SEMANTIC REPLAY HAS A COORDINATE SYSTEM: CLAIM-LEVEL EVALUATION FOR DELAYED-FEEDBACK RL

Anonymous authors
Paper under double-blind review

ABSTRACT

A replay buffer can preserve every transition yet change the meaning of its stored state when an encoder update moves the representation coordinates. Delayed semantic-history RL compounds this problem: information-time, censoring, system attribution, and resource conclusions can require different evidence even when they share one return. TRACE (Temporal Replay Admissibility and Claim Evaluation) makes the evidence unit explicit—a policy row or a registered contrast—and maps deterministic provenance failures, scope projections, and residual assumptions to the claims each unit can carry. The operator has absorbing hard failures, commutative claim projections, and explicit empty-set semantics. On WebShop, checkpointed replay yields 0.340 success when old coordinates are reused, 0.398 with version-filtered sampling, and 0.402 with full re-encoding; filtering recovers 93.5% of the point-estimate gap at 0.38 times the re-encoding cost. Under the same 500K-step protocol, the evaluated LLaMA-2-7B and pretrained RoBERTa-base systems reach 0.398 and 0.386 ± 0.018 WebShop success, above the Causal Transformer's 0.358 and GRU's 0.312 point estimates; on ALFWorld they reach 0.375 ± 0.020 and 0.364 ± 0.019, with a teacher–Delay-Survival difference of 0.043 (95% CI $[0.017, 0.069]$). Post-decision positive controls score still higher on both language workloads but are excluded by field provenance, while RecSim's 0.041–0.046 conversion range remains linked to its censoring model. The result is a claim-specific account of semantic replay: a practically large replay-strategy effect, replicated system-level performance ordering across two language-mediated workloads, and conclusions whose statistical and observational conditions remain visible.

1 INTRODUCTION

Consider one stored transition with semantic state $z_t = \phi_v(h_{\leq t})$. After the projection changes from v to $v + 1$, the bytes of z_t are unchanged, but the current critic interprets them in a different coordinate system. The replay buffer has preserved the experience and lost the representation identity needed to consume it. This is an operational distinction: in our WebShop intervention, reusing old coordinates yields 0.340 success, version-filtered sampling yields 0.398, and re-encoding stored prefixes yields 0.402 at 2.65 times the filtering cost.

The same pipeline contains two other time boundaries. An action is selected before terminal text and reward exist, although both may later appear in the training record; a delayed outcome can also arrive after the logging horizon. Web interaction, recommendation, and embodied instruction following expose these boundaries because their policies compress long structured histories into learned states (Yao et al., 2022; Ie et al., 2019; Shridhar et al., 2021). Consequently, one utility row can support policy performance while remaining silent about information legality, encoder-family attribution, censoring, or total cost.

Sequence models address long-history control (Chen et al., 2021; Janner et al., 2021; Zheng et al., 2022; Du et al., 2023); off-policy methods address replay and coverage (Precup et al., 2000; Munos et al., 2016; Kapturowski et al., 2019); censoring estimators state assumptions about unobserved outcomes. What is missing is a common semantics for the *conclusion* produced when these concerns

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

meet. A row is the correct evidence unit for a field-provenance failure, whereas encoder superiority and replay effects are contrasts between rows. Treating both as undifferentiated checklist entries either lets a high score inherit unsupported claims or lets an intentionally invalid intervention arm erase the contrast it was introduced to measure.

TRACE resolves this mismatch by operating on registered evidence units. Each unit declares its estimand, compared rows, candidate claims, hard provenance checks, scope gates, uncertainty record, and residual assumptions. Deterministic post-decision fields and incompatible mixed-coordinate deployment rows trigger absorbing failures; provenance, censoring, and resource gates project the candidate set; an empty projection returns *NoClaim*. Predictive probes remain diagnostics rather than information-legality oracles. Because the gates are set intersections, their order cannot change the surviving claim set.

The experiments instantiate this semantics on Synthetic Delay, RecSim, and WebShop, then reuse the definitions on ALFWorld's official held-out tasks. Every method receives 500K environment steps and five seeds. The replay experiment compares stale, version-filtered, and fully re-encoded strategies under one insertion and update schedule; the semantic-system comparisons include LLaMA-2-7B, pretrained and scratch RoBERTa-base, GRU, and Causal Transformer rows; censoring bridges and component accounting keep observational and resource conditions attached to their own conclusions.

The primary contribution is the evidence-unit formulation and its order-independent claim operator, including decision rules for relational, descriptive, conditional, and empty claim sets. Its empirical consequence is equally concrete: version filtering recovers 93.5% of the stale-to-re-encoded WebShop point-estimate gap at 0.38 times the cost; both evaluated pretrained semantic systems rank above the nonsemantic systems on WebShop and ALFWorld, with registered paired support against GRU on WebShop and the strongest ALFWorld comparators; and deterministic provenance rejects the highest-scoring post-decision controls. TRACE therefore turns a leaderboard row into a reviewable map from evidence to exactly the conclusion it supports.

2 Related Work

Delayed outcomes and identified targets. Delayed-reward methods address credit assignment, outcome lag, and return decomposition (Katsikopoulos & Engelbrecht, 2003; Walsh et al., 2009; Ramstedt et al., 2021; Arjona-Medina et al., 2019; Harutyunyan et al., 2019). Survival and inverse-probability estimators expose positivity, overlap, and censoring assumptions (Cox, 1972; Chapelle, 2014; Robins et al., 1994). TRACE complements these estimators by carrying their identification conditions with the affected contrast. The RecSim study instantiates this connection with primary-hazard imputation, observed-only, stabilized inverse-probability, alternative-hazard, and later-reveal estimates.

Semantic histories and information time. Recurrent agents and transformer policies compress partial histories into decision state (Hausknecht & Stone, 2015; Parisotto et al., 2020; Ni et al., 2022; Chen et al., 2021). Pretrained representations can improve language-conditioned control (Du et al., 2023; Liu et al., 2019; Touvron et al., 2023b). Representation quality and information legality are distinct: a valid decision prefix may predict future return, whereas a field created after the action is unavailable regardless of probe accuracy. TRACE therefore bases legality on deterministic field provenance and uses probe families only to test whether deliberately contaminated controls remain distinguishable.

Replay correction and off-policy evaluation. Off-policy and offline RL study coverage, distribution shift, conservative objectives, and replay reuse (Precup et al., 2000; Munos et al., 2016; Fujimoto et al., 2019; Kumar et al., 2020; Levine et al., 2020). R2D2 directly identifies representational drift and recurrent-state staleness and studies stored-state and burn-in corrections; RESeL stabilizes recurrent context encoders through a separate learning rate (Kapturowski et al., 2019; Luo et al., 2024). Continual and archive-based methods stabilize evolving representations or periodically recompute archive coordinates (Grillotti & Cully, 2021; Lillo & Cheney, 2026). TRACE advances this line at the evaluation layer: version identity is part of the evidence record, filtering and re-encoding are compared as complete replay strategies, and the resulting contrast cannot silently authorize encoder-family, censoring, or cost claims.

From diagnostic results to claim semantics. Ablations quantify component changes, and checklists expose individual risks, but neither specifies whether the evidence unit is a row or a contrast or what an empty surviving set means. TRACE supplies that missing computation: hard provenance failures are absorbing, scope gates are commutative projections, unsupported relational claims return *NoClaim*, and unresolved censoring assumptions remain attached to the affected contrast. ALFWorld tests the same mapping on a second language-mediated control problem.

3 Claim-Level Admissibility

3.1 Delayed-feedback state and evidence assumptions

We separate environment time from reward-observation time. The POMDP $\mathcal{M}_d$ has state space $\mathcal{S}$, action space $\mathcal{A}$, observation space $\mathcal{O}$, dynamics T, observation kernel Ω, reward function R, discount γ, and delay distribution $\mathcal{D}$. At environment step $t \le H$, the agent observes $o_t \sim \Omega(\cdot \mid s_t)$, chooses $a_t \sim \pi(\cdot \mid h_{\le t})$ from $h_{\le t} = (o_1, a_1, \ldots, o_t)$, and transitions to $s_{t+1} \sim T(\cdot \mid s_t, a_t)$. The outcome $r_t = R(s_t, a_t)$ is logged at observation time $\rho_t = t + d_t$, where $d_t \sim \mathcal{D}(\cdot \mid s_t, a_t)$; its arrival does not advance the environment state. For logging horizon $L \ge H$, $c_t = \mathbb{1}[\rho_t \le L]$ denotes whether the outcome is observed. Thus terminal feedback may arrive after the episode without being censored when L covers the delay support, as in the primary ALFWorld slice.

Let $\phi_v : \mathcal{H} \to \mathcal{Z}$ be a deterministic encoder version with $z_t = \phi_v(h_{\le t})$. Interpretable evaluation requires decision-time-visible inputs, a representation sufficient for expressing values as a function of (z_t, a), and replay that is coordinate-compatible with adequate action support (Precup et al., 2000; Munos et al., 2016; Gelada et al., 2019). TRACE turns these premises into explicit row transitions.

3.2 Evidence units and the claim operator

An evidence unit x is either one evaluated row r or a registered contrast $q = (r_a, r_b, m, \theta)$ naming two rows, metric m, and estimand θ. This distinction prevents a relational conclusion from being assigned to either row in isolation. For each x, the registry supplies a candidate set $C_x \subseteq \mathcal{C}$, the data split, denominators, uncertainty record, hard failures $H(x)$, scope gates $G(x)$, and residual assumptions. The evaluated vocabulary is

$$\mathcal{C} = \{c_{\mathrm{d}}, c_{\mathrm{m}}, c_{\mathrm{s}}(q), c_{\mathrm{b}}(q), c_{\mathrm{r}}(q)\} \cup \{c_{\mathrm{e}}(d, q) : d \in \mathcal{E}\}, \quad \mathcal{E} = \{\mathrm{steps}, \mathrm{online}, \mathrm{latency}, \mathrm{total}\}.$$

Here c_{d} is evaluation performance under delayed credit; c_{m} is a recommendation-outcome contrast; $c_{\mathrm{s}}(q)$ is a system-level contrast between a specified pretrained semantic system and comparator; $c_{\mathrm{b}}(q)$ is a backbone-specific contrast; and $c_{\mathrm{r}}(q)$ is the observed difference between two complete replay strategies. Indexing c_{s} and c_{r} by q preserves those system and strategy estimands throughout the operator.

Candidate claims are generated by fixed rules. A directional relational claim requires a registered paired contrast on a shared evaluation set and a two-sided 95% interval whose lower endpoint is positive; a displayed point estimate without that interval remains descriptive. A replay-strategy difference requires a shared environment-step and learner-update schedule, named arms, and arm-specific estimates; its record states whether paired uncertainty or a noninferiority margin is available. Information legality is determined from field provenance, not predictive accuracy. An efficiency advantage requires a common utility target and a directly measured, comparator-matched resource dimension. These rules produce C_x before any gate is applied.

Let $H(x) \subseteq \mathcal{H}_{\mathrm{fail}} = \{h_{\mathrm{vis}}, h_{\mathrm{rep}}\}$. The first failure marks a policy row that consumes a post-decision field; the second marks a standalone row that mixes incompatible coordinates. An intentionally stale row can still be an arm of a registered replay contrast: its hard failure blocks standalone policy or semantic claims but does not erase the measured between-strategy difference. Let $G(x) \subseteq \mathcal{G} = \{g_{\mathrm{cens}}, g_{\mathrm{prov}}, g_{\mathrm{cost}}\}$. Each gate defines an authorized subset $K_g \subseteq \mathcal{C}$, projection $\Gamma_g(C) = C \cap K_g$, and optional assumption record S_g. The provenance gate removes backbone-general and causal semantic interpretations not licensed by q; the cost gate retains only common-target resource contrasts; the censoring gate retains the affected outcome contrast and appends S_{cens}. Thus

$$C'_x = C_x \cap \bigcap_{g \in G(x)} K_g, \qquad S(x) = \bigcup_{g \in G(x)} S_g.$$

Table 1: TRACE decision rules. Predictive probes diagnose separation but do not determine information legality.

Control	Trigger	Transition
Visibility	Field created after action a_t	Reject standalone row via h_{vis}
Replay version	One row mixes coordinate versions	Reject standalone row via h_{rep}
Predictive probe	Held-out AUC under a future target	Attach diagnostic record; no verdict transition
Censoring	Outcome contrast requires a bridge	Condition affected claim on S_{cens}
Provenance	Contrast identifies named systems	Retain the registered system-level claim
Resource	No common utility target or matched meter	Remove efficiency-superiority claim

TRACE returns

$$\mathcal{A}(x, C_x) = \begin{cases} \mathrm{Reject}, & H(x) \neq \varnothing, \\ \mathrm{NoClaim}, & H(x) = \varnothing \ \wedge \ C'_x = \varnothing, \\ \mathrm{Conditional}(S(x), C'_x), & H(x) = \varnothing \ \wedge \ C'_x \neq \varnothing \ \wedge \ S(x) \neq \varnothing, \\ \mathrm{Allow}(C'_x), & H(x) = \varnothing \ \wedge \ C'_x \neq \varnothing \ \wedge \ S(x) = \varnothing. \end{cases} \tag{1}$$

Hard failures are monotone, while claim projections are contractive, idempotent, and commutative. The final claim set is therefore independent of diagnostic display order, and an empty projection cannot be mistaken for approval.

For example, the paired WebShop teacher–GRU contrast generates $c_{\mathrm{s}}(q)$ because its interval excludes zero. The teacher–RoBERTa contrast maps backbone superiority to *NoClaim* because its interval includes zero. The version-filtered–stale contrast generates the descriptive $c_{\mathrm{r}}(q)$ record even though the stale arm cannot authorize a standalone semantic-system conclusion. The reported method-specific landmarks populate resource records; c_{e} remains empty because its rule requires a common utility target. For RecSim, S_{cens} remains attached to c_{m}.

3.3 Prefix serialization, provenance, and replay

Raw trajectories take the form $\{(e_t, m_t)\}_{t=1}^{T}$ of typed event records. These records hold the current observation and action, a timestamp, and metadata available at decision time. The mask $m_t \in \{0,1\}^K$ removes delayed rewards, censoring outcomes, terminal labels, and follow-up text created after action a_t. The semantic teacher uses a frozen LLaMA-2-7B backbone (Touvron et al., 2023a;b); the causal prefix representation is the projected last-token state with $d_z = 256$.

Reward-free pretraining occurs before RL on a fixed domain corpus that excludes evaluation trajectories and removes rewards, terminal outcomes, and post-decision annotations. The projection head uses

$$\mathcal{L}_{\mathrm{pre}} = \mathcal{L}_{\text{next-event}} + \lambda_1 \mathcal{L}_{\text{inv-dyn}} + \lambda_2 \mathcal{L}_{\text{delay-haz}} + \lambda_3 \mathcal{L}_{\mathrm{legal}}, \tag{2}$$

which predicts next event type, inverse dynamics, delay hazard, and action legality. The matched RoBERTa-base control (Liu et al., 2019) uses the same corpus, masking, projection dimension, objectives, optimizer, and RL budget.

Each replay transition carries encoder version v, and version-filtered updates sample

$$\mathcal{B}_v = \{(z_t, a_t, \tilde{r}_t, d_t, c_t, v) \in \mathcal{B} : v = v_{\mathrm{cur}}\}.$$

The backbone is frozen in every row. In the replay intervention, only the 256-dimensional projection is checkpointed; a new v is issued whenever that projection checkpoint changes, and the policy and critic label each insertion with the active v. Stale replay ignores the tag, version-filtered replay samples the current tag, and full re-encoding applies the current projection to stored prefixes before sampling. The insertion stream, learner-update count, and sampled batch count are shared across the three strategies. Filtering changes replay age, current-version occupancy, coverage, and sampled

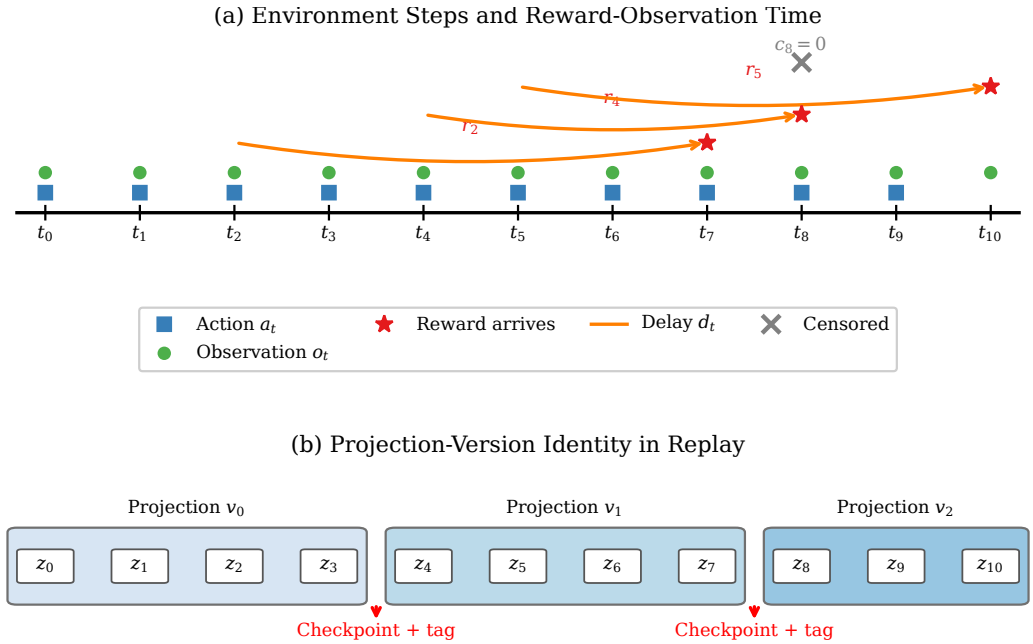

Figure 1: Two clocks and one coordinate identity. Reward observation time is separate from environment time (top); replay items carry the projection version that defines their coordinates (bottom).

transition identities relative to the other arms; $c_r(q)$ therefore measures the complete version-aware replay strategy. The headline semantic-system rows use the final fixed projection, whereas this checkpointed experiment evaluates replay under online projection updates. Figure 1 shows the prefix and version boundaries.

3.4 Censored Targets and Contamination Screen

Reward arrival attaches an outcome to the environment transition (z_t, a_t, z_{t+1}); it does not create an extra state transition. Let $b_{t+1} = 0$ for a terminal or absorbing successor and 1 otherwise. The learning target is

$$y_t = \begin{cases} r_t + \gamma b_{t+1} Q_{\bar{\theta}}(z_{t+1}, \pi_\psi(z_{t+1})) & c_t = 1, \\ g_\omega(z_t, a_t, L - t) + \gamma b_{t+1} Q_{\bar{\theta}}(z_{t+1}, \pi_\psi(z_{t+1})) & c_t = 0, \end{cases} \tag{3}$$

where $g_\omega(z_t, a_t, L - t) \approx \hat{\mathbb{E}}[r_t \mid z_t, a_t, \rho_t > L]$ is the primary bridge fit on the training partition. Its identifying conditions are consistency of the logged outcome, conditional exchangeability of outcome observation given the recorded state, action, remaining window, and delay covariates, and positivity of observation over that support. The discrete-time hazard estimates the probability that ρ_t falls inside the remaining logging window; g_ω regresses observed outcomes on the same covariates and extrapolates to censored records under those conditions. The stabilized-IPW bridge instead weights observed outcomes by the inverse fitted observation probability, the alternative bridge changes the hazard parameterization, and the later-reveal slice evaluates outcomes observed after extending L. None of these constructions indexes an undefined post-horizon state. Section D evaluates this primary target, observed-only estimation, stabilized inverse-probability weighting, an alternative hazard parameterization, and a later-reveal slice.

The post-decision contamination screen trains linear, 2-layer MLP, 4-layer residual, and temporal probes on a fixed held-out partition. Targets are episode outcome, above-median delay, and above-median future cumulative reward. The diagnostic summary is $\max_{p,y} \widehat{\mathrm{AUC}}(p, y)$, reported against the reference value 0.60 and two deliberately contaminated controls. A legal decision prefix can predict future outcome, delay, or return, so this statistic does not activate $H(x)$; deterministic field

provenance supplies the hard visibility decision. Section C reports the complete probe family, bootstrap intervals, class counts, and threshold sensitivity from 0.58 through 0.64.

4 EXPERIMENTS

4.1 RESEARCH QUESTIONS AND PROTOCOL

The experiments answer four questions: (RQ1) how much WebShop utility changes across stale, version-filtered, and re-encoded replay strategies; (RQ2) which registered row and contrast claims satisfy the TRACE decision rules; (RQ3) whether the system-level ordering and claim computation recur on ALFWorld; and (RQ4) what resource profile accompanies each reported operating point.

The first evaluation group contains CartPole-v1 and MountainCar-v0 with $d_t \sim \mathrm{Uniform}(5, 50)$ and 10% training-log censoring, RecSim with 1,000 items and 50-step sessions (Ie et al., 2019), and WebShop with terminal task reward (Yao et al., 2022). Synthetic Delay retains simulator outcomes on an evaluator-only partition, so c_{d} is measured on complete evaluation outcomes even though the learner observes censored training targets. The independent evaluation uses ALFWorld's official held-out task split (Shridhar et al., 2021), delays terminal feedback under the same schedule, and sets the logging horizon to observe the complete delay support in the primary slice. Every method receives 500K environment steps, a replay capacity of 10^6, Adam learning rates $\alpha_Q = \alpha_\pi = 3 \times 10^{-4}$, batch size 256, Polyak coefficient 0.005, and five seeds. Policy updates use SAC-style entropy regularization (Haarnoja et al., 2018).

Compared systems are a GRU critic, Causal Transformer, RUDDER, Delay-Survival, reward-free pretrained RoBERTa-base, and the LLaMA-based semantic teacher evaluated by TRACE. An intentionally leaky teacher includes post-decision fields and functions as a positive leakage control. The GRU and Causal Transformer test recurrent versus causal sequence state under the shared SAC budget; RUDDER adds return decomposition; Delay-Survival adds an explicit delayed-outcome model; and the two pretrained semantic encoders share serialization, masking, projection dimension, reward-free objectives, and RL optimization. Architectures and capacities differ across system families, so TRACE registers system-level contrasts rather than attributing their differences to one architectural factor. All reported "$\pm$" values are standard errors, computed as the standard deviation of five seed-level metrics divided by $\sqrt{5}$. Central method contrasts use paired seed differences and two-sided t_4 95% intervals. WebShop uses 100 evaluation episodes per seed; ALFWorld uses 134 held-out tasks per seed, giving 500 and 670 pooled attempts per method.

Synthetic Delay validation error is the RMSE of target y_t after scaling each task's return target by its training-split standard deviation; the table reports the arithmetic mean of the two task-level RMSEs. Its composite score is

$$\frac{\text{CartPole mean return} + 500 \times \text{MountainCar completion rate}}{2},$$

placing both terms on a 0–500 scale. Composite uncertainty is computed directly from the five seed-level composite scores.

4.2 RQ1–RQ2: REPLAY STRATEGIES AND FIRST-GROUP EVIDENCE

Replay strategy produces the largest observed pipeline change. Checkpointed projection updates yield 0.340 WebShop success with stale embeddings, 0.398 with version filtering, and 0.402 with full re-encoding. Filtering therefore recovers 93.5% of the stale-to-re-encoded point-estimate gap at normalized accelerator cost 1.00 rather than 2.65. TRACE records this result as a descriptive complete-strategy contrast with arm-level uncertainty.

Table 2 places that intervention in the first-group utility context. On WebShop, the version-consistent semantic teacher reaches 0.398 success versus 0.358 for the Causal Transformer and 0.312 for the GRU critic. The Causal-Transformer differences are descriptive; the registered teacher–GRU and pretrained-RoBERTa–GRU contrasts have paired intervals in Table 4. On the nonsemantic controls, the teacher and Delay-Survival differ by 11 composite points with a paired 95% CI of $[-3, 25]$, supporting delayed-credit performance without attributing that near-tie to semantic history. RecSim conversion depends on the outcome bridge and is carried as a conditional result.

Table 2: First-group utility (mean ± SE over five seeds). Synthetic Delay uses the defined 0–500 composite; RecSim rates use 5,000 sessions; WebShop uses 500 evaluations. The post-decision teacher is a positive provenance control.

	Synthetic Delay		RecSim		WebShop	
Method	Val. err.	Composite	CTR@10	Conv.	Success	Reward
GRU Critic	3.42 ± 0.31	412 ± 12	0.142 ± 0.008	0.031 ± 0.004	0.312 ± 0.024	42.1 ± 3.8
Causal Trans.	2.87 ± 0.24	445 ± 10	0.158 ± 0.007	0.037 ± 0.003	0.358 ± 0.021	48.7 ± 3.2
RUDDER	2.95 ± 0.28	438 ± 11	0.151 ± 0.009	0.034 ± 0.004	0.334 ± 0.023	45.3 ± 3.5
Delay-Survival	2.63 ± 0.22	461 ± 8	0.163 ± 0.006	0.039 ± 0.003	0.345 ± 0.020	47.2 ± 3.1
Leaky teacher	1.84 ± 0.15	498 ± 2	0.189 ± 0.005	0.052 ± 0.003	0.421 ± 0.018	58.6 ± 2.7
Semantic teacher	2.21 ± 0.19	472 ± 7	0.178 ± 0.006	0.046 ± 0.003	0.398 ± 0.019	54.8 ± 2.9

Table 3: ALFWorld results on 134 official held-out tasks per seed (five seeds; 670 attempts per method). Values are mean ± SE; paired success differences and 95% CIs use pretrained RoBERTa-base as the table comparator.

Method	Success	Norm. return	Successes	Δ vs. RoBERTa [95% CI]	Fields
GRU Critic	0.276 ± 0.019	0.341 ± 0.022	$185/670$	-0.088 $[-0.119, -0.057]$	Prefix
Causal Trans.	0.319 ± 0.021	0.382 ± 0.021	$214/670$	-0.045 $[-0.075, -0.015]$	Prefix
RUDDER	0.301 ± 0.020	0.366 ± 0.021	$202/670$	-0.063 $[-0.094, -0.032]$	Prefix
Delay-Survival	0.331 ± 0.020	0.397 ± 0.020	$222/670$	-0.033 $[-0.060, -0.006]$	Prefix
RoBERTa-base	0.364 ± 0.019	0.426 ± 0.019	$244/670$	—	Prefix
Semantic teacher	0.375 ± 0.020	0.438 ± 0.020	$251/670$	0.010 $[-0.015, 0.035]$	Prefix
Leaky teacher	0.454 ± 0.018	0.506 ± 0.018	$304/670$	0.090 $[0.066, 0.114]$	Post-dec.

Table 2 also exposes a utility–admissibility reversal: the post-decision control leads WebShop at 0.421 success but is rejected by h_{vis}, whereas the prefix-only semantic teacher at 0.398 remains eligible for its registered system-level contrasts.

Across censoring treatments, semantic-teacher conversion ranges from 0.041 ± 0.003 to 0.046 ± 0.003 and remains above Delay-Survival's 0.039 point estimate. The primary-hazard, stabilized-IPW, alternative-hazard, and later-reveal intervals exclude zero against Delay-Survival; the observed-only interval includes zero. The shared point-estimate direction and varying uncertainty produce $\text{Conditional}(S_{\text{cens}}, \{c_{\text{m}}\})$ rather than an unconditional recommendation claim.

4.3 RQ3: INDEPENDENT-WORKLOAD TRANSPORT

Table 3 applies the same operator, masks, and claim vocabulary to ALFWorld. The semantic teacher records $251/670$ successes and pretrained RoBERTa-base records $244/670$; equal per-seed denominators give a paired mean difference of 0.010 with 95% interval $[-0.015, 0.035]$. The teacher exceeds Delay-Survival by $29/670 = 0.043$ with interval $[0.017, 0.069]$, and pretrained RoBERTa exceeds the Causal Transformer by $30/670 = 0.045$ with interval $[0.015, 0.075]$. These registered contrasts reproduce the WebShop system-level ordering without generating a LLaMA-specific superiority claim; c_{r} is absent because ALFWorld has no replay-version intervention. The post-decision positive control reaches 0.454 ± 0.018 and remains separable under every probe family, while deterministic field provenance supplies its rejection.

Table 4: WebShop system comparison (five seeds, 500 pooled evaluations). Rows share serialization, masking, optimizer, 500K RL steps, and evaluation split. Values are mean ± SE; displayed intervals are paired 95% CIs against GRU.

Encoder	Success	Reward	Successes	Δ success vs. GRU [95% CI]
GRU Critic	0.312 ± 0.024	42.1 ± 3.8	$156/500$	—
RoBERTa-base, scratch	0.362 ± 0.020	49.6 ± 3.0	$181/500$	0.050 $[0.017,\ 0.083]$
RoBERTa-base, pretrained	0.386 ± 0.018	53.2 ± 2.8	$193/500$	0.074 $[0.041,\ 0.107]$
Semantic teacher	0.398 ± 0.019	54.8 ± 2.9	$199/500$	0.086 $[0.052,\ 0.120]$

4.4 Semantic-system contrasts and claim survival

Table 4 separates registered system-level contrasts from backbone and pretraining attributions. Pretrained RoBERTa-base exceeds the GRU critic by 0.074 success with paired 95% CI $[0.041,\ 0.107]$, and the semantic teacher exceeds GRU by 0.086 with interval $[0.052,\ 0.120]$. The semantic-teacher–RoBERTa success difference is 0.012 with paired 95% CI $[-0.010,\ 0.034]$; their reward difference is 1.6 with interval $[-2.1,\ 5.3]$. The scratch RoBERTa row reaches 0.362 ± 0.020, and pretraining raises the displayed point estimate by 0.024. The paired teacher–GRU and RoBERTa–GRU intervals establish their named $c_{\mathrm{s}}(q)$ system contrasts. The Causal Transformer supplies the strongest nonsemantic WebShop point comparator at 0.358, and the scratch row quantifies a 0.024 displayed pretraining difference within RoBERTa. The teacher–RoBERTa interval yields *NoClaim* for LLaMA-specific superiority.

The complete class map is concise: registered semantic-system contrasts survive on both language workloads, the replay-strategy difference is evaluated only on WebShop, and RecSim carries its bridge record. Table 13 gives the exact row and contrast units and returns *NoClaim* for backbone superiority and unmatched efficiency.

4.5 RQ4: utility and resource profiles

The semantic teacher reaches its method-specific 90%-of-final WebShop threshold at 212K steps and 4.43 measured online hours per seed; GRU reaches its lower method-specific threshold at 318K steps and 7.15 hours. TRACE reports these as distinct convergence landmarks and reserves efficiency claims for common-utility comparisons. The teacher package consumes 210.8 A100-hours through its landmark, including 181.30 hours of reward-free pretraining, whereas the GRU package consumes 36.8 A100-hours. The 125M student reaches 0.393 success at 18.0 ms batch-1 median latency and has 56.4 A100-hours of incremental cost once the teacher representation exists, retaining 94.2% of the teacher's point-estimate gain over GRU. Including the one-time teacher pretraining dependency gives 237.7 A100-hours for the first student; for n students sharing that teacher, the attributable cost per student is $56.4 + 181.3/n$. Figure 4 reports step-indexed learning trajectories, and Appendix F keeps method-specific landmark time, batch-1 latency, incremental student cost, and end-to-end attributable cost as separate quantities.

5 Discussion

The central empirical result is a replay-strategy separation: stale-coordinate replay reaches 0.340 WebShop success, version filtering reaches 0.398, and full re-encoding reaches 0.402 at 2.65 times the filtering cost. Version filtering is attractive as a complete systems intervention because it captures most of the point-estimate recovery without re-encoding the full prefix store. TRACE keeps that result separate from the semantic-system contrasts, the censored recommendation contrast, and the resource profiles. Equation 1 makes these distinctions reproducible at the evidence-unit level instead of delegating them to informal interpretation.

All RecSim bridge treatments preserve the point-estimate direction while spanning 0.041–0.046 conversion, so the bridge assumption remains attached to magnitude. Across four probe families, prefix-only, no-mask, and post-decision representations preserve the same dependence ordering; the deterministic field ledger, rather than an AUC threshold, determines legality. The simulator, recommendation, web-interaction, and embodied-text results therefore distinguish four outputs: c_{d}

on complete Synthetic Delay evaluation outcomes, conditional c_{m} on RecSim, comparator-specific $c_{\mathrm{s}}(q)$ on both language workloads, and descriptive $c_{\mathrm{r}}(q)$ for the WebShop replay strategy.

The evaluated regime also defines the interpretation boundary. The semantic comparisons are system-level because architecture, capacity, pretraining, and auxiliary objectives are not jointly factor-matched; the WebShop replay arms differ in age, coverage, and sampled identities as part of filtering; and the reported arm-level SEs do not establish equivalence between filtering and re-encoding. RecSim additionally relies on conditional exchangeability and positivity for its outcome bridge. These conditions focus the contribution on claim semantics and the measured strategy separation while preserving the practical conclusion: representation version must be designed and reported as part of semantic replay.

6 Conclusion

Semantic replay results need a coordinate identity and a claim identity. TRACE supplies both: registered rows and contrasts enter deterministic provenance checks, commutative scope projections, and explicit conditional or empty outcomes. On WebShop, version filtering moves success from 0.340 to 0.398 and captures 93.5% of the re-encoding point-estimate gap at 0.38 times the cost. On ALFWorld, the strongest registered system contrasts reproduce the language-workload ordering, including 0.043 teacher improvement over Delay-Survival and 0.045 pretrained-RoBERTa improvement over Causal Transformer. The resulting evidence map preserves the replay-strategy result, comparator-specific semantic-system results, and the conditional RecSim result without allowing one scalar score to overstate any of them.

Reproducibility Statement

Sections 4.1 and A specify workloads, splits, masks, baselines, optimization, seeds, uncertainty, hardware, and evaluation denominators. Appendices B–G report task-native outcomes, contamination probes, bridge alternatives, replay updates, component cost accounting, and the complete evidence-unit registry.

AI Use Statement

Generative AI tools assisted manuscript editing and presentation throughout the paper, including the title, abstract, section organization, and prose. This statement covers the documented use of generative AI in manuscript preparation.

References

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13566–13577, 2019. URL https://mlanthology.org/neurips/2019/arjonamedina2019neurips-rudder/.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1097–1105, 2014. doi: 10.1145/2623330.2623634. URL https://doi.org/10.1145/2623330.2623634.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision Transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x. URL https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language models. In *International Conference on Machine Learning*, volume 202, pp. 8657–8677, 2023. URL https://proceedings.mlr.press/v202/du23f.html.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, volume 97, pp. 2052–2062, 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, volume 97, pp. 2170–2179, 2019. URL https://proceedings.mlr.press/v97/gelada19a.html.

Luca Grillotti and Antoine Cully. Unsupervised behaviour discovery with quality-diversity optimisation. *arXiv preprint arXiv:2106.05648*, 2021. URL https://arxiv.org/abs/2106.05648.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, volume 80, pp. 1861–1870, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Rémi Munos. Hindsight credit assignment. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1912.02503.

Matthew Hausknecht and Peter Stone. Deep recurrent Q-learning for partially observable MDPs. In *AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*, 2015. URL https://arxiv.org/abs/1507.06527.

Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. RecSim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019. URL https://arxiv.org/abs/1909.04847.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1273–1286, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/099fe6b0b444c23836c4a5d07346082b-Abstract.html.

Steven Kapturowski, Georg Ostrovski, John Quan, Rémi Munos, and Will Dabney. Recurrent experience replay in distributed reinforcement learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1lyTjAqYX.

Konstantinos V Katsikopoulos and Sascha E Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE Transactions on Automatic Control*, 48(4):568–574, 2003. doi: 10.1109/TAC.2003.809799. URL https://doi.org/10.1109/TAC.2003.809799.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020. URL https://arxiv.org/abs/2006.04779.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. URL https://arxiv.org/abs/2005.01643.

Lute Lillo and Nick Cheney. Beyond single-model optimization: Preserving plasticity in continual reinforcement learning. *arXiv preprint arXiv:2604.15414*, 2026. URL https://arxiv.org/abs/2604.15414.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Fan-Ming Luo, Zuolin Tu, Zefang Huang, and Yang Yu. Efficient recurrent off-policy RL requires a context-encoder-specific learning rate. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1536. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5706668422bd0d82588998ebe1067133-Abstract-Conference.html.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 1054–1062, 2016. URL https://mlanthology.org/neurips/2016/munos2016neurips-safe/.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. In *International Conference on Machine Learning*, volume 162, pp. 16691–16723, 2022. URL https://proceedings.mlr.press/v162/ni22a.html.

Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphaël Lopez Kaufman, Aidan Clark, Seb Noury, Matthew Botvinick, Nicolas Heess, and Raia Hadsell. Stabilizing transformers for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7487–7498, 2020. URL https://proceedings.mlr.press/v119/parisotto20a.html.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*, pp. 759–766, 2000. URL https://mlanthology.org/icml/2000/precup2000icml-eligibility/.

Simon Ramstedt, Yann Bouteiller, Giovanni Beltrame, Christopher Pal, and Jonathan Binas. Reinforcement learning with random delays. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QFYnKlBJYR.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89 (427):846–866, 1994. doi: 10.1080/01621459.1994.10476818. URL https://doi.org/10.1080/01621459.1994.10476818.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03768.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey

Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b. URL https://arxiv.org/abs/2307.09288.

Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems*, 18(1): 83–105, 2009. doi: 10.1007/s10458-008-9056-7. URL https://doi.org/10.1007/s10458-008-9056-7.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. In *International Conference on Machine Learning*, volume 162, pp. 27042–27059, 2022. URL https://proceedings.mlr.press/v162/zheng22c.html.

A Implementation and Evaluation Details

All online runs use an 8×A100 80GB node with CUDA 12.4. Each seed is pinned to one A100, and the five seeds for a method run concurrently; reported hours per seed are elapsed hours for one pinned device. Semantic-teacher reward-free pretraining uses four A100s, while online teacher inference uses one A100 per seed. Batch-1 latency uses 256-token prefixes after 500 warm-up calls and reports the median and p95 over $10,000$ synchronized forward passes. Online hours are end-to-end measurements of the implemented rollout and batched learner execution, whereas batch-1 latency is a separate serving microbenchmark and is not multiplied by the environment-step count. For example, the measured teacher landmark of 212K decisions in 4.43 hours corresponds to 13.3 completed environment decisions per second; the 198 ms value characterizes isolated batch-1 inference, not that end-to-end execution rate.

WebShop and ALFWorld serialize HTML-derived or textual observations, legal-action metadata, and executed actions before each decision. The same field-level mask is applied during pretraining, replay insertion, and evaluation. ALFWorld uses the official held-out task split, 134 tasks per seed, and terminal outcomes delayed by $\mathrm{Uniform}(5, 50)$ environment steps. The primary ALFWorld logging horizon covers the complete imposed delay support, so delayed terminal rewards are observed without right censoring.

RecSim uses 5,000 sessions, 1,000 items, 50 decisions per session, immediate clicks, and conversion delays drawn from $\mathrm{LogNormal}(\mu = 2.5, \sigma = 1.2)$ with 15% censoring. WebShop uses 100 held-out episodes per seed. For binary outcomes, pooled counts are reported alongside seed-level means; standard errors and paired confidence intervals continue to use seeds as the independent unit.

Table 5 states the role of each learning system. Every row receives the same legal-action interface, environment-step budget, optimizer family, replay capacity, and evaluation split. The architectures, capacities, and method-specific objectives differ, which is why the main text reports named system contrasts rather than single-factor effects.

Table 5: Comparator specification under the shared protocol.

Method	Decision state	Additional training signal	Comparison role
GRU critic	Recurrent prefix	SAC actor–critic	Recurrent nonsemantic state
Causal Transformer	Causal prefix sequence	SAC actor–critic	Transformer history model
RUDDER	Recurrent prefix	Return decomposition	Delayed credit assignment
Delay-Survival	Prefix and delay covariates	Survival/hazard target	Censored delayed outcomes
RoBERTa-base	Masked 256-d semantic prefix	Equation 2	Semantic encoder family
LLaMA-2-7B teacher	Masked 256-d semantic prefix	Equation 2	Semantic encoder family
Post-decision control	Prefix plus future fields	Teacher objective	Information-time positive control

B Synthetic Delay Task-Native Results

Table 6 exposes both task scales used by the composite in Table 2. Every displayed point estimate reproduces the main-table composite exactly. The within-task pattern agrees across CartPole and MountainCar: the semantic teacher is competitive with Delay-Survival, and their paired interval overlaps zero.

The semantic teacher minus Delay-Survival is 10 CartPole points with paired 95% CI $[-2,\ 22]$, 0.024 MountainCar completion probability with interval $[-0.006,\ 0.054]$, and 11 composite points with interval $[-3,\ 25]$.

Table 6: Task-native Synthetic Delay results (mean ± SE over five seeds). MountainCar counts pool 100 episodes per seed. Validation error is normalized target RMSE; the composite follows Section 4.1.

Method	CartPole return	MountainCar success	MC count	CP val. err.	MC val. err.	Comp. score
GRU	430 ± 12	0.788 ± 0.025	$394/500$	2.86 ± 0.25	3.98 ± 0.37	412 ± 12
Causal Trans.	460 ± 9	0.860 ± 0.020	$430/500$	2.38 ± 0.20	3.36 ± 0.28	445 ± 10
RUDDER	452 ± 11	0.848 ± 0.022	$424/500$	2.44 ± 0.23	3.46 ± 0.33	438 ± 11
Delay-Survival	476 ± 7	0.892 ± 0.018	$446/500$	2.15 ± 0.18	3.11 ± 0.26	461 ± 8
Leaky teacher	499 ± 1	0.994 ± 0.004	$497/500$	1.51 ± 0.12	2.17 ± 0.18	498 ± 2
Semantic teacher	486 ± 6	0.916 ± 0.018	$458/500$	1.80 ± 0.16	2.62 ± 0.22	472 ± 7

C Post-Decision Contamination Stress Test

Table 7 uses identical held-out embeddings and targets across four probe families. Each AUC is followed by a seed-stratified bootstrap 95% interval. The held-out targets contain $2480/2520$ positive/negative episode outcomes, $2500/2500$ above/below-median delays, and $2491/2509$ above/below-median future returns.

Table 7: Held-out future-target probes (mean AUC ± SE over five seeds, followed by bootstrap 95% CI). "No mask" omits field filtering; "Post-decision" explicitly adds future fields. Probes share one split and serve as diagnostics; field provenance determines legality.

Probe	Encoder	Outcome AUC [CI]	Delay AUC [CI]	Future reward AUC [CI]
Linear	Prefix-only	0.52 ± 0.02 $[0.48, 0.56]$	0.51 ± 0.01 $[0.49, 0.53]$	0.55 ± 0.02 $[0.51, 0.59]$
Linear	No mask	0.60 ± 0.03 $[0.54, 0.66]$	0.57 ± 0.02 $[0.53, 0.61]$	0.65 ± 0.03 $[0.59, 0.71]$
Linear	Post-decision	0.89 ± 0.01 $[0.87, 0.91]$	0.84 ± 0.02 $[0.80, 0.88]$	0.92 ± 0.01 $[0.90, 0.94]$
2-layer MLP	Prefix-only	0.53 ± 0.02 $[0.49, 0.57]$	0.51 ± 0.01 $[0.49, 0.53]$	0.57 ± 0.02 $[0.53, 0.61]$
2-layer MLP	No mask	0.62 ± 0.03 $[0.56, 0.68]$	0.59 ± 0.02 $[0.55, 0.63]$	0.68 ± 0.03 $[0.62, 0.74]$
2-layer MLP	Post-decision	0.91 ± 0.01 $[0.89, 0.93]$	0.87 ± 0.02 $[0.83, 0.91]$	0.94 ± 0.01 $[0.92, 0.96]$
4-layer res.	Prefix-only	0.54 ± 0.02 $[0.50, 0.58]$	0.53 ± 0.02 $[0.49, 0.57]$	0.58 ± 0.02 $[0.54, 0.62]$
4-layer res.	No mask	0.65 ± 0.02 $[0.61, 0.69]$	0.62 ± 0.02 $[0.58, 0.66]$	0.71 ± 0.02 $[0.67, 0.75]$
4-layer res.	Post-decision	0.93 ± 0.01 $[0.91, 0.95]$	0.89 ± 0.02 $[0.85, 0.93]$	0.95 ± 0.01 $[0.93, 0.97]$
Temporal	Prefix-only	0.56 ± 0.02 $[0.52, 0.60]$	0.54 ± 0.02 $[0.50, 0.58]$	0.59 ± 0.02 $[0.55, 0.63]$
Temporal	No mask	0.67 ± 0.02 $[0.63, 0.71]$	0.64 ± 0.02 $[0.60, 0.68]$	0.73 ± 0.02 $[0.69, 0.77]$
Temporal	Post-decision	0.94 ± 0.01 $[0.92, 0.96]$	0.90 ± 0.02 $[0.86, 0.94]$	0.96 ± 0.01 $[0.94, 0.98]$

Prefix-only $<$ no-mask $<$ explicitly leaky holds for every target and probe family. Reference thresholds from 0.60 through 0.64 separate the prefix-only representation from both contaminated variants on at least one target. At 0.58 the temporal future-reward probe also flags the prefix-only representation, illustrating why predictive AUC cannot define information legality.

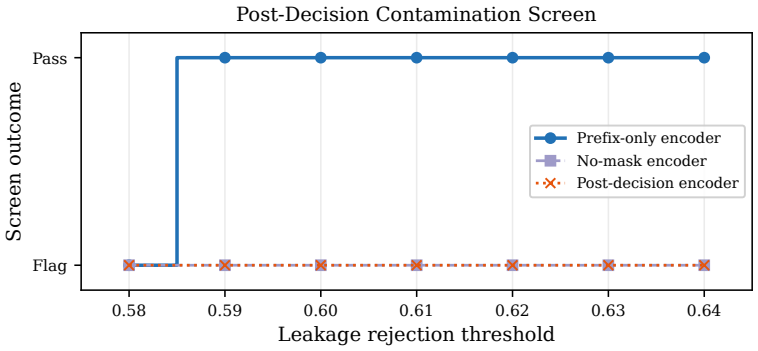

Figure 2: Diagnostic sensitivity from reference threshold 0.58 through 0.64. The change for a legal prefix demonstrates threshold dependence; both future-field controls remain separated. TRACE verdicts use deterministic field provenance instead.

D RecSim Censoring-Bridge Validation

The bridge analysis uses $750/5000$ censored and $4250/5000$ observed sessions. Table 8 compares the primary imputation target with an observed-only estimate, stabilized inverse-probability weighting, an alternative discrete hazard, and a later-reveal slice. Primary imputation and IPW invoke the conditional-exchangeability and positivity conditions stated with Equation 3; observed-only makes no extrapolation to the 750 censored sessions, and later reveal replaces model-based recovery with outcomes observed under the extended window. All conversion point estimates remain above Delay-Survival, while their spread and the observed-only interval motivate the conditional label.

Table 8: RecSim bridge sensitivity over 5,000 sessions (mean $\pm$ SE over five seeds). Conversion differences and 95% CIs are paired against Delay-Survival at 0.039; ECE and Brier score assess bridge calibration.

Bridge	CTR@10	Conversion	Δ conversion [95% CI]	ECE	Brier
Primary hazard	0.178 ± 0.006	0.046 ± 0.003	$0.007\ [0.002, 0.012]$	0.028	0.143
Observed-only	0.174 ± 0.006	0.041 ± 0.003	$0.002\ [-0.002, 0.006]$	0.034	0.149
Stabilized IPW	0.176 ± 0.006	0.044 ± 0.003	$0.005\ [0.001, 0.009]$	0.025	0.145
Alternative hazard	0.177 ± 0.005	0.045 ± 0.003	$0.006\ [0.002, 0.010]$	0.023	0.141
Later reveal	0.179 ± 0.005	0.045 ± 0.002	$0.006\ [0.003, 0.009]$	0.019	0.139

The seed-level conversion shifts from the primary bridge to observed-only are

$$-0.004,\ -0.006,\ -0.005,\ -0.004,\ -0.006.$$

Corresponding shifts for IPW, alternative hazard, and later reveal remain within $[-0.003,\ 0.001]$.

E Replay, Distillation, and Learning-Curve Diagnostics

Table 9: WebShop replay strategies under checkpointed projection updates (mean $\pm$ SE over five seeds). The LLaMA backbone remains frozen; relative accelerator use is normalized to version-filtered replay. The complete strategy includes its resulting current-version support and sampled identities.

Replay strategy	Success	Repr. drift (cosine)	Relative A100-h
Stale, no versioning	0.340 ± 0.021	0.31 ± 0.02	0.94
Version-filtered	0.398 ± 0.019	0.08 ± 0.01	1.00
Full re-encoding	0.402 ± 0.018	—	2.65

Table 10: WebShop distillation (mean ± SE over five seeds). Gain retention is relative to the teacher improvement over GRU; latency is the batch-1 median.

Encoder	Success	Latency (ms/step)	Gain retention
GRU baseline	0.312 ± 0.024	2.0	—
LLaMA-2-7B teacher	0.398 ± 0.019	198.0	100%
125M student	0.393 ± 0.020	18.0	94.2%

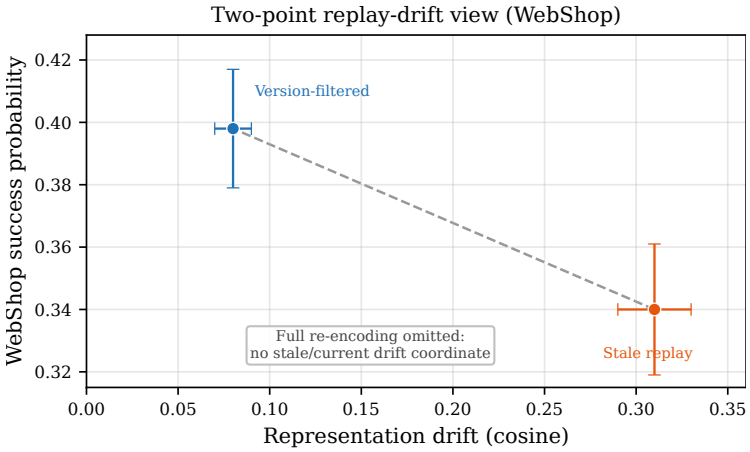

Figure 3: Two replay operating points: measured drift and WebShop success for stale and version-filtered strategies. The line connects the registered operating points; full re-encoding recomputes current coordinates and therefore has no stale/current drift coordinate.

The trajectories show where each system's final utility is acquired over environment interaction. They complement, rather than replace, the distinct serving-latency and end-to-end resource measurements below.

F Resource Accounting by Operating Point

Table 11 separates final 500K-step utility, time to each method's 90%-of-final landmark, batch-1 latency, and accelerator cost accumulated through that landmark. Because the final utilities differ, these rows populate resource profiles; TRACE reserves efficiency superiority for a common-target comparison. The teacher's 4.43 landmark hours per seed corresponds to $5 \times 4.43 = 22.15$ A100-hours of online RL through the landmark. The student total is incremental conditional on the teacher representation; adding the one-time 181.30 A100-hour teacher pretraining stage yields the end-to-end and amortized totals reported in Section 4.5.

Table 11: WebShop resource profiles over five seeds. The landmark is the first interpolated step reaching 90% of each method's own final 500K-step success. Cost is accumulated through that landmark; the student's value is incremental after teacher pretraining.

Method	Landmark steps (K)	Hours/seed	Latency median/p95 (ms)	Landmark A100-h	Final success
GRU Critic	318	7.15	2.0/2.6	36.8	0.312
Causal Transformer	286	6.05	8.3/10.9	43.9	0.358
Delay-Survival	260	5.42	12.1/15.8	50.2	0.345
Semantic teacher	212	4.43	198.0/231.4	210.8	0.398
Distilled student	219	4.88	18.0/22.7	56.4	0.393

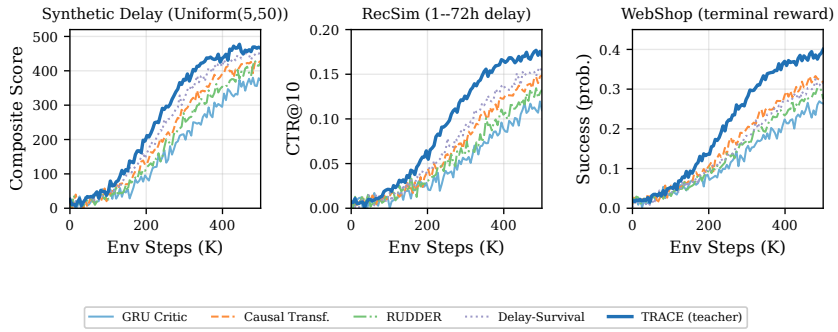

Figure 4: Step-indexed learning trajectories under the 500K interaction protocol. Table 11 separately reports elapsed online time and method-specific utility landmarks.

The component ledger below explains why the teacher combines fewer measured online hours with a larger end-to-end accelerator total: reward-free pretraining dominates its package cost. The student reverses that serving-latency profile while inheriting a share of the teacher dependency.

Table 12: A100-hour components through the reported landmark. Online RL equals five one-device seed runs; rows sum disjoint stages. The student row is incremental and excludes the shared 181.30-hour teacher pretraining dependency.

Method	Online RL	Pretrain	Distill./fit	Probe	Eval.	Total
GRU	35.75	0	0	0.45	0.60	36.80
Causal Trans.	30.25	0	10.80	1.05	1.80	43.90
Delay-Survival	27.10	0	19.80	1.10	2.20	50.20
Semantic teacher	22.15	181.30	0	3.20	4.15	210.80
Distilled student	24.40	0	28.20	1.10	2.70	56.40

G Evidence Paths

The manuscript-level trace closes through six paths. *Visibility* starts from field creation time: the positive controls consume future fields and fail h_{vis}, while Table 7 shows that the probes remain sensitive without controlling that verdict. *Replay* registers version-filtered–stale and re-encoded–filtered strategy contrasts under checkpointed projections; Table 9 supplies utility, drift, and cost while retaining the arm-level uncertainty status. *Censoring* attaches the stated exchangeability and positivity assumptions to RecSim and tests them through the five bridge analyses in Table 8.

System comparison uses named paired contrasts rather than encoder-family attribution. Table 4 supplies WebShop comparisons against GRU, and Table 3 supplies the independent ALFWorld comparisons; the teacher–RoBERTa intervals return *NoClaim* for backbone superiority. *Delayed-credit performance* uses evaluator-observed Synthetic Delay outcomes, with task-native scales and paired intervals in Table 6. *Resources* separates serving latency, method-specific convergence landmarks, incremental cost, and end-to-end attributable cost in Tables 11–12; these values remain resource profiles under the common-target decision rule.

The final registry makes the evidence-unit distinction executable for every headline conclusion. Rows carry field and coordinate provenance, while ordered pairs carry the relational estimand and its uncertainty record.

Table 13: TRACE registry for the principal rows and contrasts. Candidate generation follows Section 3; q always names the displayed ordered pair.

Evidence unit	C_x	$H(x)$	Gate or record	Output
Post-decision controls	row claims	h_{vis}	Field provenance	Reject
Stale WebShop deployment row	row claims	h_{rep}	Version provenance	Reject
Filtered–stale WebShop q	$c_{\mathrm{r}}(q)$	—	Descriptive arm SEs	Allow $c_{\mathrm{r}}(q)$
RecSim teacher–Delay-Survival q	c_{m}	—	$g_{\mathrm{cens}}, S_{\mathrm{cens}}$	Conditional c_{m}
Synthetic teacher row	c_{d}	—	Complete evaluator outcomes	Allow c_{d}
WebShop RoBERTa–GRU q	$c_{\mathrm{s}}(q)$	—	Paired CI $[.041, .107]$	Allow $c_{\mathrm{s}}(q)$
WebShop teacher–GRU q	$c_{\mathrm{s}}(q)$	—	Paired CI $[.052, .120]$	Allow $c_{\mathrm{s}}(q)$
WebShop teacher–RoBERTa q	$\varnothing$	—	CI contains zero	NoClaim
WebShop semantic–Causal q	$\varnothing$	—	No paired CI	NoClaim
ALFWorld RoBERTa–Causal q	$c_{\mathrm{s}}(q)$	—	Paired CI $[.015, .075]$	Allow $c_{\mathrm{s}}(q)$
ALFWorld teacher–Delay-Survival q	$c_{\mathrm{s}}(q)$	—	Paired CI $[.017, .069]$	Allow $c_{\mathrm{s}}(q)$
Method-specific landmark contrasts	$\varnothing$	—	g_{cost}	NoClaim

The registry exposes the paper's main decision consequence: version-aware replay supports a measured strategy difference, the strongest semantic results remain comparator-specific, and unsupported superiority or efficiency interpretations terminate as *NoClaim*.