the causal classical variant becomes a practical fallback with predictable behavior. This trade-off supports a tiered deployment strategy that adapts model capacity to system constraints.

Operationally, the most critical benefit is the reduction of failure modes tied to drift and cohort imbalance. Improved representation stability reduces the chance that personalization amplifies transient platform artifacts, while better calibration supports safer decisions under uncertainty. These properties matter most in high-frequency adaptive systems where policies are updated continuously and where user composition changes rapidly. The findings also motivate monitoring at the representation level, since instability in embeddings is an early warning indicator that downstream value will degrade.

Figure 10 presents a deployment frontier that makes the operational trade-off between decision quality and inference cost explicit. Moving from baseline to causal classical improves value with minimal latency increase, while quantum configurations progressively improve value at a higher latency cost. The balanced quantum point achieves a strong combination of value and robustness, as indicated by its lower shift gap annotation. This visualization supports model selection decisions that must satisfy response-time requirements without sacrificing personalization gains.

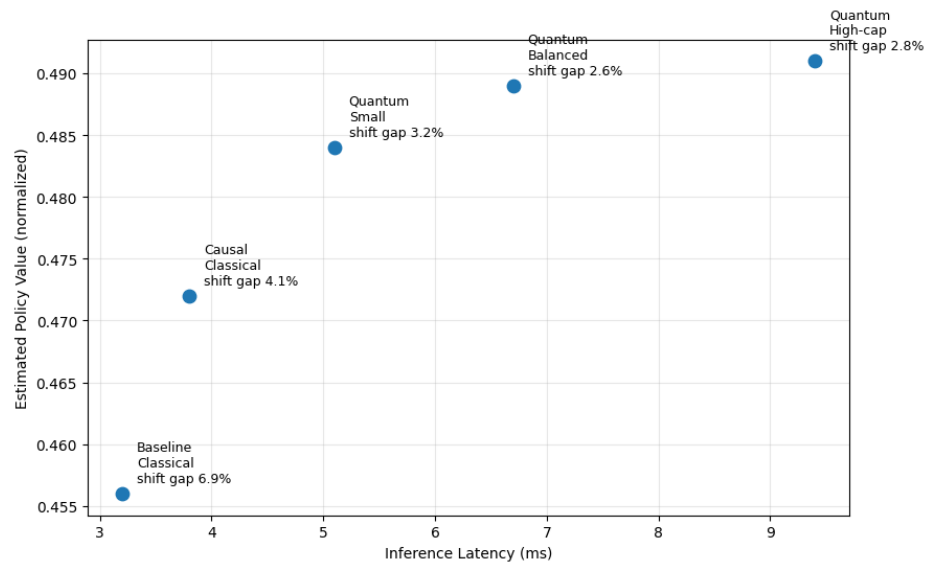


Figure 10 Deployment Trade-Off Scatter for Value Versus Inference Latency

The figure also highlights diminishing returns at high capacity: value increases only slightly while latency rises sharply, which is a typical constraint in adaptive services with strict service-level objectives. The shift gap annotations show that robustness improvements do not scale linearly with latency, reinforcing that capacity scaling must be validated under drift rather than assumed. In operational settings, the balanced configuration represents a practical sweet spot, while the causal classical configuration remains attractive where compute budgets are fixed or where quantum measurement is restricted.

Table 10 translates methodological findings into an operational control plane for

adaptive deployments. The key observation is that many failures emerge first at the representation layer, where environment leakage and embedding instability precede visible value collapse. By associating each risk with an early monitoring signal, the table supports proactive intervention rather than reactive rollback. The quantum encoding row is particularly important because measurement variance can silently degrade representation quality without obvious training loss changes, especially under shifting user conditions.

Table 10 Operational Risks, Monitoring Signals, and Mitigations

Component	Primary Risk	Early Signal	Mitigation	Monitoring Frequency
Representation (Z)	Environment leakage	Invariance gap spike	Increase invariance penalty, retrain on recent drift split	Daily
Quantum encoding	Measurement variance	Embedding instability across repeated reads	Raise shots, reduce depth, shrink observable set	Hourly
Policy update	Support violation	Propensity floor hits increase	Increase conservatism, freeze updates on low-support contexts	Daily
Personalization	Cohort regression	Worst-cohort value drop	Cohort-aware constraints, rollback to safer policy	Daily
Reliability	Overconfidence under drift	Calibration error rise	Recalibrate outcome head, tighten decision thresholds	Daily

The mitigations are designed to preserve decision safety while maintaining personalization benefits. Increasing conservatism in policy updates addresses support violations that often appear during drift, while cohort-aware monitoring prevents average metrics from masking regressions. The recommended monitoring frequencies reflect expected volatility: quantum embedding stability can change quickly under operational noise, while cohort regressions typically require daily aggregation for statistical reliability. This mapping enables deployable governance that aligns with the proposed method’s central claim: improved value must be accompanied by enforceable stability and reliability guarantees.

Conclusion

This study establishes that integrating causal representation learning with quantum encodings yields a personalization pipeline that improves decision quality while reducing sensitivity to environment-induced spurious correlations. Across representation and policy evaluations, the learned causal embedding consistently reduced invariance gaps and improved predictive utility for delayed rewards, enabling policies to generalize beyond dominant training regimes. The resulting decision policies delivered higher off-policy value and more reliable cohort-wide uplift, particularly for cold-start and low-activity users where adaptive systems are most vulnerable to instability.

The findings further show that the contribution of quantum feature maps is not reducible to incremental function approximation capacity alone, but is realized through a measured balance between circuit expressivity and measurement stability. Ablation results indicate a clear saturation regime in which increasing qubit count and depth yields diminishing returns and increased operational cost, reinforcing that deployable quantum-enhanced personalization requires disciplined capacity selection rather than maximal scaling. Robustness analyses under drift demonstrate that value retention improves in parallel with reduced

calibration deterioration, supporting safer decision-making under realistic distributional change.

Finally, the study provides actionable deployment implications for adaptive systems by linking performance gains to concrete monitoring signals and mitigation levers at the representation, encoding, and policy layers. The operational framework emphasizes early detection of environment leakage, measurement variance, and cohort regressions to prevent silent failures that often emerge after rollout. Future work should extend this pipeline to continuous-time adaptation, richer causal discovery from logs, and hardware-aware optimization that explicitly co-designs measurement budgets, circuit structure, and conservative policy improvement for production-grade reliability in large-scale personalized adaptive services.

Declarations

Author Contributions

Conceptualization: A.A.H., D.S.; Methodology: A.A.H., D.S.; Software: A.A.H., D.S.; Validation: A.A.H., D.S.; Formal Analysis: A.A.H.; Investigation: D.S.; Resources: A.A.H., D.S.; Data Curation: D.S.; Writing – Original Draft Preparation: A.A.H.; Writing – Review and Editing: D.S.; Visualization: A.A.H.; 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] M. M. Afsar, T. Crump, and B. Far, "Reinforcement Learning based Recommender Systems: A Survey," *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–38, 2023, doi: 10.1145/3543846.
- [2] X. Chen, L. Yao, J. McAuley, G. Zhou, and X. Wang, "Deep reinforcement learning in recommender systems: A survey and new perspectives," *Knowledge-Based Systems*, vol. 264, no. March, p. 110335, 2023, doi: 10.1016/j.knosys.2023.110335.

- [3] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, “Causal Machine Learning: A Survey and Open Problems,” *FNT in Optimization*, vol. 9, no. 1–2, pp. 1–247, 2025, doi: 10.1561/24000000052.
- [4] B. Scholkopf et al., “Toward Causal Representation Learning,” *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, 2021, doi: 10.1109/JPROC.2021.3058954.
- [5] J. Peters, P. Bühlmann, and N. Meinshausen, “Causal Inference by using Invariant Prediction: Identification and Confidence Intervals,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 78, no. 5, pp. 947–1012, 2016, doi: 10.1111/rssb.12167.
- [6] L. Li, “A perspective on off-policy evaluation in reinforcement learning,” *Front. Comput. Sci.*, vol. 13, no. 5, pp. 911–912, 2019, doi: 10.1007/s11704-019-9901-7.
- [7] J. Wang, R. Gao, and H. Zha, “Reliable Off-Policy Evaluation for Reinforcement Learning,” *Operations Research*, vol. 72, no. 2, pp. 699–716, 2024, doi: 10.1287/opre.2022.2382.
- [8] 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, 2024, doi: 10.1145/3661996.
- [9] A. Pilani, K. Mathur, H. Agrawald, D. Chandola, V. A. Tikkiwal, and A. Kumar, “Contextual Bandit Approach-based Recommendation System for Personalized Web-based Services,” *Applied Artificial Intelligence*, vol. 35, no. 7, pp. 489–504, 2021, doi: 10.1080/08839514.2021.1883855.
- [10] C. Subramanian and B. Ravindran, “Causality in Bandits: A Survey,” *ACM Comput. Surv.*, vol. 57, no. 12, pp. 1–30, 2025, doi: 10.1145/3744917.
- [11] Y. Saito and T. Joachims, “Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances,” in *Fifteenth ACM Conference on Recommender Systems*, Amsterdam Netherlands: ACM, vol. 2021, no. September, pp. 828–830, 2021, doi: 10.1145/3460231.3473320.
- [12] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature*, vol. 549, no. 7671, pp. 195–202, 2017, doi: 10.1038/nature23474.
- [13] V. Havlíček et al., “Supervised learning with quantum-enhanced feature spaces,” *Nature*, vol. 567, no. 7747, pp. 209–212, 2019, doi: 10.1038/s41586-019-0980-2.
- [14] M. Schuld and N. Killoran, “Quantum Machine Learning in Feature Hilbert Spaces,” *Phys. Rev. Lett.*, vol. 122, no. 4, p. 040504, 2019, doi: 10.1103/PhysRevLett.122.040504.
- [15] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, “Parameterized quantum circuits as machine learning models,” *Quantum Sci. Technol.*, vol. 4, no. 4, p. 043001, 2019, doi: 10.1088/2058-9565/ab4eb5.
- [16] S. Qassimi and S. Rakrak, “Multi-objective contextual bandits in recommendation systems for smart tourism,” *Sci Rep*, vol. 15, no. 1, p. 13669, 2025, doi: 10.1038/s41598-025-89920-2.
- [17] S. Brahmachari, J. Lumbreras, and M. Tomamichel, “Quantum contextual bandits and recommender systems for quantum data,” *Quantum Mach. Intell.*, vol. 6, no. 2, p. 58, 2024, doi: 10.1007/s42484-024-00189-6.

- [18] F. Torresan and M. Baltieri, “Disentangled representations for causal cognition,” *Physics of Life Reviews*, vol. 51, no. December, pp. 343–381, 2024, doi: 10.1016/j.plrev.2024.10.003.
- [19] G. Valverde, D. Quesada, P. Larrañaga, and C. Bielza, “Causal reinforcement learning based on Bayesian networks applied to industrial settings,” *Engineering Applications of Artificial Intelligence*, vol. 125, no. October, p. 106657, 2023, doi: 10.1016/j.engappai.2023.106657.
- [20] Z. Bian, C. Shi, Z. Qi, and L. Wang, “Off-Policy Evaluation in Doubly Inhomogeneous Environments,” *Journal of the American Statistical Association*, vol. 120, no. 550, pp. 1102–1114, 2025, doi: 10.1080/01621459.2024.2395593.
- [21] C. H. Leung, Y. Huang, Y. Li, and Q. Wu, “Distributionally Robust Policy Evaluation and Learning for Continuous Treatment with Observational Data,” *AAAI*, vol. 39, no. 17, pp. 18209–18217, 2025, doi: 10.1609/aaai.v39i17.34003.
- [22] H. Kiyohara, Y. Saito, T. Matsuhira, Y. Narita, N. Shimizu, and Y. Yamamoto, “Doubly Robust Off-Policy Evaluation for Ranking Policies under the Cascade Behavior Model,” in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, Virtual Event AZ USA: ACM, vol. 2022, no. February, pp. 487–497, 2022, doi: 10.1145/3488560.3498380.
- [23] H. Morioka and A. Hyvärinen, “Causal Representation Learning Made Identifiable by Grouping of Observational Variables,” vol. 2023, no. October, pp. 1-45, 2023, arXiv, doi: 10.48550/ARXIV.2310.15709.
- [24] C. Uhler, R. Welch, and J. Zhang, “Identifiability Guarantees for Causal Disentanglement from Purely Observational Data,” in *Advances in Neural Information Processing Systems 37*, Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), vol. 2024, no. December, pp. 102796–102821, 2024, doi: 10.52202/079017-3265.
- [25] M. Cerezo et al., “Variational quantum algorithms,” *Nat Rev Phys*, vol. 3, no. 9, pp. 625–644, 2021, doi: 10.1038/s42254-021-00348-9.
- [26] J. Schnabel and M. Roth, “Quantum kernel methods under scrutiny: a benchmarking study,” *Quantum Mach. Intell.*, vol. 7, no. 1, p. 58, 2025, doi: 10.1007/s42484-025-00273-5.
- [27] X. Liu, Y. Qiu, O. Dahlsten, and V. Vedral, “Quantum causal inference with extremely light touch,” *npj Quantum Inf*, vol. 11, no. 1, p. 54, 2025, doi: 10.1038/s41534-024-00956-0.