DYNACAST: INTERVENTION-GROUNDED STATE FOR LONG-HORIZON LANGUAGE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon language agents need compact persistent state that preserves distinctions in past experience relevant to future decisions. Existing memory objectives emphasize retention or prediction without specifying which state differences should matter when goals, world conditions, or plans change. Learning such a criterion also requires comparing open-ended language actions across histories and keeping component-specific targets consistent with a global recursive metric. We introduce DYNACAST, which grounds persistent goal, world, and plan registers in controlled futures generated by executable goal, physical-world, and route interventions in resettable environments. We make these futures comparable through canonical action classes and frozen argument-binding distributions, and derive global and component Bellman targets from shared transport couplings and a common maximizing control, ensuring exact target compatibility while establishing contraction and covered-value preservation for the exact operator. Experiments across eight benchmarks with nine matched baseline/control comparisons across four model families—Qwen2.5-7B, Llama-3.1-8B, Gemma-2-9B-IT, and Mistral-7B-Instruct—isolate 1.6–3.1-point macro-averaged task-score gains from matched finite-horizon Bellman bootstrapping over direct Monte Carlo metrics in both state architectures, with intervention and register ablations identifying the contributions of executable controls and modular state, and exact checks on 24 enumerated delivery SCMs validating operator implementation.

1 INTRODUCTION

A transcript records what an agent did; a control state must retain what still matters for what it will do. Long-horizon histories interleave user constraints, grounded world facts, rejected plans, tool failures, and intermediate commitments, so any compression asserts which distinctions are safe to discard. Outcome-level and belief-centric objectives sharpen credit assignment (Shao et al., 2024; Feng et al., 2025; Tang et al., 2026), and recurrent and retrieval memories widen access to earlier information (Bulatov et al., 2022; Packer et al., 2023); these mechanisms govern what remains available, whereas state learning must decide which retained differences alter future control.

Two obstacles separate that requirement from a trainable objective. First, retention, reconstruction, and observational prediction never define interchangeability under controlled changes to the task, the physical world, or the chosen route; architectural labels such as "goal" and "world" carry no behavioral semantics on their own (Scholkopf et al., 2021; Locatello et al., 2019). Second, the natural formal criterion—bisimulation—assumes the compared states share an action space, whereas language agents emit open-ended tool arguments, and independently optimized component distances need not agree with any single global recursive target.

Each obstacle fixes a design requirement. State equivalence must be defined by executable controlled futures: goal changes, physical-world mutations, and alternative routes with the same terminal effect are executed in resettable environments, while internal register swaps remain representation tests rather than samples from a physical kernel. Recursive comparison must in turn operate on a closed canonical control space and attribute one global Bellman witness across state components. Satisfying both joins causal interchange interventions (Geiger et al., 2022; 2025) to bisimulation metrics (Ferns et al., 2011) in recurrent language-agent control.

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.

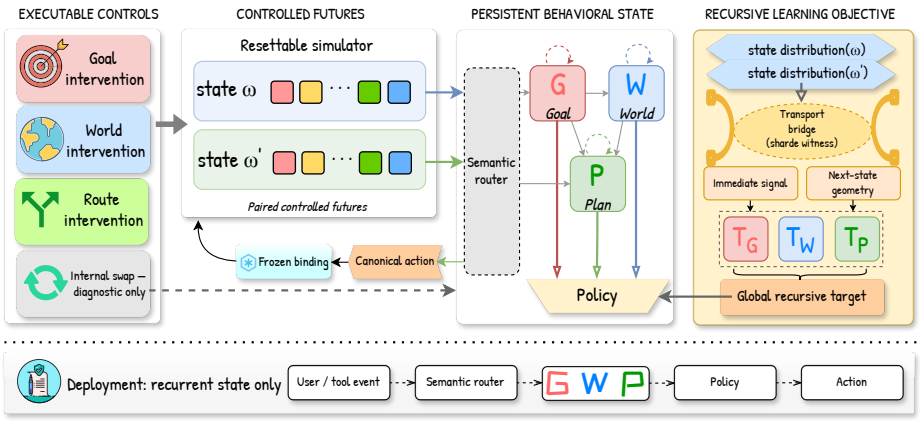

Figure 1: **DynaCAST overview.** Executable goal, world, and route controls produce paired controlled futures from resettable simulator states. A semantic router updates persistent goal (G), world (W), and plan (P) registers; a shared transport witness links immediate-signal and next-state geometry to compatible component targets. Internal register swaps are used only as a diagnostic. Deployment retains the recurrent state and policy, without simulator resets or transport.

We make three contributions. **(1)** We formulate behavioral state as an intervention-indexed equivalence over histories and realize it with executable goal, physical-world, and route kernels (Section 2); persistent $G/W/P$ registers parameterize this equivalence, and their names do not supply its semantics. **(2)** We close open-ended language actions with canonical controls and frozen within-class bindings, then learn an action-conditioned recursive pseudometric in which a single optimal-transport witness defines the global target and its component attribution at once, so register targets cannot drift into mutual inconsistency; on the covered process the operator is contractive and preserves covered value (Section 3). **(3)** We evaluate DynaCAST against nine matched baseline/control comparisons using Qwen2.5-7B, Llama-3.1-8B, Gemma-2-9B-IT, and Mistral-7B-Instruct across eight embodied, web, scientific, application, retail, and tool-use benchmarks, and check the implemented operator against exact kernels on 24 enumerated delivery SCMs; a matched objective factorial, identity-only and no-transport controls, complete register-subset ablations, and mechanism and deployment studies attribute the improvement to each claimed source separately (Section 4).

The ingredients we combine exist individually and answer different questions: structured memory decides where information is stored, DeepMDP and world models learn predictive latent dynamics, bisimulation and DBC decide which factual transitions should be close, and causal-state methods predict future tests (Zhang et al., 2020; Gelada et al., 2019; Singh et al., 2004; Zhang et al., 2019). DynaCAST instead makes executable intervention programs the index set of a recursive history metric, closes open tool arguments with a frozen binding law, and derives every register target from the same future coupling; no ingredient in isolation specifies this learning object. The experiments separate its parts: 1.8- and 1.6-point matched finite-horizon bootstrap effects in the two state architectures, a 2.1-point effect of adding executable controls to an otherwise identical recursive learner, and a 1.7-point gain over an equal-budget predictor, with selective component responses and calibrated behavior under the deployed policy.

2 Problem and Method

Figure 1 separates three easily conflated things: the training-time intervention process, the persistent policy-visible state, and the deployment path. The object being learned is the metric over histories under executable intervention–action controls; the register layout is that metric's recurrent parameterization, not its semantic definition.

2.1 Executable interventions and recurrent registers

We model a benchmark by latent physical/database state X_t, task Y_t, executable route directive R_t, action a_t, observation o_t, typed effect E_t, and exogenous noise. Because a reset intervention must

preserve unmanipulated latent coordinates, the exact training-time state is

$$\omega_t = (h_t, X_t, Y_t, R_t) \in \Omega_k, \qquad h_t = (x^{\mathrm{usr}}_{\le t}, a_{<t}, o_{\le t}). \tag{1}$$

The deployed agent observes only h_t; simulator snapshots expose ω_t solely for reset, pairing, and target construction, and no hidden-state posterior is estimated. Three intervention families are admitted: $j_G = \mathrm{do}(Y = y')$, $j_X = \mathrm{do}(X^S = x'^S)$, and $j_R = \mathrm{do}(R = r')$ over two prevalidated routes with the same terminal effect. An internal W or P swap is a representation test, never a draw from the physical kernel.

Open-ended tool arguments would otherwise prevent two histories from sharing an action set, leaving the recursive comparison undefined. We therefore form a control $u = (j, c)$ from an intervention j and a canonical action class c, and let a development-frozen binding distribution $\kappa_c(b \mid h)$ map c to concrete arguments b. Each u then induces exactly one signal distribution $\mu^\rho_u(\cdot \mid \omega)$ and one next-state kernel $K_u(\cdot \mid \omega)$; cohorts $\mathcal{C}_k$ retain only pairs whose covered control set $\mathcal{U}_k$ is common and nonempty.

The observable recurrent state holds four hidden tokens per register. A semantic router assigns mixed user/environment events to goal-bearing r_G, world-bearing r_W, and structural r_S channels, and masked updates enforce

$$G_{t+1} = F_G(G_t, a_t, r_G, r_S), \qquad W_{t+1} = F_W(W_t, r_W, r_S, a_t), \tag{2}$$

$$P_{t+1} = F_P(P_t, G_{t+1}, W_{t+1}, r_S, a_t). \tag{3}$$

The plan register therefore reads the already-updated G_{t+1} and W_{t+1}, whereas G and W are updated only from their own routed channels and the action. Adapters validate entity support, SCM agreement, executability, and donor independence before a cross-trajectory splice is accepted, and the unrestricted-parent control receives identical routed inputs, isolating modular dynamics from router supervision. Full construction rules appear in Appendix C, with the frozen training and evaluation contract in Appendix E.25; Appendix Table 3 collects the notation and Algorithm 1 gives the complete training loop.

2.2 Action-conditioned recursive metric

Let $\rho = (r, E) \in [0, 1]^m$ be the normalized immediate signal and $c_\rho(x, x') = \sum_m w_m |x_m - x'_m|$, where the fixed signal weights sum to one. For a cohort with common controls,

$$(\mathcal{B}_k d)(\omega, \omega') = \max_{u \in \mathcal{U}_k} \left[(1-\gamma)\mathcal{W}_{c_\rho}(\mu^\rho_u, \mu^{\rho'}_u) + \gamma \mathcal{W}_d(K_u, K'_u) \right]. \tag{4}$$

All measures are conditioned on their displayed state and control. The practical target pushes M paired simulator continuations through an EMA encoder; Sinkhorn iterations compute shared entropic signal and next-state couplings, and the target evaluates the unregularized ground costs under those couplings rather than the regularized or debiased Sinkhorn objective. Each component distance reads only its own register, $D_\phi = \sum_g \beta_g d^g_\phi$ for $g \in \{G, W, P\}$.

Transporting each component independently would in general select different couplings and conflict with the global maximum. We instead fix component costs c_g partitioning the global cost, $c_\rho = \sum_g \beta_g c_g$, with their coordinate-level construction in Appendix C. For every candidate control, a single global entropic coupling $\Lambda^\star_{u,\varepsilon}$ of the two empirical signal measures gives

$$\widehat{\Delta}^{\varepsilon,M}_g = \mathbb{E}_{\Lambda^\star_{u,\varepsilon}} c_g, \qquad \widehat{\Delta}^{\varepsilon,M}_\rho = \sum_g \beta_g \widehat{\Delta}^{\varepsilon,M}_g. \tag{5}$$

Here $\widehat{\Delta}^{\varepsilon,M}_\rho$ is the unregularized ground cost under the computed coupling, not the exact Wasserstein term Δ_ρ in Equation 4; their discrepancy is part of estimation error. The globally maximizing $u^\star$ and the same entropic next-state coupling $\Gamma^\star_\varepsilon$ used by the practical global target define

$$T_g = (1-\gamma)\widehat{\Delta}^{\varepsilon,M}_g + \gamma \mathbb{E}_{(\tilde\omega, \tilde\omega') \sim \Gamma^\star_\varepsilon} d^g_{\bar\phi}(q^g_{\bar\phi}(h(\tilde\omega)), q^g_{\bar\phi}(h(\tilde\omega'))). \tag{6}$$

Consequently $\sum_g \beta_g T_g = \widehat{\mathcal{B}}^{\varepsilon,M}_k D_{\bar\phi}$ holds exactly for every sampled witness—Sinkhorn supplies the shared coupling and the target is its unregularized ground cost. Boundary invariance preserves only

declared non-descendant registers. During representation training, Core minimizes the following sum after normalizing every term to one development loss unit:

$$\mathcal{L}_{\text{core}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{allow}} + \mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{semantic}}, \qquad \mathcal{L}_{\text{semantic}} = \tfrac{1}{2}(\mathcal{L}_{\text{typed}} + \mathcal{L}_{\text{inv}}). \tag{7}$$

$\mathcal{L}_{\text{cons}}$ aligns direct encodings, recurrent tokens, and policy-visible register geometry, so the learned geometry is the one the policy consumes. The primary nonrecursive control regresses the same scalar distance to a matched finite-horizon Monte Carlo target. In the H1 factorial, the recursive cell uses the same horizon, normalized discount weights, terminal-zero boundary, and per-step control re-selection; the direct cell removes only the EMA next-distance bootstrap:

$$\widehat{T}^{(H)}_{\text{MC}} = \max_{\mathbf{u}\in\widehat{\mathcal{U}}_{k,H}} \sum_{\ell=1}^{H} \chi_\ell \mathbb{E}_{\Lambda^\star_{u_\ell,\ell,\varepsilon}}[c_\rho(\rho,\rho')], \qquad \chi_\ell = \frac{(1-\gamma)\gamma^{\ell-1}}{1-\gamma^H}. \tag{8}$$

Here $\mathbf{u} = (u_1, \ldots, u_H)$ and $\widehat{\mathcal{U}}_{k,H}$ reselects an admissible covered control after each paired successor. For the recursive H1 cell, the terminal-zero H-step Bellman unrolling is divided by $1-\gamma^H$ before regression, so its discounted mass matches the direct target above. This matched truncated target defines the confirmatory H1 factorial, whereas the full deployment metric retains the untruncated Bellman operator. A secondary prediction control decodes per-step signals and terminal effects from the same continuations. Full losses and normalization are in Appendix C.

Training and secondary extensions. Stage 1 trains factual state; Stage 2 freezes the reference policy and learns the counterfactual metric; Stage 3 freezes the registers and optimizes the policy. Two extensions sit outside Core: a route extension trains differentiable route/terminal heads on stop-gradient rollout distributions, admitting gradients into P/F_P, and optional sensitivity weighting reruns the frozen policy after register restoration under shared exogenous noise. Both are excluded from the confirmatory 2×2 cells.

3 Guarantees on the Covered Process

The statements below fix a cohort, a finite $\mathcal{U}_k$, and bounded signals, encoder, and kernels; they therefore characterize the covered process rather than the benchmark process as a whole. Proofs and the policy-conditioned operator appear in Appendix D, and coverage and policy-shift analyses (Section 4.5) measure behavior beyond coverage.

Covered fixed point and value (Appendix Theorems 1–2). $\mathcal{B}_k$ is a γ-contraction on bounded pseudometrics and has a unique fixed point d^*_k. For normalized scalarized return $V = (1-\gamma)\mathbb{E}[\sum_{t\geq 0}\gamma^t w^\top \rho_t]$,

$$|V^*_{k,\text{agent}}(\omega) - V^*_{k,\text{agent}}(\omega')| \leq d^*_k(\omega,\omega'). \tag{9}$$

The agent value restricts controls to the identity intervention and covered environment actions, so adding intervention controls makes d^*_k a conservative refinement: it can separate histories that identity-only bisimulation merges, never the reverse.

Learned-distance diagnostics (Appendix Theorems 3–4). Let $\mathcal{T}$ denote either the action-conditioned operator or its fixed-policy counterpart, $d_{\mathcal{T}}$ its fixed point, and $\widehat{\mathcal{T}}$ the held-out empirical operator. For $\widetilde{d}_\phi = D_\phi(q_\phi(h), q_\phi(h'))$,

$$\|\widetilde{d}_\phi - d_{\mathcal{T}}\|_\infty \leq \frac{\epsilon_{\text{res}} + \epsilon_{\text{est}} + \epsilon_{\text{miss}}}{1-\gamma}. \tag{10}$$

The three terms are operator residual, continuation-estimation error, and missed-control error; for a deployment witness rule frozen before evaluation, a finite-horizon occupancy version bounds time-zero error by $\sum_{t=0}^{H-1}\gamma^t\epsilon^{\text{occ}}_t$. Push-forward is exact for the learned pull-back metric at the population Wasserstein level, so finite-sample and entropic-coupling errors enter ϵ_{est} while encoder aliasing remains visible as operator residual. Because the witness is shared, $e_D = \sum_g \beta_g e_g$ and $|e_D| \leq \sum_g \beta_g |e_g|$: typed attribution refines the global Bellman target without defining a competing one.

4 EVALUATION

4.1 PROTOCOL AND COMPARISONS

The core benchmarks are ALFWorld, WebShop, τ^2-Retail, and ComplexMCP; AppWorld, ScienceWorld, WebArena-Lite, and ToolSandbox test transfer across interaction types. The main comparison includes Base/SFT, GRPO, GiGPO, ReBel, factorized state without interventions, matched Monte Carlo metric and future-prediction controls, RMT, and MemGPT; generic-latent and matched-splice controls remain in the complete appendix comparison. Every entry in the main table is rerun within benchmark under the same backbone, task split, policy budget, and paired evaluation; released numbers appear only in a separate protocol-qualified scale audit. Every DYNACAST configuration uses GRPO for Stage 3, so GiGPO enters solely as a standalone baseline. Exact sizes, metrics, adapters, and baseline contracts are in Appendices E and E.25.

Confirmation is organized around three named, falsifiable claims fixed before the runs. **H1 (bootstrap effect)** asks whether Bellman recursion improves on the matched finite-horizon Monte Carlo metric by more than one point in *both* the single-state and the $G/W/P$ architecture. **H2 (modular-interface effect)** asks whether the complete $G/W/P$ interface—register boundaries, typed readouts, and masked update parents—improves selectivity over a capacity-matched DAS subspace by more than four points under *both* objectives. **H3 (covered deployment audit)** is conjunctive and operational: coverage, action-bound violations, and policy-conditioned calibration must each clear a prespecified threshold. H1 and H2 are intersection–union tests, so a single failing simple effect falsifies the claim.

Each score is normalized within benchmark, averaged over tasks, then macro-averaged with equal benchmark weight. Training seed and task are crossed rather than nested, and the confirmatory interval uses a paired multiway cluster bootstrap over both factors with environment replicates resampled within each seed–task cell; ComplexMCP retains all 47 scenario-level paired differences. After the protocol freeze, a three-seed validity gate—requiring improved physical-do consistency, plan response, distance correlation, and long-horizon retention over both matched controls—makes a stop/go decision only, and passing it triggers five-seed confirmation. H1 and H2 are tested at one-sided $\alpha = .05$ in the fixed sequence H1$\rightarrow$H2$\rightarrow$H3; H3 is decided on the prespecified equal-weight core-four aggregate and requires 95% lower bounds above 90% and 80% for class and binding coverage, an action-bound violation upper bound below 5%, policy ECE below ten, and under three points of calibration degradation. Benchmark rows are heterogeneity diagnostics, not gates.

4.2 MAIN RESULTS ACROSS BACKBONES

Because AppWorld reports task-weighted TGC over the prespecified Test-N/Test-C union and ToolSandbox reports milestone similarity rather than exact success, scores are compared within benchmark before equal-benchmark aggregation, and Appendix Table 6 audits the primary blocks against released scale anchors. The headline comparison is a within-backbone, protocol-matched comparison rather than a cross-protocol leaderboard claim: DYNACAST-Full is compared against the future-prediction control expanded to the same $3.49\times$ envelope: 1.7 points on Qwen2.5 (95% CI $[0.9, 2.6]$), still 1.5 without ComplexMCP, and 1.8, 1.8, and 2.6 points on Llama, Gemma, and Mistral, respectively. The matched factorial preserves positive portfolio-level bootstrap effects under every backbone (Appendix Table 11). Full leads five to seven of eight columns per backbone, while the cross-family profile is distinctly nonuniform: Qwen is strongest on ALFWorld, Mistral has the highest DynaCAST score on both WebShop and ScienceWorld, and Gemma is second on ScienceWorld. Within the matched factorial $G/W/P$ recursion trails the Monte Carlo metric on WebShop. Appendix Figure 4 and Table 11 report benchmark-level uncertainty, factorial heterogeneity, and paired intervals across backbones; H1 is evaluated only on the matched cells, never on Full.

4.3 IDENTIFYING RECURSION, INTERVENTIONS, AND FACTORIZATION

Table 2 separates the three mechanisms the method claims. *Recursion.* Against the matched Monte Carlo metric—same scalar distance, ground cost, candidate controls, worst-control maximum, and continuation budget—Bellman bootstrapping raises the core-four task score by 1.5 points for a single state and 1.3 for $G/W/P$, i.e. all-eight H1 effects of 1.8 and 1.6 points. *Executable interventions.* Holding architecture and recursive loss fixed and varying only whether controls are executable, the core-four macro moves from 66.6 to 68.0 and the all-eight macro by 2.1 points. *Shared witness.*

Table 1: Eight-benchmark results for four fixed backbones. Each block uses paired tasks and the same policy budget. The Full-budget predictor receives DYNACAST-Full's $3.49\times$ compute envelope. Baseline citations appear in the first block; bold marks the best score within each backbone and benchmark.

Method	Core benchmarks				Extended benchmarks			
	ALF	WebShop	τ^2	CMCP	AppWorld	SciWorld	WA-Lite	ToolSB
Qwen2.5								
Base/SFT	65.7	49.3	28.6	20.0	24.8	61.3	18.7	36.6
GRPO (Shao et al., 2024)	73.4	56.1	33.2	29.7	30.1	65.8	23.6	42.7
GiGPO (w/o std.) (Feng et al., 2025)	85.4	68.0	33.1	36.7	29.5	71.6	30.1	48.9
ReBel (Tang et al., 2026)	92.8	74.7	43.1	36.2	33.1	72.8	29.4	54.5
Factorized (no intervention)	87.6	69.2	38.9	44.8	35.0	73.6	28.8	49.7
CF prediction (matched)	93.7	75.9	41.8	54.3	37.5	76.3	31.7	54.0
CF prediction (Full budget)	94.7	75.1	43.3	55.9	38.6	76.5	32.0	54.7
RMT (Bulatov et al., 2022)	89.4	72.2	39.2	47.9	39.3	**78.2**	29.7	50.6
MemGPT (Packer et al., 2023)	91.0	**77.2**	39.8	50.2	36.5	75.1	**34.1**	51.8
DYNACAST-Full	**95.0**	76.8	**44.5**	**59.1**	**41.0**	76.2	34.0	**57.8**
Llama-3.1-8B								
Base/SFT	63.4	51.6	27.9	17.8	23.5	60.8	18.3	35.8
GRPO	74.1	59.8	33.8	27.1	29.6	65.4	22.8	42.1
GiGPO (w/o std.)	83.9	67.1	34.5	35.2	29.0	70.9	29.2	48.0
ReBel	91.7	73.8	42.6	34.4	32.5	72.5	30.1	53.1
Factorized (no intervention)	86.9	69.9	39.7	42.9	34.0	73.1	30.5	49.2
CF prediction (matched)	92.6	75.6	43.4	48.1	37.0	75.5	32.7	54.1
CF prediction (Full budget)	93.9	74.6	45.9	50.4	37.0	75.6	32.9	54.0
RMT	89.8	72.9	39.4	44.9	39.0	**77.8**	30.0	51.0
MemGPT	90.8	**76.7**	40.9	46.8	36.1	74.8	34.3	52.0
DYNACAST-Full	**93.9**	76.4	**46.9**	**53.4**	**40.1**	75.7	**35.2**	**57.2**
Gemma-2-9B-IT								
Base/SFT	67.4	48.8	32.6	17.9	21.8	66.7	16.2	32.9
GRPO	75.9	55.6	37.8	26.4	27.1	70.5	20.4	39.1
GiGPO (w/o std.)	85.1	63.7	39.1	33.5	26.7	75.8	26.1	44.3
ReBel	92.5	70.8	46.8	33.9	30.2	77.2	27.0	49.5
Factorized (no intervention)	88.0	66.4	43.0	40.8	31.8	77.9	25.7	45.2
CF prediction (matched)	93.0	71.9	46.7	47.8	34.0	79.3	28.4	49.9
CF prediction (Full budget)	94.2	72.1	47.5	50.8	34.9	78.8	28.8	49.8
RMT	90.2	68.6	43.8	43.1	36.1	80.3	26.7	46.5
MemGPT	91.7	73.1	44.6	44.7	33.4	77.6	**31.0**	48.4
DYNACAST-Full	**94.6**	**73.9**	**49.2**	**52.8**	**37.4**	**80.8**	30.4	**52.3**
Mistral-7B-Instruct								
Base/SFT	58.7	54.8	24.3	19.2	24.6	60.5	21.7	37.2
GRPO	68.9	62.1	29.4	28.7	31.6	66.2	27.8	44.6
GiGPO (w/o std.)	78.1	69.6	30.7	35.8	32.4	72.9	34.2	51.0
ReBel	86.8	75.3	38.1	37.2	36.8	75.0	35.1	56.2
Factorized (no intervention)	81.7	72.0	35.6	45.1	37.2	75.6	34.4	53.6
CF prediction (matched)	88.3	76.8	39.6	52.2	40.4	79.1	37.5	58.5
CF prediction (Full budget)	90.1	76.0	42.0	54.8	41.0	78.4	38.0	58.5
RMT	84.2	73.6	34.9	47.0	40.8	80.7	35.2	54.1
MemGPT	87.0	76.4	37.3	49.3	38.9	77.8	**40.8**	57.2
DYNACAST-Full	**90.8**	**78.4**	**43.1**	**57.6**	**44.0**	**82.3**	40.2	**63.1**

Replacing the product coupling with the shared OT witness adds 1.1 task, 1.9 physical-do, and 3.2 selectivity points—the empirical counterpart of the exact compatibility in Equation 6. The modular interface improves selectivity by 6.5 points under the MC metric and 7.0 under recursion against a DAS control exchanging the same activation dimension as one four-token register; since attribution and invariance losses enter only after the 2×2 comparison, H1 gives $G/W/P$ no typed supervision the single state lacks. Across all seven nonempty register subsets, $G + W + P$ exceeds the best two-register macro ($G + W$) by 0.7 points, while $G + P$ is stronger on WebShop, so the ranking is not uniform.

The confirmatory intervals resolve the same contrasts in the same direction: H1 clears its one-point threshold in both architectures (lower bounds 1.1 single, 1.0 $G/W/P$), H2 clears its four-point threshold under both objectives (selectivity lower bounds 5.0 MC, 5.4 recursive), and the executable-control effect has a 1.3-point lower bound. H3 passes as a conjunction: class/binding-mass lower bounds are $91.2/85.8$ against thresholds of $90/80$, and the action-violation, policy-ECE, and ECE-increase upper bounds are $4.0/8.1/1.9$ against thresholds of $5/10/3$. The all-eight factorial interaction is -0.2 points (95% CI $[-0.9, 0.6]$), indicating a bootstrap gain of similar size in both architectures. Held-out residual and expanded-search diagnostics quantify covered-process approximation and candidate-search saturation; the exact finite SCM audit supplies the theorem-linked check when the control set is enumerable. Appendix Tables 42–44 give all per-benchmark configurations, register subsets, and replacements.

Table 2: Selected core-four ablations, each changing one factor at a time. Rows 1–3 and 6 form the matched 2×2 objective factorial (H1); row 4 removes executable interventions from the otherwise identical recursive learner; row 5 replaces the shared transport witness with a product coupling; the last three rows summarize register composition and the Core-to-Full extension. All three metrics are higher-is-better. Complete configurations and uncertainty appear in Appendix Tables 42–44.

State/objective	Task	Phys.-do	Select.
Objective and intervention contrasts			
Single + MC metric	64.7	82.2	73.2
Single + recursive	66.2	85.0	75.0
$G/W/P$ + MC metric	66.7	84.0	79.7
$G/W/P$ + recursive, identity only	66.6	74.1	73.5
$G/W/P$ + recursive, product coupling	67.1	83.9	78.8
$G/W/P$ + global recursive	68.0	85.8	82.0
Register composition and extensions			
Best two-register Core ($G + W$)	67.3	–	–
$G/W/P$ Core	68.0	87.4	85.1
$G/W/P$ Full	**68.9**	**89.1**	**87.3**

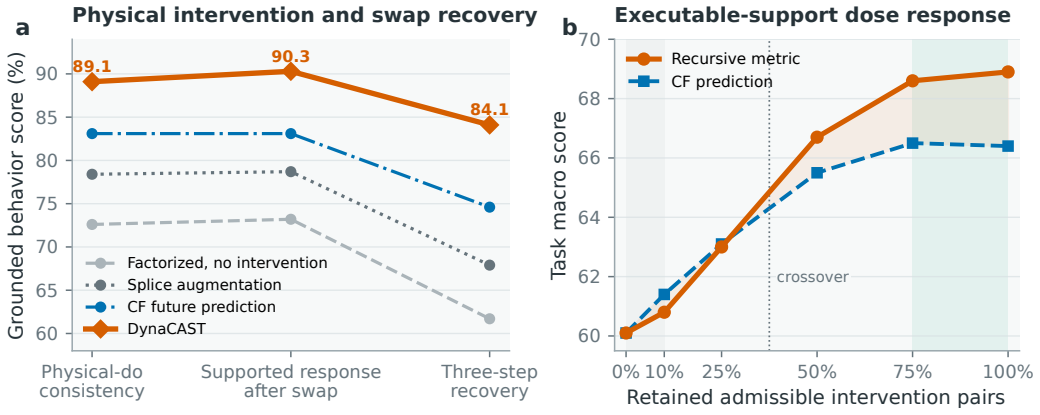

Figure 2: Executable-intervention mechanism evidence on the core four, comparing DynaCAST-Full with the matched future-prediction control. **Left:** physical-do consistency, supported response after an internal swap ($100-$unsupported effects), and three-step recovery; Full leads the matched predictor on all three. **Right:** task score as the retained admissible intervention-pair coverage increases, with factual data and optimization updates held fixed at every point; prediction leads at 10–25% support and the recursive metric overtakes it between 25% and 50%.

4.4 Intervention Semantics and Long-Horizon Gains

Figure 2 asks whether the task gains follow executable intervention semantics rather than donor copying, by contrasting two manipulations a copying account would not distinguish: a physical j_X changes simulator state and regenerates observations, whereas a belief-only W swap leaves the environment untouched and should be corrected once new evidence arrives. Full reaches 89.1 physical-do consistency, 9.7 unsupported effects, and 84.1 three-step recovery against 83.1/16.9/74.6 for matched future prediction—same continuations and budget, different learning object. The coverage sweep is a second matched test with factual data and optimization fixed: prediction leads while only 10–25% of admissible pairs are retained, the recursive metric overtakes between 25% and 50%, and saturates beyond 75%. Goal selectivity is scored against the new goal's executable terminal set and preserves only declared non-descendants, so correct goal-directed changes to X are not violations; route response uses differentiable route heads in training and closed-loop rollouts only at evaluation. Appendix Tables 16 and 18, Figure 5, and Appendix E give the physical/belief and aliasing comparisons, both coverage curves, and every denominator.

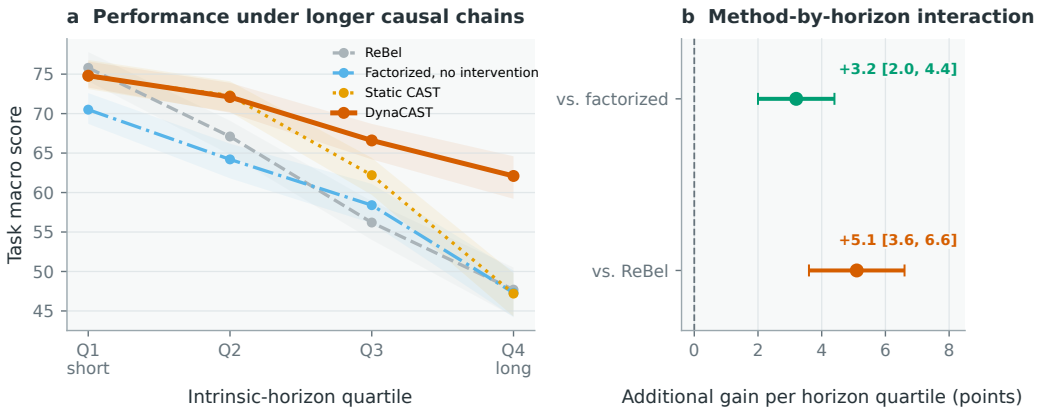

Figure 3: Long-horizon evidence. **Left:** performance across method-independent intrinsic-horizon quartiles, with crossed task–seed bootstrap 95% intervals. **Right:** the direct method-by-horizon test, reported as additional gain per quartile with 95% intervals; the interaction against factorized/no-intervention is $+3.2$ points per quartile (95% CI $[2.0, 4.4]$). Quartiles are frozen from method-independent oracle/reference traces before evaluation.

Figure 3 tests how the gain varies with intrinsic horizon. DynaCAST does not lead everywhere: ReBel leads Q1 by 1.0 point and Static CAST leads Q2 by 0.2, while DynaCAST leads Q3 and Q4 with a Q4 advantage of 14.4 points over ReBel and 14.8 over the factorized non-intervention control. The crossovers place the gain where a recursive intervention-grounded state should matter—when distant histories look locally similar but carry different future effects—and the method–horizon interaction confirms the trend at $+3.2$ points per quartile (95% CI $[2.0, 4.4]$), although adjacent Q1–Q2 differences need not be individually significant. Because the four equally weighted quartiles reconstruct each method's core-four macro, this refines the main estimand rather than replacing it with a favorable subset. Aliasing strata and plan-pair support appear in Appendix E.

4.5 Final-policy coverage and calibration

Appendix Figure 7 separates four evaluation operations, each isolating one source of change. Post-hoc metric refitting improves ECE from 10.8 to 8.6 while class and binding mass stay at 84.6/77.8, precisely because neither the policy nor the control library changes; development-only expansion raises coverage to 90.7/84.9 without moving policy TV. Alternating metric–policy refresh reaches 93.1/87.6 coverage and lowers class/binding TV to 0.066/0.089 under the same counterfactual budget, at a policy-distance ECE of 6.9 versus 6.1 under the frozen reference policy—an increase of 0.8 points (one-sided upper bound 1.9), inside the three-point H3 allowance. This alternating checkpoint supplies the DynaCAST-Full scores in Table 1.

Action-conditioned upper-bound violations and policy-conditioned ECE measure different objects, so coverage, calibration, and residual diagnostics are reported separately. Appendix Table 30 gives the alternating-policy checkpoint, benchmark rows, and resampling unit behind the H3 audit; Table 31 extends it to the expansion benchmarks and all-eight aggregate, while Tables 32 and 34 report residual median, tail, and maximum.

4.6 Matched controls and exact finite validation

The primary nonrecursive comparison used by H1 is Equation 8. Both H1 cells use a four-step truncated target with the same normalized discount weights, terminal-zero boundary, and per-step covered-control re-selection. The direct cell therefore removes only the EMA next-distance and Bellman bootstrap from the recursive learner; the secondary future-decoding control (Appendix Equation 45) predicts signals, canonical actions, typed events, termination, and EMA states from the same continuations.

Across 24 enumerated delivery SCMs, the observed fixed-point and $H = 4$ gaps remain below their exact theorem RHS at $\gamma \in \{.80, .90, .95\}$; Appendix Table 47 reports the complete check. Where

exact enumeration is unavailable, Appendix D reports held-out operator residuals and expanded-search gaps instead.

Compute interpretation. All confirmatory factorial cells share the $2.84\times$ forward-equivalent envelope: recursive and Monte Carlo cells use identical transport and candidate scoring, the decoder control spends the same budget on passes over the cached continuation tensors, and optimizer steps, physical resets, and continuation samples are identical. Full adds route calibration and 10% sampled closed-loop reruns, reaching $3.49\times$ (decomposed in the appendix), and the main table reports a predictor expanded to that same envelope, so the 1.7-point gain is not a compute artifact; the native-budget $1.64\times$ predictor appears only in the efficiency analysis. Deployment retains one masked update per register.

5 Positioning and Discussion

DynaCAST differs from structured memory, predictive world models, factual bisimulation, and causal interchange tests in its state criterion: executable intervention programs index the controlled futures compared by one recursive metric (Appendix Table 48). Frozen bindings make open-ended actions comparable, and the shared witness makes the global target exactly attributable to $G/W/P$ components. The registers implement this criterion; their names alone do not define it.

The matched contrasts isolate gains from recursion, executable controls, and shared transport. Physical resets, internal swaps, and equal-terminal route pairs then test different consequences of the learned state: adaptation to a changed environment, use of a policy-visible register, and route choice at a fixed terminal effect. Positive method–horizon interactions locate the largest gains in long-horizon, intervention-supported tasks, while the exact-kernel check and deployment diagnostics connect the covered theory to learned-state behavior. Together, these tests link register-level attribution to observable control without treating an internal swap as an environmental intervention.

6 Conclusion

DynaCAST grounds compact agent state in executable controlled futures: intervention programs index the equivalence being learned, frozen bindings close the open language action space, and one transport witness yields global and typed targets that agree by construction. Across eight benchmarks and four model families, matched contrasts attribute the improvement to these specific ingredients rather than to scale or supervision. Executable adapters and simulator resets remain training-time instruments: they build the controlled futures that define the state and are then discarded, leaving deployment with three recurrent registers, their masked updates, and the policy. Intervention access thus becomes a state-learning signal rather than a permanent execution dependency, and the learned memory stays auditable through the very behaviors that define it—a practical basis for long-horizon agent control in instrumented, resettable environments.

The empirical pattern also clarifies where the method should be used. Gains are small on short, nearly fully observed tasks because their histories already determine a narrow set of successful continuations. They widen when the same local observation is compatible with different goals, physical states, or routes, precisely the cases in which a factual memory can merge histories that require different actions. The horizon and aliasing analyses therefore provide an operational criterion for selecting the intervention library: coverage should be concentrated on distinctions that change executable futures, rather than on all differences in the observation stream. This criterion keeps the state compact while preserving the distinctions that affect control.

The resulting interface is compatible with existing agent stacks. An adapter only needs to expose typed interventions, canonical action classes, and resettable continuations; the deployed policy consumes the same recurrent state update as a standard language agent. The training machinery can consequently be audited at three independent levels: exact finite-state operator checks, held-out residuals on benchmark continuations, and closed-loop task outcomes. Agreement across these levels is the basis for the reported claims. It also makes failure analysis direct: a coverage deficit, an operator residual, or a task-level error points to a different part of the pipeline and can be addressed without changing the definition of the learned state.

AI USE STATEMENT

We used LLMs as a general-purpose writing assist tool during the preparation of this submission. Specifically, LLMs were employed for polishing the clarity and readability of text (e.g., refining sentence structure, improving grammar, and shortening overly verbose phrasing), and for assisting with the generation and editing of portions of utility code and formatting scripts. All research ideas, methodology design, experiments, analyses, and final writing decisions were conceived, implemented, and validated solely by the authors.

ETHICS STATEMENT

The experimental protocol confines agent actions and executable interventions to resettable simulators and benchmark sandboxes. Goal, world, and route modifications operate on benchmark state within these environments, with resets separating counterfactual branches from the original trajectory. The protocol excludes interventions on live user accounts or production services and does not require private user data. Third-party datasets, models, and tools remain subject to their respective licenses and access conditions. For applications beyond the evaluation environments, execution must respect user authorization and data-access boundaries, with human review for consequential actions. Learned state representations complement these operational safeguards; they do not confer additional execution permissions.

REPRODUCIBILITY STATEMENT

The main text specifies the learning objective, identifying comparisons, and statistical estimands; the appendix documents the proofs, experimental settings, and evaluation definitions. The anonymous supplementary artifact provides an executable DynaCAST reference implementation with recurrent registers, executable intervention controls, shared-witness objectives, training and evaluation entry points, checkpoint loading, and a runnable CPU configuration. It also includes unit tests for transport, interventions, target additivity, gradients, and checkpoint recovery; scripts for artifact verification, finite-state Bellman checks, control export, and evaluation; machine-readable protocol and benchmark manifests; tabulated result files; and a task-level result schema. Command-line instructions describe both the short verification run and the longer training/evaluation workflow.

REFERENCES

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Recurrent memory transformer, 2022. URL https://arxiv.org/abs/2207.06881.

Pablo Castro and Doina Precup. Using bisimulation for policy transfer in MDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):1065–1070, 2010. doi: 10.1609/aaai.v24i1.7751.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(6):10069–10076, 2020. doi: 10.1609/aaai.v34i06.6564.

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. MICo: Improved representations via sampling-based state similarity for markov decision processes, 2021. URL https://arxiv.org/abs/2106.08229.

Alejandro Cuadron, Pengfei Yu, Yang Liu, and Arpit Gupta. SABER: Small actions, big errors—safeguarding mutating steps in LLM agents, 2025. URL https://arxiv.org/abs/2512.07850.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019. doi: 10.18653/v1/P19-1285.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training, 2025. URL https://arxiv.org/abs/2505.10978.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011. doi: 10.1137/10080484X. URL https://doi.org/10.1137/10080484X.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks, 2022. URL https://arxiv.org/abs/2112.00826.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations, 2023. URL https://arxiv.org/abs/2303.02536.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning, 2019. URL https://arxiv.org/abs/1906.02736.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

David Ha and Jurgen Schmidhuber. World models, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels, 2019. URL https://arxiv.org/abs/1811.04551.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination, 2020. URL https://arxiv.org/abs/1912.01603.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *International Symposium on Artificial Intelligence and Mathematics*, 2006. URL http://anytime.cs.umass.edu/aimath06/proceedings/P21.pdf. Online proceedings paper; no page range assigned.

Yuanyang Li, Xue Yang, Longyue Wang, Weihua Luo, and Hongyang Chen. ComplexMCP: Evaluation of LLM agents in dynamic, interdependent, and large-scale tool sandbox, 2026. URL https://arxiv.org/abs/2605.10787.

Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. SwiftSage: A generative agent with fast and slow thinking for complex interactive tasks, 2023. URL https://arxiv.org/abs/2305.17390.

Xiao Liu, Tianjie Zhang, Yu Gu, et al. VisualAgentBench: Towards large multimodal models as visual foundation agents, 2024. URL https://arxiv.org/abs/2408.06327.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Ratsch, Sylvain Gelly, Bernhard Scholkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019. URL https://arxiv.org/abs/1811.12359.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities, 2025. URL https://arxiv.org/abs/2408.04682.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. URL https://arxiv.org/abs/2310.08560.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Wenyi Zhao, Yu Yang, Xinyue Yang, Jiadai Sun, Shuntian Yao, Tianjie Zhang, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM web agents via self-evolving online curriculum reinforcement learning, 2024. URL https://arxiv.org/abs/2411.02337.

Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling, 2019. URL https://arxiv.org/abs/1911.05507.

Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Benjamin Shih, John Winnicki, and Eric Darve. When does activation steering change what a model computes from?, 2026. URL https://arxiv.org/abs/2606.29522v2.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning, 2021. URL https://arxiv.org/abs/2010.03768.

Satinder Singh, Michael James, and Matthew Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pp. 512–519, 2004. URL https://arxiv.org/abs/1207.4167.

Stony Brook NLP. AppWorld official benchmark leaderboard. https://appworld.dev/appworld/leaderboard, 2026. Accessed 2026-08-28.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022. URL https://www.jmlr.org/papers/v23/20-1165.html.

Wenjie Tang, Minne Li, Sijie Huang, Liquan Xiao, and Yuan Zhou. Rewarding beliefs, not actions: Consistency-guided credit assignment for long-horizon agents, 2026. URL https://arxiv.org/abs/2605.20061.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents, 2024. URL https://arxiv.org/abs/2407.18901.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Cote, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader?, 2022. URL https://arxiv.org/abs/2203.07540.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca, 2023. URL https://arxiv.org/abs/2305.08809.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents, 2022. URL https://arxiv.org/abs/2207.01206.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Amy Zhang, Zachary C. Lipton, Luis Pineda, Kamyar Azizzadenesheli, Anima Anandkumar, Laurent Itti, Joelle Pineau, and Tommaso Furlanello. Learning causal state representations of partially observable environments, 2019. URL https://arxiv.org/abs/1906.10437.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction, 2020. URL https://arxiv.org/abs/2006.10742.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents, 2024. URL https://arxiv.org/abs/2307.13854.

APPENDIX CONTENTS

A Related Work 15

B Problem Formulation 16

C DynaCAST 18

D Covered Interventional Bisimulation: Guarantees 24

E Experimental Design 26

A Related Work

Long-horizon agent training. ALFWorld and ScienceWorld test language-mediated embodied control (Shridhar et al., 2021; Wang et al., 2022), WebShop and WebArena test grounded web interaction (Yao et al., 2022; Zhou et al., 2024), and AppWorld evaluates cross-application programs (Trivedi et al., 2024). ToolSandbox evaluates stateful conversational tool use (Lu et al., 2025), while τ^2-bench and ComplexMCP expose dual-control and interdependent-tool failures (Barres et al., 2025; Li et al., 2026). GRPO and GiGPO improve outcome- and step-level optimization, respectively (Shao et al., 2024; Feng et al., 2025). ReAct interleaves language reasoning with environment actions (Yao et al., 2023), Reflexion stores verbal feedback across trials (Shinn et al., 2023), and ReBel groups similar structured belief states for credit assignment (Tang et al., 2026). These systems make memory useful without defining a state equivalence under physical interventions. DynaCAST supplies this intervention-grounded equivalence while remaining compatible with existing agent scaffolds.

Long-context and recurrent memory. Transformer-XL reuses hidden states across segments (Dai et al., 2019), Compressive Transformers compress older activations (Rae et al., 2019), and Recurrent Memory Transformers propagate learned memory tokens (Bulatov et al., 2022). MemGPT instead manages explicit memory tiers around an LLM (Packer et al., 2023). These methods address capacity and access. They do not constrain two memories to be interchangeable under the same controlled future, so they motivate the capacity-matched recurrent baselines in our study.

State abstraction, bisimulation, and world models. MDP state abstraction characterizes which distinctions can be removed while preserving control (Li et al., 2006), and bisimulation metrics give a graded reward-and-transition notion (Ferns et al., 2011). Bisimulation-guided policy transfer uses distances between source and target states to select actions with bounded transfer loss (Castro & Precup, 2010). Predictive-state representations encode a controlled system by predictions of future tests (Singh et al., 2004); this motivates our history-measurable deployed representation, while the exact operator acts on simulator-augmented states and compares intervention–action-indexed distributions rather than a finite vector of tests. Approximate information states similarly characterize history compressions through future-control sufficiency (Subramanian et al., 2022). Sampling-based state-similarity objectives avoid explicit Wasserstein computation (Castro, 2020; Castro et al., 2021); our shared-witness construction keeps transport because one coupling must support exact global-to-component attribution. DeepMDP learns latent reward and transition models (Gelada et al., 2019); DBC directly matches latent distances to bisimulation targets (Zhang et al., 2020); and causal-state representations predict futures from partially observed histories (Zhang et al., 2019). World Models, PlaNet, and Dreamer learn compact dynamics for prediction, planning, or imagination (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020). DynaCAST changes the index set and factorization of the state metric. Instead of comparing only factual action transitions, it compares futures under executable task, environment, and route interventions paired with canonical actions. A frozen binding law makes those controls common across histories with open tool arguments, and one transport witness yields both the global target and its register components. Consequently, the $G/W/P$ labels are learned consequences of controlled futures rather than architectural annotations.

Causal representation and interchange interventions. Causal representation learning emphasizes that observational prediction alone does not identify causal variables (Scholkopf et al., 2021); unrestricted nonlinear disentanglement is not identifiable without inductive bias or supervision (Locatello et al., 2019). Interchange intervention training aligns neural variables with a supplied causal model (Geiger et al., 2022), and causal abstraction formalizes when such variables faithfully realize a lower-level system (Geiger et al., 2025). Activation-steering studies distinguish an output change from subsequent computation that uses an edited state according to its semantics (Shih et al., 2026). In contrast to one-shot interchange tests, DynaCAST recurs after environment transitions and compares multi-step, action-conditioned distributions. This construction turns interchange semantics into a trainable recurrent control metric with a covered, directly testable behavioral-abstraction claim. Distributed Alignment Search and its boundless variant search for causal variables in learned subspaces instead of assuming coordinate alignment (Geiger et al., 2023; Wu et al., 2023). We use that machinery to construct the best-development subspace-swap baseline for the unrestricted single-state H2 cell; DynaCAST's register boundaries are trained through recurrent executable-intervention targets.

Table 48 locates the novelty in a single learning object rather than three parallel modules: intervention programs determine which futures define equivalence, the binding kernel makes that definition well-posed for language actions, and the shared witness makes the global metric and typed recurrent state jointly trainable.

B Problem Formulation

Table 3: Notation used throughout the method and theory.

Symbol	Meaning
$h_t,\ \omega_t$	observable agent history; simulator-augmented state $\omega_t = (h_t, X_t, Y_t, R_t)$
$Z_t = (G_t, W_t, P_t)$	recurrent goal, world, and plan registers
$j \in \{j_G, j_X, j_R, j_{\mathrm{id}}\}$	executable goal, physical-world, route, or identity intervention
c, b, κ_c	canonical action class, concrete binding, and frozen within-class binding kernel
$u = (j, c),\ \mathcal{U}_k$	intervention–control and the covered controls in cohort k
$K_u,\ \mu_u^\rho$	augmented next-state kernel and immediate signal distribution
$D_\phi,\ d_\phi^g$	learned global distance and register-attributed distance, $g \in \{G, W, P\}$
$\mathcal{B}_k,\ T_g$	recursive Bellman operator and shared-witness component target
$\widehat{T}_{\mathrm{MC}}^{(H)},\ \mathcal{L}_{\mathrm{MC}}$	matched finite-horizon Monte Carlo target and its pairwise metric loss
$M, N_{\mathrm{ctrl}}, N_{\mathrm{adv}}$	continuations, policy candidates, and adversarial candidates per pair
$\epsilon_{\mathrm{res}}, \epsilon_{\mathrm{est}}, \epsilon_{\mathrm{miss}}$	operator residual, continuation-estimation error, and missed-control error

We model the benchmark as an environment structural causal model with latent physical or database state X_t, task specification Y_t, an executable route directive R_t, exogenous noise ξ_{t+1}, ν_{t+1}, and typed effects E_t:

$$X_{t+1} = f_X(X_t, a_t, \xi_{t+1}), \qquad o_{t+1} = f_O(X_{t+1}, \nu_{t+1}), \tag{11}$$

$$R_{t+1} = f_{\mathsf{R}}(R_t, X_{t+1}, Y_t, a_t), \qquad E_t = f_E(X_t, a_t, X_{t+1}), \tag{12}$$

$$r_t = f_R(Y_t, E_{\le t}). \tag{13}$$

The agent observes neither X_t nor the exogenous variables. A route directive is an adapter-validated environment instruction or milestone automaton emitted into the observable event stream; it is not the learned plan register. Its history is $h_t = (x_{\le t}^{\mathrm{usr}}, a_{<t}, o_{\le t})$, where x_t^{usr} carries task or user updates. We reserve $u = (j, c)$ exclusively for an intervention–control pair. Executable reset interventions are not, in general, Markov on h_t alone: the unmanipulated coordinates of (X_t, Y_t, R_t) must also be specified. We therefore define the exact theoretical state as the simulator-augmented program state $\omega_t = (h_t, X_t, Y_t, R_t) \in \Omega_k$. Simulator snapshots and frozen task programs provide exact samples of ω_t and support reset-and-continue operations; no posterior over hidden physical states is estimated. This augmented state is a training-time object, not an input to the deployed agent. Reset identifiers and physical states are used only to execute, pair, and audit targets. We seek an observable encoder and recurrent transition

$$q_\phi : h_t \mapsto Z_t, \qquad Z_{t+1} = F_\phi(Z_t, x_{t+1}^{\mathrm{usr}}, a_t, o_{t+1}), \tag{14}$$

that preserve covered future control while making typed interventions explicit.

Adapter-routed factorized state. The compact state is $Z_t = (G_t, W_t, P_t)$. G_t stores the active task and user constraints. W_t is a goal-independent predictive belief over the covered environment ontology, rather than a "task-relevant" belief whose definition would depend on G_t. P_t stores route or control progress. Raw dialogue events are not assumed to have a single semantic type. A frozen benchmark adapter routes each event e_{t+1} into goal-bearing $r_G(e_{t+1})$, world-evidence $r_W(e_{t+1})$, and shared discourse/provenance $r_S(e_{t+1})$ channels. One user utterance may populate both r_G and r_W; an environment observation may populate r_G only when the adapter marks it as resolving a task constraint. The update parents are

$$G_{t+1} = F_G(G_t, a_t, r_G(e_{t+1}), r_S(e_{t+1})), \tag{15}$$

$$W_{t+1} = F_W(W_t, a_t, r_W(e_{t+1}), r_S(e_{t+1})), \tag{16}$$

$$P_{t+1} = F_P(P_t, G_{t+1}, W_{t+1}, a_t, r_S(e_{t+1})). \tag{17}$$

The frozen adapter defines the graph semantics for each benchmark. W is grounded in routed action–observation evidence; goal-dependent relevance is computed by the plan and policy readouts.

Because P receives (G, W), it can still redundantly encode them. Selective interventions, router corruption, and leakage audits evaluate the resulting component semantics directly.

The shared channel r_S contains only source identity, turn boundary, timestamp/order, call identifier, and schema identifier. It excludes lexical content, entity values, arguments, result payloads, requested effects, and route milestones. Those semantic fields occur only in their routed channels. This restriction prevents r_S from becoming a full-event bypass around the parent masks; the unrestricted-parent control tests the remaining architectural gain.

Typed interventions. Three typed operations define the intervention semantics. A goal intervention j_G changes Y_t and emits the corresponding user update while holding X_t fixed. A physical world intervention j_X executes $\mathrm{do}(X_t^S = x'^S)$ in a resettable simulator and regenerates the observation from the modified state. An internal belief swap $\mathrm{do}(W_t = w')$ leaves X_t unchanged and is used only to test unsupported effects; it is not a sample from the physical causal kernel. An executable route intervention j_R sets R_t to one of two prevalidated route directives, emits that directive into the history, and is admitted only when both routes are executable from the same (Y_t, X_t) and realize the same terminal effect. The internal interchange $\mathrm{do}(P_t = p')$ is a representation-level test supervised by j_R continuations; it is not itself an environment kernel. Replacing P_t while forcing the same action class cannot alter the physical transition and is never used to define $K_{j,c}$.

For each intervention cohort $\mathcal{C}_k$, all histories use a finite, predeclared class library $\mathcal{U}_k = \mathcal{J}_k \times \mathcal{C}_k^A$. $\mathcal{J}_k$ includes the identity intervention. A canonical class $c \in \mathcal{C}_k^A$ is an environment-native action template or a tool schema with typed argument slots; raw language strings are not distinct classes. Let $\mathcal{B}(c, h)$ contain its concrete argument bindings at history h. Before representation training, the adapter freezes a history-dependent binding kernel $\kappa_c(db \mid h)$: deterministic for benchmark-native templates and a development-estimated categorical distribution over typed/binned bindings for open arguments. It never depends on the learned policy. Unsupported bindings map to a common typed invalid-action outcome with its declared cost. The unique class-conditioned augmented-state kernel is therefore

$$K_{j,c}(B \mid \omega) = \int_{\mathcal{B}(c,h(\omega))} K_{j,b}(B \mid \omega)\, \kappa_c(db \mid h(\omega)). \tag{18}$$

Thus common controls do not require identical raw legal-action lists, and class semantics cannot drift with policy argument selection. We separately report class coverage, binding-support coverage, and within-class binding drift between the deployed policy and κ_c.

When the class library is small we enumerate it exactly. Otherwise we form the deduplicated union of classes proposed at the two histories, rank it by the maximum of the two reference-policy probabilities, retain N_{ctrl} classes in total, and add N_{adv} classes proposed to maximize the current residual. This is a lower bound on the exact maximum. Small libraries permit an exact miss calculation; on large libraries, an audit with $4N_{\mathrm{ctrl}}$ retained policy classes plus $4N_{\mathrm{adv}}$ proposals measures only the expanded-search gap, not an upper bound on unseen controls. Coverage is reported over raw states, effect types, final-policy states, class probability mass, and concrete binding probability mass.

Desired equivalence. For $u = (j, c) \in \mathcal{U}_k$, let $K_u(d\omega' \mid \omega)$ be the exact intervention–control kernel on Ω_k, and let $\mu_u^\rho(\cdot \mid \omega)$ be the marginal distribution of the normalized immediate reward/effect vector on a separate signal space. Environment, observation, effect, physical-state, and route-program uncertainty are integrated in K_u; action-class sampling is not. The learned distance is restricted to the history-measurable family

$$\widetilde{d}_\phi(\omega, \omega') = D_\phi(q_\phi(h(\omega)), q_\phi(h(\omega'))). \tag{19}$$

Consequently, two augmented states with the same observable history receive the same representation. If their executable futures differ, the restriction produces a nonzero held-out Bellman residual; this directly audits partial-observability aliasing. A frozen reference policy $\pi_0(c \mid h(\omega), j)$ defines the policy-conditioned augmented kernel. The action-conditioned kernel supports the covered-control theorem; the π_0 mixture is a deployment diagnostic and baseline.

C DynaCAST

C.1 Intervention-grounded recurrent registers

Each register is a small set of persistent hidden tokens attached after a fixed backbone layer ℓ_s. A shared encoder first caches the current user update and action–observation transition. Two distinct query modules then update G and W in parallel from their permitted cached keys; a third module updates P after the new (G, W) are available. The modules share the frozen backbone attention and MLP weights but use separate query tokens, low-rank update parameters, and hard attention masks. Ordinary transcript keys are read-only during these passes. Gradients flow through all three new register values; donor values and exponential-moving-average targets are stop-gradient. This specifies the layer, parameter sharing, cache use, and gradient path without requiring a symbolic state at inference.

The token budget is matched to the generic-latent baseline. Static CAST, which extracts the three registers once and supervises one-step swaps, is retained as an ablation. The intervention objectives, rather than register names alone, establish the registers' behavioral semantics.

C.2 Typed bottleneck and leakage control

Let s_G, s_W, s_P be environment-provided labels for requested effects, goal-independent grounded attributes, and route milestones. Categorical targets use cross entropy, continuous targets are scaled to $[0, 1]$ and use squared error, and the mean within each interface is normalized to one loss unit. The allowed-head loss is

$$\mathcal{L}_{\text{allow}} = \sum_{g \in \{G,W,P\}} \mathbb{E}[\ell_g(q_g(s_g \mid Z^g), s_g)]. \tag{20}$$

For each forbidden pair $g \neq g'$, an adversary predicts $s_{g'}$ from Z^g. Define $\mathcal{L}_{\text{leak}} = -\sum_{g \neq g'} \ell(v_{g,g'}(Z^g), s_{g'})$ and solve $\min_{\phi,q} \max_v \mathcal{L}_{\text{allow}} + \lambda_L \mathcal{L}_{\text{leak}}$, implemented with a gradient-reversal layer. Forbidden labels exclude causal parents needed by the route head; in particular, P is not required to be statistically independent of (G, W). This loss regularizes cross-register leakage; semantic validity is evaluated through intervention behavior.

To prevent the registers from becoming an unbounded side channel, the default design allocates four hidden tokens to each component. At inference time, the policy receives the twelve persistent tokens, the current user update, the latest action–observation transition, and a short rolling transcript window. Older transcript tokens can be discarded after their task-relevant content is absorbed into Z_t. Token count and window size serve as capacity controls and are matched in the latent-state baselines.

C.3 Admissible multi-step trajectory splicing

Training examples contain two trajectory prefixes $(h_t, h'_{t'})$, their compact states (z, z'), an environment intervention program, and matched executable continuations. A goal program changes Y in the recipient reset. A world program changes declared coordinates of X and regenerates o. The corresponding internal register splice is supervised to reproduce that physical continuation, but is not itself called an environment intervention. An executable route program sets R to one of two validated directives with equal terminal effect; the paired internal P swap is evaluated separately. Programs that import an unsupported entity or break a precondition are excluded.

Formally, a proposed program is admitted before representation training when three environment-level checks hold. *Support* requires every referenced entity and action schema to exist in the recipient reset. *SCM agreement* requires equality of the non-intervened task/environment variables listed by the benchmark adapter, never equality under a learned abstraction. *Executability* requires at least one reference continuation that reaches a declared terminal effect; a plan pair requires two successful routes with the same terminal effect and a predeclared divergent action or milestone. These conditions define admissibility entirely at the environment level, independently of the learned abstraction. Ambiguous pairs are discarded and coverage is reported.

Physical goal/world programs are executed by resetting the simulator SCM; register splices are then trained to reproduce the resulting continuation. Belief-only swaps never substitute for these targets. A learned transition head may interpolate between executed points, but all primary losses and held-out

metrics use executable continuations. The two supervision-matched controls receive the identical resets, trajectories, and metadata.

Supervision provenance and effect vocabulary. Task templates supply requested-effect labels for G; reset APIs and transition logs supply grounded attributes and effect deltas for W; successful reference traces supply route milestones for P. Human annotation is used only to define the adapter ontology and to audit a sample of programs, not to label hidden registers per trajectory. Effects are benchmark-local typed tuples $(\text{entity}, \text{attribute}, \text{operation})$. The dedicated transfer study shares recurrent metric and update parameters only across runs using the same backbone, hidden width, and layer placement; it replaces the router, tuple embedding, and prediction heads and does not assume a universal aligned effect vocabulary.

C.4 Action-conditioned interventional bisimulation

The random immediate signal is $\rho = (r, E)$ on $\mathcal{R} = [0,1]^m$ after benchmark-declared normalization. With fixed nonnegative w_m summing to one, define $c_\rho(x, x') = \sum_m w_m |x_m - x'_m|$ and

$$\Delta_\rho(\omega, \omega'; u) = \mathcal{W}_{c_\rho}\left(\mu_u^\rho(\cdot \mid \omega), \mu_u^\rho(\cdot \mid \omega')\right). \tag{21}$$

Stochastic effects are therefore compared as distributions, not as unspecified sample or expectation differences. A common-random-number estimate is used only as a variance-reduced estimator when the simulator exposes its seed. For a cohort $\mathcal{C}_k$ with common nonempty control set $\mathcal{U}_k$, the operator on bounded pseudometrics is

$$(\mathcal{B}_k d)(\omega, \omega') = \max_{u \in \mathcal{U}_k} \left[(1-\gamma)\Delta_\rho(\omega, \omega'; u) + \gamma \mathcal{W}_d(K_u(\cdot \mid \omega), K_u(\cdot \mid \omega'))\right]. \tag{22}$$

The class c and frozen binding kernel are explicit in $u = (j, c)$; environment and observation randomness are integrated by K_u. The policy-conditioned object is defined separately, including its immediate signal. Fix a cohort-level intervention distribution $\nu_k(j)$ before state pairs or outcomes are examined, and a reference class policy $\pi_0(c \mid h(\omega), j)$ on the same covered library. The mixture weights never depend on the other history in a comparison. Let

$$K_k^{\nu,\pi_0}(\cdot \mid \omega) = \sum_{j,c} \nu_k(j)\pi_0(c \mid h(\omega), j) K_{j,c}(\cdot \mid \omega), \tag{23}$$

$$\mu_k^{\rho,\nu,\pi_0}(\cdot \mid \omega) = \sum_{j,c} \nu_k(j)\pi_0(c \mid h(\omega), j) \mu_{j,c}^\rho(\cdot \mid \omega), \tag{24}$$

with unsupported mass mapped to the typed invalid outcome. The fixed-policy operator is

$$(\mathcal{B}_k^{\nu,\pi_0} d)(\omega, \omega') = (1-\gamma)\mathcal{W}_{c_\rho}(\mu_k^{\rho,\nu,\pi_0}(\cdot \mid \omega), \mu_k^{\rho,\nu,\pi_0}(\cdot \mid \omega')) + \gamma \mathcal{W}_d(K_k^{\nu,\pi_0}(\cdot \mid \omega), K_k^{\nu,\pi_0}(\cdot \mid \omega')). \tag{25}$$

This defines one Markov signal/transition kernel for every augmented state in the cohort, so the same operator applies consistently to every pair and triangle. ν_k is uniform over the cohort's declared intervention families in the confirmatory study; an occupancy-weighted mixture is a separate diagnostic. This policy-conditioned distance changes with the policy; deployment diagnostics quantify that shift.

Each component distance only accesses its own register,

$$d_\phi^g(z_g, z'_g) = \frac{1}{m_g} \|\sigma(f_g(z_g)) - \sigma(f_g(z'_g))\|_1, \qquad D_\phi(z, z') = \sum_g \beta_g d_\phi^g(z_g, z'_g), \tag{26}$$

where fixed $\beta_g > 0$ sum to one. Thus D_ϕ is a bounded pseudometric and cannot rescale an arbitrary target. The practical loss regresses D_ϕ to a stopped-gradient sample of Equation 22. Independent typed fixed points would generally conflict with the global fixed point because maximization and convex combination do not commute. We instead use a *shared-witness Bellman attribution*. Let $g \in \{G, W, P\}$ index registers. The adapter freezes nonnegative signal assignments a_{gm} satisfying $\sum_g a_{gm} = w_m$, sets $\beta_g = \sum_m a_{gm}$, and defines

$$c_g(x, x') = \beta_g^{-1} \sum_m a_{gm} |x_m - x'_m|, \qquad c_\rho(x, x') = \sum_g \beta_g c_g(x, x'). \tag{27}$$

For each candidate control u, choose one global optimal signal coupling

$$\Lambda_u^\star \in \operatorname*{arg\,min}_{\Lambda \in \Pi(\mu_u^\rho(\cdot|\omega), \mu_u^\rho(\cdot|\omega'))} \mathbb{E}_{(x,x')\sim\Lambda} c_\rho(x, x'), \quad \Delta_g(\omega, \omega'; u) = \mathbb{E}_{\Lambda_u^\star} c_g(x, x'). \tag{28}$$

A deterministic lexicographic rule resolves nonunique empirical optima. Hence $\Delta_\rho = \sum_g \beta_g \Delta_g$ holds for the same coupling; no component solves an independent transport problem. For a pair and stopped-gradient distance, let $u^\star$ be selected by the practical global backup, let $\Lambda_\varepsilon^\star$ be its entropic signal coupling, and let $\Gamma_\varepsilon^\star$ be its entropic next-state coupling. Sinkhorn iterations obtain these couplings, but the target discards the entropy objective and self-transport corrections:

$$\widehat{\mathcal{B}}_k^{\varepsilon,M} D_{\bar\phi} = (1-\gamma)\mathbb{E}_{\Lambda_\varepsilon^\star} c_\rho + \gamma \mathbb{E}_{\Gamma_\varepsilon^\star} D_{\bar\phi}. \tag{29}$$

The resulting target is the unregularized ground cost under a regularized coupling, distinct from both OT_ε and the debiased Sinkhorn divergence $OT_\varepsilon(P,Q) - \frac{1}{2}OT_\varepsilon(P,P) - \frac{1}{2}OT_\varepsilon(Q,Q)$. This definition preserves exact shared-witness additivity while retaining a stable regularized coupling computation. The component target is

$$T_g(\omega, \omega') = (1-\gamma)\mathbb{E}_{\Lambda_\varepsilon^\star} c_g + \gamma \mathbb{E}_{(\tilde\omega,\tilde\omega')\sim\Gamma_\varepsilon^\star} d_{\bar\phi}^g(q_{\bar\phi}^g(h(\tilde\omega)), q_{\bar\phi}^g(h(\tilde\omega'))). \tag{30}$$

By construction, $\sum_g \beta_g T_g = (\widehat{\mathcal{B}}_k^{\varepsilon,M} D_{\bar\phi})(\omega, \omega')$ for every sampled pair, rather than only in expectation. Thus typed supervision attributes the same global Bellman witness; it does not introduce three incompatible operators. At the intervention boundary, the graph declares $N(j_G) = \{W\}$, $N(j_X) = \{G\}$, and $N(j_R) = \{G, W\}$. We impose invariance only on these causally preserved components:

$$\mathcal{L}_{\text{inv}} = \mathbb{E}_{(\omega,j)} \sum_{g\in N(j)} d_\phi^g(q_\phi^g(h(\omega)), q_\phi^g(h(\omega^j))). \tag{31}$$

In particular, a world or goal intervention may change P after replanning; Equation 31 does not impose false future invariance. Define the normalized global and typed residuals

$$\mathcal{L}_{\text{global}} = \mathbb{E}\left[\left(D_\phi(q_\phi(h(\omega)), q_\phi(h(\omega'))) - \text{sg}[(\widehat{\mathcal{B}}_k^{\varepsilon,M} D_{\bar\phi})(\omega, \omega')]\right)^2\right], \tag{32}$$

$$\mathcal{L}_{\text{typed}} = \frac{1}{3} \sum_{g\in\{G,W,P\}} \mathbb{E}\left[\left(d_\phi^g(q_\phi^g(h(\omega)), q_\phi^g(h(\omega'))) - \text{sg}[T_g(\omega, \omega')]\right)^2\right]. \tag{33}$$

We distinguish the architecture-neutral recursive objective from semantic attribution losses:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{global}}, \qquad \mathcal{L}_{\text{semantic}} = \tfrac{1}{2}(\mathcal{L}_{\text{typed}} + \mathcal{L}_{\text{inv}}). \tag{34}$$

All terms are normalized to unit development scale; the two semantic terms have fixed equal weight, so typed semantics add no tuning dimension. Because the typed targets sum exactly to the global target, their weighted absolute residual upper-bounds the aggregate Bellman residual; the held-out global residual is nevertheless measured directly for every named recursive configuration in Table 34. The strict factorial uses $\mathcal{L}_{\text{rec}}$ in both architectures; typed and invariance terms are then added sequentially only within $G/W/P$. Thus the factorial does not confound state structure with supervision unavailable to a single state. The global term preserves task-level behavior; the component terms and boundary invariance make semantic attribution a trained and falsifiable property rather than an interpretation of the aggregate metric. For sampled next histories, the transition target is explicitly

$$\widehat{\mathcal{W}}_{D_{\bar\phi}}\left(Q_{\bar\phi\#}\widehat{K}_u(\cdot \mid \omega), Q_{\bar\phi\#}\widehat{K}_u(\cdot \mid \omega')\right), \tag{35}$$

where $Q_{\bar\phi}(\omega) = q_{\bar\phi}(h(\omega))$. We estimate this term with Equation 29 over M independent executable continuations and register ground metric $D_{\bar\phi}$. The exact theorem is on augmented states, and Equation 35 evaluates the sampled transport for the history-measurable pull-back metric. Representation aliasing remains visible in the global operator residual. Finite M and entropic coupling regularization introduce estimation error. Table 32 varies M, and Table 33 varies the coupling scale; theoretical claims refer to exact Wasserstein distance rather than this estimator.

The recurrent state used after a splice can depend on intervention provenance, whereas $q_\phi(h)$ is a direct encoding. Metric equality alone cannot close this gap because D_ϕ has a policy-visible null

Table 4: Held-out direct–recurrent alignment on factual and executable-splice transitions.

Consistency objective	Token RMSE	Policy JS	Metric error
Metric only	0.184	0.071	**0.018**
Token + metric	0.061	0.029	0.021
Token + policy + metric (Full)	**0.057**	**0.011**	0.020

space. Let $Z^{\mathrm{dir}}_{t+1} = q_\phi(h_{t+1})$ and $Z^{\mathrm{rec}}_{t+1} = F_\phi(q_\phi(h_t), x^{\mathrm{usr}}_{t+1}, a_t, o_{t+1})$. We align both the raw register interface and the policy readout:

$$\mathcal{L}_{\mathrm{tok}} = \frac{1}{3} \sum_{g \in \{G,W,P\}} \frac{\|Z^{\mathrm{dir},g}_{t+1} - Z^{\mathrm{rec},g}_{t+1}\|_2^2}{n_g d_r}, \tag{36}$$

$$\mathcal{L}_{\mathrm{pol}} = \mathrm{JS}\big(\pi_\theta(\cdot \mid Z^{\mathrm{dir}}_{t+1}, h_{\mathrm{win}}), \pi_\theta(\cdot \mid Z^{\mathrm{rec}}_{t+1}, h_{\mathrm{win}})\big)\,, \tag{37}$$

$$\mathcal{L}_{\mathrm{met}} = D_\phi(Z^{\mathrm{dir}}_{t+1}, Z^{\mathrm{rec}}_{t+1})^2, \qquad \mathcal{L}_{\mathrm{cons}} = \tfrac{1}{3}(\bar{\mathcal{L}}_{\mathrm{tok}} + \bar{\mathcal{L}}_{\mathrm{pol}} + \bar{\mathcal{L}}_{\mathrm{met}}). \tag{38}$$

Bars denote deterministic unit-development-scale normalization; the terms have fixed equal weight and add no tuned coefficient. Both arguments see the same routed event and intervention provenance, donor identity is excluded, and the policy distribution is measured over canonical classes and binding bins. Held-out token error, policy JS, and metric error are all reported. The raw-token term removes affine differences that LayerNorm-only alignment would leave unconstrained. Together, the terms constrain every quantity consumed by deployment rather than inferring raw-token agreement from a non-injective distance head.

C.5 Secondary extensions: route control and sensitivity weighting

A plan register should change the route without redefining success. For a prevalidated executable pair (R, R') with common (Y, X), let (p, p') denote the induced internal plan-register values; these values are supervised targets, not environment interventions. Define a route signature $s = (a_{\mathrm{div}}, \eta_{1:L})$: the first divergent canonical action class and the ordered benchmark-level milestone sequence. Executable rollouts provide stop-gradient label distributions q_p over route signatures and e_p over terminal effects, excluding the donor history used for the current swap. They are not differentiable model outputs. Two trainable heads read the complete post-swap recurrent state:

$$\widehat{q}_\psi(p; z) = Q^{\mathrm{route}}_\psi(z^{P \leftarrow p}), \qquad \widehat{e}_\psi(p; z) = Q^{\mathrm{term}}_\psi(z^{P \leftarrow p}). \tag{39}$$

Both heads are masked to the benchmark-local valid signature/effect vocabulary. We use

$$\mathcal{L}_{\mathrm{term}} = \tfrac{1}{2} \sum_{v \in \{p,p'\}} \mathrm{CE}(\mathrm{sg}[e_v], \widehat{e}_\psi(v; z)) + \mathcal{W}_{d_E}(\widehat{e}_\psi(p; z), \widehat{e}_\psi(p'; z)), \tag{40}$$

$$\mathcal{L}_{\mathrm{route}} = \tfrac{1}{2} \sum_{v \in \{p,p'\}} \mathrm{CE}(\mathrm{sg}[q_v], \widehat{q}_\psi(v; z)) + [\delta_{\mathrm{path}} - \mathrm{JS}(\widehat{q}_\psi(p; z), \widehat{q}_\psi(p'; z))]_+. \tag{41}$$

The first term anchors each prediction to executable behavior; the terminal transport enforces equal outcomes, while the route margin requires distinct predicted paths. CE uses the reset, not an action token, as its statistical unit. The terminal-effect transport is an entropically regularized Sinkhorn distance over the finite benchmark-local effect vocabulary, so its gradient flows through both terminal-head distributions while the rollout labels remain stopped. Because the JS term compares full head distributions rather than two empirical rollout histograms, it has no plug-in sampling bias. We nevertheless freeze $\delta_{\mathrm{path}} = \mathrm{Quantile}_{.95}(\mathrm{null}) + 0.05$, where the null is obtained by permuting plan labels within development resets. Multiple raw routes may share one signature, and donor histories are held out from q_p, reducing route cloning. During Stage 2, gradients pass through Q^{route}_ψ, Q^{term}_ψ, P, and F_P; rollout labels, environment transitions, and the frozen policy are stop-gradient. Closed-loop executable rollouts are used again only for held-out route/terminal evaluation. Thus the extension requires neither a differentiable environment nor a policy-gradient estimator.

Register updates are deterministic, so they have no log-probability. The optional *closed-loop counterfactual sensitivity weighting* extension restores only one pre-update register value and reruns

the frozen policy from that point. The factual and restored rollouts share the same environment exogenous-noise stream and the same base uniform variates used to sample policy actions, but actions are regenerated from their respective logits. They may therefore take different actions while avoidable Monte Carlo variation is removed; holding actions fixed would make physical effects independent of the restored register under our SCM and is not used. To avoid weighting an objective by its own current error, a lagged frozen policy and EMA encoder define behavioral influence $I_{t,g}$ as the normalized sum of action-distribution JS, terminal-effect Wasserstein distance, and route-signature divergence between factual and restored reruns. We set $w_{t,g} \propto \exp(I_{t,g}/\tau)$, impose a development-frozen entropy floor of $0.6 \log N$ by mixing with the uniform distribution, stop its gradient, and refresh it once per 500 optimizer steps. Sensitivity reruns use a predeclared 10% sample of eligible updates. Let $\ell_{t,g}^{\text{future}}$ be the component continuation and attribution loss; we optimize

$$\mathcal{L}_{\text{sens}} = \sum_{t,g} \text{sg}(w_{t,g}) \ell_{t,g}^{\text{future}}. \tag{42}$$

Gradients pass through F_g and the g-specific heads, but not through donor states or weights. A swap is on-manifold only when a held-out register-density model places it above a development-frozen fifth-percentile threshold and the adapter validates its typed fields. Training weights use only on-manifold restorations; evaluation reports the all-swap estimand first, followed by the support-filtered diagnostic and exclusion rate. The resulting estimand is closed-loop behavioral sensitivity to an internal-state intervention, distinct from environment-level causal credit assignment and policy-gradient estimation.

C.6 Matched nonrecursive controls

The primary H1 control removes Bellman bootstrapping while retaining the same scalar pairwise metric, signal ground cost, candidate controls, worst-control maximum, continuations, and optimizer budget. Both H1 cells use a four-step truncated target, normalized discount weights, a terminal-zero boundary, and control re-selection after each paired successor. For control u_ℓ at continuation step ℓ, let $\Lambda^\star_{u_\ell,\ell,\varepsilon}$ be the entropic coupling computed from the two empirical signal distributions and define

$$\begin{aligned} \widehat{T}^{(H)}_{\text{MC}}(\omega, \omega') &= \max_{\mathbf{u} \in \widehat{\mathcal{U}}_{k,H}} \sum_{\ell=1}^{H} \chi_\ell \mathbb{E}_{(\rho,\rho') \sim \Lambda^\star_{u_\ell,\ell,\varepsilon}} c_\rho(\rho, \rho'), \\ \mathcal{L}_{\text{MC}} &= \text{Huber}\Big(D_\phi(q_\phi(h), q_\phi(h')), \text{sg}[\widehat{T}^{(H)}_{\text{MC}}] \Big). \end{aligned} \tag{43}$$

For the recursive H1 cell, the terminal-zero Bellman target is unrolled for H steps with per-step control re-selection. Writing $R_0 = 0$ and $R_{h+1} = (1-\gamma)\widehat{\Delta}_{\rho,h+1} + \gamma \mathbb{E}[R_h \mid \Gamma^\star_\varepsilon]$, the regression target is $R_H/(1-\gamma^H)$. Thus the recursive and direct H1 cells have the same truncated discount mass; they differ in the EMA next-distance bootstrap. The weights χ_ℓ are defined below and sum to one. The sequence $\mathbf{u} = (u_1, \ldots, u_H)$ reselects from the same covered library after each paired successor, so the direct target has the same control-selection rule and discount scale as the truncated recursive target. It contains neither an EMA next-state distance nor a Bellman backup. The maximizing sequence and its global couplings supply component diagnostic costs, so their weighted sum equals $\widehat{T}^{(H)}_{\text{MC}}$ for every sampled pair. These component costs are not optimized in the strict factorial. H1 therefore changes only finite-horizon direct regression versus recursive bootstrapping.

The secondary decoder control compares the recursive metric with fully supervised decoding of the same executable futures. It replaces pairwise metric regression with the following multi-step predictive objective while preserving every executable continuation. For horizon H, let $v_{t+\ell}$ denote the normalized reward/effect vector, $a^\sharp_{t+\ell}$ the canonical action class, $e_{t+\ell}$ the typed event tuple, $b_{t+\ell}$ the termination indicator, and $z^-_{t+\ell} = \text{sg}[q_{\bar{\phi}}(h_{t+\ell})]$ the EMA target state. The decoder at step ℓ receives the current compact state, intervention identifier, selected canonical control, and horizon embedding. With

$$\chi_\ell = \frac{(1-\gamma)\gamma^{\ell-1}}{1-\gamma^H}, \tag{44}$$

the matched prediction loss is

$$\mathcal{L}_{\mathrm{FP}} = \sum_{\ell=1}^{H} \chi_\ell \big[\mathrm{Huber}(\widehat{v}_{t+\ell}, v_{t+\ell}) + \mathrm{CE}(\widehat{a}^{\sharp}_{t+\ell}, a^{\sharp}_{t+\ell}) + \mathrm{CE}(\widehat{e}_{t+\ell}, e_{t+\ell}) + \mathrm{BCE}(\widehat{b}_{t+\ell}, b_{t+\ell}) + \mathcal{L}_{\mathrm{NCE}}(\widehat{z}_{t+\ell}, z^{-}_{t+\ell}) \big]. \tag{45}$$

Huber coordinates and categorical event fields are averaged within their type and normalized to one development loss unit. InfoNCE uses cosine similarity, temperature 0.07, and all other resets at the same horizon and event type in the batch as negatives; continuations from the same recipient–donor reset pair are excluded. All five prediction terms have unit weight after normalization; we perform no mined-negative or test-dependent calibration. Single-state and $G/W/P$ cells use the same five head families, targets, masks over missing targets, rollouts, optimizer steps, and forward-equivalent budget. Single-state heads read the full 12-token state; factorized heads read the register declared by the same frozen target schema. Thus $\mathcal{L}_{\mathrm{FP}}$ has access to every future variable used by the recursive cell but never receives a Bellman target, a learned ground metric, or a worst-control maximum.

The representation-stage Core objective is a sum of normalized losses:

$$\mathcal{L}_{\mathrm{core}} = \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{allow}} + \mathcal{L}_{\mathrm{cons}} + \mathcal{L}_{\mathrm{semantic}}. \tag{46}$$

For the four H1 factorial cells, the semantic term is absent and Stage 2 uses $\mathcal{L}_{\mathrm{factorial}} = \mathcal{L}^{(2)}_{\mathrm{agent}} + \mathcal{L}_{\mathrm{allow}} + \mathcal{L}_{\mathrm{cons}}$, with $\mathcal{L}^{(2)}_{\mathrm{agent}} \in \{\mathcal{L}_{\mathrm{MC}}, \mathcal{L}_{\mathrm{rec}}\}$. This is the exact single-variable objective contrast used by H1; the semantic ladder begins only after the factorial comparison. The allowed-interface loss has unit coefficient because each interface was already normalized to one loss unit. Continuation depth was fixed to four from development-only operator convergence before the protocol freeze; neither the subsequent validity gate nor confirmation task scores can revise it. The extension objectives are

$$\mathcal{L}_{\text{route-ext}} = 0.3(\mathcal{L}_{\mathrm{term}} + \mathcal{L}_{\mathrm{route}}), \quad \mathcal{L}_{\text{sens-ext}} = 0.3\mathcal{L}_{\mathrm{sens}}, \quad \mathcal{L}_{\text{leak-ext}} = 0.3\mathcal{L}_{\mathrm{leak}}. \tag{47}$$

The ablation adds each family alone and route plus sensitivity jointly; the 0.3 coefficient is fixed through development-scale normalization before confirmation. Each term is normalized by observed target count. The following configuration names are used throughout: *DynaCAST-Global* is the $G/W/P$ factorial cell trained with $\mathcal{L}_{\mathrm{rec}}$; *DynaCAST-Core* adds Bellman attribution and boundary invariance through $\mathcal{L}_{\mathrm{semantic}}$; and *DynaCAST-Full* adds route and closed-loop sensitivity terms. Leakage remains a separately reported audit regularizer (Table 24) and is not in the primary configuration. Main task tables report DynaCAST-Full. H1–H2 and the recursive-metric claim are decided by the strict factorial; the subsequent attribution-loss ladder tests the component-semantic claim. Unless a row explicitly says Global or Core, an unqualified DynaCAST result denotes Full.

C.7 Training stages and inference

Training has three stages with an explicit stage-dependent agent objective. Stage 1 minimizes $\mathcal{L}^{(1)}_{\mathrm{agent}}$, the teacher-forced language and tool-action negative log-likelihood plus factual versions of the signal, event, and termination heads, to initialize the policy and register updates. Stage 2 freezes that policy as π_0 and sets $\mathcal{L}^{(2)}_{\mathrm{agent}} = \mathcal{L}_{\mathrm{MC}}$ in the primary nonrecursive cells, $\mathcal{L}_{\mathrm{FP}}$ in the secondary decoder control, or $\mathcal{L}_{\mathrm{rec}}$ in recursive cells; the $G/W/P$ semantic ladder then adds its explicitly listed attribution and invariance terms. The route row jointly trains the differentiable heads in Equation 39 and backpropagates their supervised losses into P/F_P while treating executable rollouts as fixed labels. Route, leakage, and sensitivity extensions occur only in their designated rows. Stage 3 freezes the registers and metric and minimizes $\mathcal{L}^{(3)}_{\mathrm{agent}}$, the GRPO policy loss in every DynaCAST configuration. GiGPO is evaluated only as a standalone baseline. In Equation 46, $\mathcal{L}_{\mathrm{agent}}$ denotes the objective of the active stage. This schedule makes the reference kernel stationary. An optional alternating variant refreshes π_0 only between phases and refits the metric before the next policy phase. No loss differentiates through off-policy environment samples. The default configuration uses four-step counterfactual continuations, one factual batch per intervention batch, and an EMA target update slower than the encoder. The sensitivity analysis varies these choices independently.

Inference requires no donor trajectory and no counterfactual rollout. After each real transition, the three masked update modules revise G, W, and P; the policy then samples an ordinary language or tool action conditioned on the updated state. Thus trajectory splicing increases training cost but does not multiply environment calls at deployment. An optional diagnostic decoder can render each register as text for analysis, but rendered text is not fed back to the policy in the default method.

D COVERED INTERVENTIONAL BISIMULATION: GUARANTEES

The guarantees apply to a cohort $\mathcal{C}_k$ with a common finite control set $\mathcal{U}_k$, a fixed encoder and augmented-state kernels K_u, bounded signals, and $0 < \gamma < 1$. The space of $[0,1]$-valued pseudometrics is equipped with the supremum norm. Coverage and deployment-policy shift are evaluated by the dedicated audits in Section E. During learning, the EMA target approximates successive fixed-encoder operator evaluations; the analysis characterizes each frozen operator and its converged representation. Building on bisimulation machinery, we establish a well-typed augmented intervention-control process, a history-measurable representation class, an exact shared-witness attribution identity, and complementary worst-case and occupancy residual analyses for executable language-agent interventions.

Theorem 1 (fixed point). If $\Delta_\rho(\cdot,\cdot;u)$ is a pseudometric on Ω_k for each $u \in \mathcal{U}_k$, then $\mathcal{B}_k$ maps bounded pseudometrics to bounded pseudometrics and is a γ-contraction. Hence Equation 22 has a unique fixed point d_k^*. The identity intervention guarantees that the ordinary environment dynamics are included.

Proposition 1 (valid fixed-policy operator). For fixed cohort-level $\nu_k(j)$ and fixed observable-history policy $\pi_0(c \mid h(\omega), j)$, Equations 25 define a single Markov signal and transition kernel on $\mathcal{C}_k$. Consequently $\mathcal{B}_k^{\nu,\pi_0}$ maps bounded pseudometrics to bounded pseudometrics, is a γ-contraction, and has a unique fixed point d_k^{ν,π_0}. This conclusion does not hold for pair-dependent mixture weights $\nu(j \mid h, h')$, which are excluded by definition.

Lemma 1 (push-forward is exact for the learned pull-back metric). Fix an encoder q and a bounded pseudometric D on its representation space, let $Q_q(\omega) = q(h(\omega))$, and define $\widetilde{d}(\omega,\omega') = D(Q_q(\omega), Q_q(\omega'))$. If the augmented and representation spaces are standard Borel, then for any augmented-state distributions P_0, P_1,

$$\mathcal{W}_{\widetilde{d}}(P_0,P_1) = \mathcal{W}_D(Q_{q\#}P_0, Q_{q\#}P_1). \tag{48}$$

At the population Wasserstein level, push-forward therefore exactly evaluates augmented-kernel transport for the learned history-measurable pull-back metric. The practical entropic-coupling estimator additionally contributes finite-M and regularization error. The representation can still alias augmented states that the fixed point separates; such aliasing appears as a nonzero operator residual.

Theorem 2 (covered value preservation). Let the normalized discounted signal be $V = (1-\gamma)\mathbb{E}[\sum_{t\geq 0}\gamma^t w^\top \rho_t]$, using the same fixed weights as Δ_ρ. For the augmented MDP whose controls are $u \in \widetilde{\mathcal{U}}_k$, the optimal scalarized value satisfies

$$|V_k^*(\omega) - V_k^*(\omega')| \leq d_k^*(\omega,\omega'). \tag{49}$$

For a fixed reference policy, the same bound holds with the policy-conditioned fixed point d_k^{ν,π_0} from Proposition 1 and V^{ν,π_0}. An unnormalized return introduces the factor $1/(1-\gamma)$. Proposition 3 handles deployment under a changed state-dependent policy through explicit coverage, class-policy, and binding-policy shift terms.

Executable interventions are training-time probes, not actions available to the deployed agent. Let $\mathcal{U}_k^{\mathrm{id}} \subset \mathcal{U}_k$ retain only the identity intervention and covered environment action classes. Because the maximum in Equation 22 includes this subset, the same induction gives

$$|V_{k,\mathrm{agent}}^*(\omega) - V_{k,\mathrm{agent}}^*(\omega')| \leq d_k^*(\omega,\omega'). \tag{50}$$

Thus the intervention-grounded metric preserves ordinary covered-agent value, while potentially distinguishing more states than the identity-only metric. This refinement exposes latent differences that factual occupancy can miss. Action-bound tightness and policy-conditioned calibration then measure how this additional resolution transfers to deployment behavior.

Theorem 2 controls the declared scalarized effect signal. For adapters with binary terminal success, ρ includes an absorbing terminal-success coordinate, and the theorem controls its discounted counterpart. Benchmark- native pass[1], exact-success, and milestone scores are reported as separate task-level outcomes with their corresponding metric definitions.

Theorem 3 (learned-distance approximation). Let $\mathcal{T}$ be either $\mathcal{B}_k$ or its fixed-π_0 mixture, with fixed point $d_{\mathcal{T}}$, and let $\widetilde{d}_\phi(\omega,\omega') = D_\phi(q_\phi(h(\omega)), q_\phi(h(\omega')))$. Let $\widehat{\mathcal{T}}$ be its held-out empirical estimate on proposed controls $\widetilde{\mathcal{U}}_k \subseteq \mathcal{U}_k$, and define

$$\epsilon_{\mathrm{res}} = \|\widetilde{d}_\phi - \widehat{\mathcal{T}}\widetilde{d}_\phi\|_\infty, \tag{51}$$

$$\epsilon_{\mathrm{est}} = \|\widehat{\mathcal{T}}\widetilde{d}_\phi - \mathcal{T}_{\widetilde{\mathcal{U}}_k}\widetilde{d}_\phi\|_\infty, \tag{52}$$

$$\epsilon_{\mathrm{miss}} = \|\mathcal{T}_{\widetilde{\mathcal{U}}_k}\widetilde{d}_\phi - \mathcal{T}\widetilde{d}_\phi\|_\infty. \tag{53}$$

Then

$$\|\widetilde{d}_\phi - d_{\mathcal{T}}\|_\infty \le \frac{\epsilon_{\mathrm{res}} + \epsilon_{\mathrm{est}} + \epsilon_{\mathrm{miss}}}{1-\gamma}. \tag{54}$$

The theorem is deterministic. On finite, exhaustively enumerable control sets, all three terms can be upper bounded and the result is directly instantiated. In large language-action spaces, enlarged candidate search estimates search saturation, while Equation 54 retains the population missed-control term ϵ_{miss}. Held-out residuals, continuation- sampling error, and expanded-search gaps are reported separately to preserve the distinction between observed candidate coverage and the full control set.

Theorem 4 (support-wise and fixed-witness occupancy propagation). For $t = 0, \dots, H-1$, let $\mathcal{D}_t \subseteq \Omega_k^2$ be a prespecified pair cohort, let exact and learned recursions share terminal distance zero, and write $d^{(H)}$ and $\widehat{d}^{(H)}$ for their time-zero outputs. Require product-support closure: for every $(\omega,\omega') \in \mathcal{D}_t$ and every control u used by either backup,

$$\operatorname{supp} K_u(\cdot \mid \omega) \times \operatorname{supp} K_u(\cdot \mid \omega') \subseteq \mathcal{D}_{t+1}. \tag{55}$$

Thus both the exact- and learned-cost optimal couplings are supported on the next cohort. If the total one-step residual is at most $\varepsilon_t^{\mathrm{sup}}$ on $\mathcal{D}_t$, backward induction gives

$$\|\widehat{d}^{(H)} - d^{(H)}\|_{\infty,\mathcal{D}_0} \le \sum_{t=0}^{H-1} \gamma^t \varepsilon_t^{\mathrm{sup}}. \tag{56}$$

For the occupancy statement, fix before evaluation a measurable witness rule w_t that selects one covered control and coupling for each pair, and let $\mathcal{T}_t^w$ denote its non-maximizing backup. Write $d_w^{(H)}$ and $\widehat{d}_w^{(H)}$ for the corresponding exact and learned recursions. Let η_0 be a frozen deployment-visitation law over pairs and propagate η_t through w_t. If $\mathbb{E}_{\eta_t}|\widehat{d}_t - \mathcal{T}_t^w \widehat{d}_{t+1}| \le \varepsilon_t^{\mathrm{occ}}$, then

$$\mathbb{E}_{\eta_0}|\widehat{d}_w^{(H)} - d_w^{(H)}| \le \sum_{t=0}^{H-1} \gamma^t \varepsilon_t^{\mathrm{occ}}. \tag{57}$$

The first statement controls the full action-conditioned maximum on its covered support; the second controls the prespecified witness process. Both follow from backward recursion and provide statewise and visitation-weighted views of approximation quality. For a common error, the multiplier is $S_H = (1-\gamma^H)/(1-\gamma)$. Numerical instantiation requires an upper bound for every control used by the stated backup. The enumerated-SCM experiment satisfies this requirement; the large-benchmark evaluation instead reports held-out candidate-search saturation and operator fit.

Proposition 2 (Bellman-attribution compatibility). For one sampled global witness $(u^\star, \Lambda_\varepsilon^\star, \Gamma_\varepsilon^\star)$, let $e_g = d_\phi^g - T_g$ and $e_D = D_\phi - \widehat{\mathcal{B}}_k^{\varepsilon,M} D_{\bar{\phi}}$. Equation 30 gives

$$e_D = \sum_g \beta_g e_g, \qquad |e_D| \le \sum_g \beta_g |e_g|. \tag{58}$$

Consequently, component residuals bounded by ε_g imply an aggregate empirical Bellman residual bounded by $\sum_g \beta_g \varepsilon_g$. This result uses the shared global witness; independently maximized component operators lie outside the construction. Boundary invariance adds a semantic constraint without changing the target identity. Theorem 2 characterizes the exact global fixed point targeted by Global, Core, and Full. Table 34 evaluates the global residual of each learned configuration under the same held-out operator audit; route and sensitivity extensions retain the same global target.

Proposition 3 (explicit deployment-shift penalty). For a deployed policy π, let $m_k(\omega) = \sum_j \nu_k(j) \sum_{c:(j,c)\in\mathcal{U}_k} \pi(c \mid h(\omega), j)$ be covered joint intervention–class mass. Let $\bar{\pi}_k(j, c \mid \omega)$ be the renormalized covered joint distribution proportional to $\nu_k(j)\pi(c \mid h(\omega), j)$, and define $\bar{\pi}_{0,k}$ analogously. Let $\mathcal{S} \subseteq \Omega_k$ contain the two compared initial states and be forward invariant under both the deployed concrete-action kernel and the reference class-plus-binding kernel. Assume normalized scalar signals in $[0, 1]$ and define

$$\epsilon_{\text{out}} = \sup_{\omega\in\mathcal{S}}[1 - m_k(\omega)], \tag{59}$$

$$\epsilon_{\text{class}} = \sup_{\omega\in\mathcal{S}} \text{TV}(\bar{\pi}_k(\cdot \mid \omega), \bar{\pi}_{0,k}(\cdot \mid \omega)), \tag{60}$$

$$\epsilon_{\text{bind}} = \sup_{\omega\in\mathcal{S}} \mathbb{E}_{(j,c)\sim\bar{\pi}_k(\cdot|\omega)} \text{TV}(\pi(\cdot \mid c, h(\omega)), \kappa_c(\cdot \mid h(\omega))). \tag{61}$$

Then the learned-distance deployment bound is

$$|V^\pi(\omega) - V^\pi(\omega')| \le \widetilde{d}_\phi(\omega, \omega') + \frac{\epsilon_{\text{res}} + \epsilon_{\text{est}} + \epsilon_{\text{miss}} + 2(\epsilon_{\text{class}} + \epsilon_{\text{bind}} + \epsilon_{\text{out}})}{1 - \gamma}. \tag{62}$$

Here the three approximation errors are evaluated for the fixed-π_0 operator; the action-conditioned residual is reported separately. The theorem requires suprema on the common forward-reachable set $\mathcal{S}$; empirical visitation maxima and high-probability upper bounds are finite-support diagnostics rather than replacements for this condition. Visitation means are reserved for Equation 57. We report all three shift terms under final-policy visitation and test deployment with direct held-out calibration. A post-hoc metric refit can improve calibration but cannot change a fixed policy's covered mass; coverage changes only when the frozen control library is enlarged or the policy is optimized again. Equation 62 applies to the declared covered control process and its measured policy shift.

Lemma 2 (population route-predictor non-collapse). For a feasible plan pair whose reference route distributions differ and $\delta_{\text{path}} > 0$, a route predictor independent of its swapped P input has population JS zero and therefore incurs hinge loss δ_{path}. Thus ignoring P is not a zero-loss solution for the route predictor. The permutation statistic in Equation 41 uses a development-frozen null quantile, with calibration audited by held-out label permutations. This result establishes use of the plan interface by the differentiable route predictor; held-out closed-loop route swaps establish use by the deployment policy.

Semantic guarantees and validation. The theory establishes covered behavioral equivalence for the aggregate metric. Register semantics are specified by typed attribution and boundary invariance, whose shared-witness compatibility is given by Proposition 2. The component- swap and closed-loop evaluations measure semantic localization and policy stability directly, including the cross-register redundancy permitted by masked updates. Proofs appear in the appendix.

Nontrivial example. Consider a partially observed delivery task with two indistinguishable textual prefixes, hidden inventory and door state X, a requested destination Y, and two valid routes around a locked door. Goal interventions change the destination, world interventions toggle inventory or the lock in the simulator, and route interventions select one of the two successful directives. The common covered action set contains inspect, unlock, move, and deliver. The observation alone aliases the histories, yet reward and transition kernels differ after inspect or unlock; consequently d_k^* is nonzero and the abstraction problem is not a one-state or fully observed special case. This finite model satisfies the theorem assumptions without implying unique recovery of its three registers.

E Experimental Design

E.1 Benchmarks and metrics

The core suite isolates embodied partial observability (ALFWorld), transactional web interaction (WebShop), dual-control dialogue (τ^2-bench Retail), and interdependent tool use (ComplexMCP). Four transfer environments add cross-app programs (AppWorld), long physical-state trajectories (ScienceWorld), realistic multi-route websites (WebArena-Lite), and stateful conversational tools (ToolSandbox). Table 5 states the primary metric and stressor for each environment.

Table 5: Benchmark portfolio, primary metrics, and stressors.

Benchmark	Primary metric	Primary stressor
ALFWorld	success rate	partial observability
WebShop	success rate	grounded transactional search
τ^2 Retail	pass[1]	user/world dual control
ComplexMCP	exact success	effect dependencies and failures
AppWorld	task goal completion	cross-app API state
ScienceWorld	normalized score	horizon and physical dynamics
WebArena-Lite	success rate	open-ended route choice
ToolSandbox	milestone score	stateful conversational tools

Evaluation denominators follow the frozen official or cited protocol rather than dataset headline sizes: 128 ALFWorld held-out tasks in the released ReBel protocol; 500 frozen WebShop tasks; a prespecified stratified 40-task evaluation subset of the 115 released τ^2 Retail tasks; 47 ComplexMCP v2 test scenarios; AppWorld Test-N 168 plus Test-C 417 (585 total); all official ScienceWorld test variations within each of 30 subtasks (25% per subtask, analyzed as strata); the released 165-task VAB-WebArena-Lite refinement (Liu et al., 2024); and 1,032 ToolSandbox scenarios. Database users, full-dataset totals, and development variations are excluded from the independent test-task count.

E.2 Baselines and controls

The complete protocol uses eleven universal baselines or supervision-matched controls; the main table presents the nine matched baseline/control comparisons most relevant to the identification claims. All newly executed cells share the same backbone checkpoint, training-task instances, seed schedule, environment-interaction ceiling, and evaluation split; each method generates its own on-policy trajectories under this common contract: (1) Base/SFT; (2) GRPO; (3) GiGPO; (4) ReBel; (5) a generic latent state with the same persistent-token budget as DynaCAST; and (6) factorized $G/W/P$ state with modular dynamics but no trajectory intervention; (7) a capacity-matched Recurrent Memory Transformer (RMT) (Bulatov et al., 2022); and (8) a MemGPT-style tiered retrieval memory (Packer et al., 2023); (9) matched splice augmentation; (10) the matched Monte Carlo metric; and (11) matched counterfactual future prediction. RMT receives the same twelve persistent tokens and update depth as DynaCAST. MemGPT receives the same policy context window plus a retrieval store built only from the allowed trajectory, with retrieval tokens and inference calls matched to ReBel. Static one-step CAST is reported as a separate ablation. Environment-specific comparisons include the same-protocol SABER audit on mutation-sensitive τ^2 tasks (Cuadron et al., 2025) and the strongest reproducible benchmark-native agent whose released protocol can be matched at experiment freeze. These specialist results are reported in Table 6 and the paragraph following it. Released anchors are confined to that audit; every method in the main result table is rerun under the within-benchmark contract and retains the task–seed records required for paired resampling.

Mechanism controls remove the supervision confound. *Splice augmentation* receives every factual and counterfactual trajectory, effect label, reset, and optimizer update used by DynaCAST, but treats them as ordinary supervised examples. The primary decomposition is a strict 2×2: one unrestricted 12-token state versus three four-token $G/W/P$ registers, crossed with the finite-horizon Monte Carlo metric $\mathcal{L}_{\mathrm{MC}}$ versus the architecture-neutral global recursive loss $\mathcal{L}_{\mathrm{rec}}$. These objectives share scalar pairwise regression, signal ground cost, candidate controls, worst-control maximization, a four-step horizon, normalized discount weights, a terminal-zero boundary, and per-step control re-selection; only the recursive target contains an EMA next-distance Bellman bootstrap. All four cells receive identical routed channels (concatenated for the single state), raw event text, metadata, labels, metric-head count, resets, continuation samples, optimizer steps, and forward-equivalent compute. Single-state heads all read the same 12-token state; factorized heads read their declared register. Neither typed component targets nor boundary invariance appears in these four cells. Three subsequent $G/W/P$ rows add global recursion, Bellman attribution, and boundary invariance cumulatively. A further diagnostic keeps three registers and routed inputs but removes all update-parent masks, isolating modular dynamics from routing. Route and sensitivity extensions are added one at a time only after this factorial test.

All eight benchmarks use one fixed checkpoint for each of four model families: Qwen2.5-7B (Qwen et al., 2024), Meta Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Gemma-2-9B-IT (Gemma Team et al., 2024), and Mistral-7B-Instruct (Jiang et al., 2023). Each block reruns the complete main-table

method suite and the four strict-factorial cells with the same adapters, task splits, five seeds, and within-benchmark paired inference; the four estimands are reported as separate, equally weighted confirmations. Exploratory development selects the architecture, tolerances, and hyperparameters on development tasks, after which the complete protocol is frozen. A subsequent three-seed validity gate on one core environment is restricted to a stop/go decision and cannot change thresholds, contrasts, adapters, or model settings; passing it triggers five-seed confirmation on the core and expansion benchmarks. Secondary sweeps use three seeds and remain outside H1–H3. Reporting preserves method pairing and uses 5,000 multiway-cluster bootstrap replicates: training seeds and task instances are independently resampled as crossed factors, then environment replicates are resampled within each selected seed–task cell. Two-sided 95% percentile intervals use the 2.5th and 97.5th percentiles; confirmatory one-sided 95% lower or upper bounds use the 5th or 95th percentile, respectively. A crossed random-intercept model with fixed method effects and random seed and task intercepts is the prespecified sensitivity analysis. Hyperparameters are selected on development tasks; no test result may tune an intervention rule.

For cross-benchmark macro-averages, each task score is first mapped to $[0, 100]$ using the benchmark's documented minimum and maximum. We average over tasks within a benchmark and then give each benchmark equal weight. Statistical tests remain at the task level; a macro-average never pools episodes from different benchmarks as exchangeable samples. The primary estimand treats the eight-benchmark portfolio as fixed and assigns equal benchmark weight, independent of task count. Sensitivity analyses report a six-benchmark macro excluding ALFWorld and WebShop and an inverse-variance random-effects summary. The latter uses the DerSimonian–Laird moment estimator with each benchmark's paired multiway-bootstrap standard error, treats benchmark effects as random, and reports I^2 from Cochran's Q; it does not replace the fixed-portfolio estimand.

The analysis pipeline uses one task-level long-form result table keyed by benchmark, task, environment replicate, training seed, backbone, state architecture, objective, and extension set. Benchmark means, factorial contrasts, uncertainty intervals, figures, and all appendix tables are generated from that table in one analysis command; aggregate values are generated rather than manually transcribed between tables. Recurrent configurations and estimands therefore reproduce the same derived value across ablation, efficiency, and sensitivity tables, while distinct diagnostics retain separate result keys.

For the confirmatory stage, we prespecify two primary hypotheses and one deployment criterion with a fixed gatekeeping order H1 $\rightarrow$ H2 $\rightarrow$ H3 at one-sided $\alpha = 0.05$. Let $Y_{s,o}$ be the equal-benchmark macro score for state architecture $s \in \{\text{single}, \text{GWP}\}$ and objective $o \in \{\text{MC}, \text{rec}\}$. The confirmatory H1 estimand gives equal weight to all eight benchmarks; the four core benchmarks remain a separately reported replication stratum. **H1 (bootstrap effect)** is the intersection–union claim that the recursive metric improves over its direct Monte Carlo counterpart in both architectures by a minimum meaningful one point:

$$\Delta_{H1}^{\text{single}} > 1.0 \quad \wedge \quad \Delta_{H1}^{\text{GWP}} > 1.0, \qquad \Delta_{H1}^{s} = Y_{s,\text{rec}} - Y_{s,\text{MC}}. \tag{63}$$

Let $S_{s,o}$ be the predeclared core-four mean of goal and world selectivity. A raw single-state swap structurally privileges the factorized model because the single state has no named component boundary. The single-state baseline therefore receives a development-selected intervention oracle. On development pairs, a Boundless-DAS-style alignment learns separate orthogonal subspaces for goal and world content in the flattened 12-token activation. Let d_r be the activation width of one persistent token. Swapping one four-token $G/W/P$ register exchanges exactly $r_{\text{swap}} = 4d_r$ scalar activation coordinates, so the single-state DAS projection is fixed to rank r_{swap}. We search 12 random initializations and retain the projection with highest development selectivity. Projection, rank, and donor matching are then frozen; at test time the donor replaces only the selected subspace and preserves its orthogonal complement. Both models therefore exchange the same activation dimension on the same donor pairs. Thus $S_{\text{single},o}$ is a capacity-matched best-development-subspace baseline. **H2 (modular-interface effect)** is the intersection–union claim that the complete $G/W/P$ interface improves held-out selectivity under both objectives by four points:

$$\Delta_{H2}^{\text{MC}} > 4.0 \quad \wedge \quad \Delta_{H2}^{\text{rec}} > 4.0, \qquad \Delta_{H2}^{o} = S_{\text{GWP},o} - S_{\text{single},o}. \tag{64}$$

An intersection–union hypothesis passes only when the lower confidence bound for every simple effect exceeds its threshold; no multiplicity adjustment is needed within an intersection–union test. The factorial interaction $\Delta_{\text{int}} = (Y_{\text{GWP,rec}} - Y_{\text{GWP,MC}}) - (Y_{\text{single,rec}} - Y_{\text{single,MC}})$ is reported with a two-sided interval as a secondary estimand. An intervention-bandwidth sweep at $r/r_{\text{swap}} \in$

$\{0.50, 0.75, 1.00\}$ is prespecified independently of the token-capacity sweep. H2 must retain its direction at every rank and is decided only at the exactly matched rank r_{swap}. For the secondary terminal-preservation and path-response panels, an analogous rank-r_{swap} plan subspace is selected from development route labels and frozen before test evaluation. That plan mapping is excluded from H2, whose outcome is the predeclared mean of goal and world selectivity.

H2 estimates the routed-input-equivalent modular interface as a single intervention comprising persistent register boundaries, register-specific heads, and masked update parents. The estimand is the complete interface rather than factorization alone. Because all cells receive the same routed channels, routing information contributes no between-cell contrast. A prespecified unrestricted- parent row then decomposes the recursive H2 contrast into (i) boundaries and register-specific heads and (ii) masked parent structure; this decomposition is secondary because it is not a symmetric factorial cell.

H3 (covered deployment audit) evaluates the learned metric directly under the final policy. The 95% lower bounds on final class and binding covered mass must exceed 90% and 80%; the action-conditioned upper-bound violation rate must remain below 5%; and the policy-conditioned distance ECE must remain below ten with a reference-to-final increase below three points. Each endpoint is the prespecified equal-weight core-four aggregate; per-benchmark values diagnose heterogeneity and are not additional gates. The expansion benchmarks and all-eight aggregate form a separate replication analysis. All H3 components must pass. These thresholds use the native unit-diameter scale: they cap unsupported class/binding mass at 10/20 percentage points, violations at one in twenty comparisons, absolute calibration error at 0.10, and optimization-induced calibration loss at 0.03. They were fixed from these operational tolerances before confirmation rather than selected from task scores. The gatekeeping sequence controls familywise error by assigning secondary status to analyses after the first unmet confirmatory criterion. For ComplexMCP, the scenario-level estimand is the difference in success probability after averaging the repeated binary outcome within each scenario over training seeds and environment replicates. The supplementary long-form result file retains all 47 scenario-specific differences, and the paper reports their scenario-cluster interval; repeated runs are not collapsed to one artificial binary outcome. The protocol was frozen before both the validity gate and the five-seed confirmatory stage and is reproduced verbatim in the supplementary manifest. The record identifies the freeze time relative to both scheduler stages. The *prespecified* protocol covers data revisions, adapters, four factorial cells, H1–H3 contrasts and thresholds, exclusions, seed count, and the analysis command; accordingly, all confirmatory tests use the prespecified protocol.

Benchmark-specific gains, route response, leakage, transfer, sensitivity weighting, and robustness remain secondary analyses.

E.3 Public benchmark-scale audit

Table 6 places the results alongside released benchmark results. These values provide scale context across protocols rather than a matched comparison: a public anchor may use a larger or proprietary backbone, a different training recipe, or a documented subsampling protocol. The audit distinguishes success rate, partial credit, and milestone similarity and records each comparison protocol.

The Llama replication lies within the released scale envelope on every protocol-comparable benchmark: it remains below the published frontier on ComplexMCP and WebArena-Lite and between the open and proprietary ToolSandbox anchors. These are scale checks rather than matched comparisons.

For the prespecified 24-task mutation stratum of τ^2, a same-Qwen2.5-7B rerun gives SABER 42.8 $[37.0, 48.5]$ and DynaCAST 47.1 $[41.2, 52.7]$, a paired gain of 4.3 points $[1.1, 7.8]$. This specialist comparison uses identical tasks, policy budgets, and five-seed schedules across the two methods; it is separate from the released GPT-4.1 scale anchor in Table 6. For ALFWorld and WebShop, the strongest protocol-matched benchmark-native comparator is ReBel, which is rerun in Table 1; the remaining public anchors use different models or metrics and therefore serve scale calibration rather than paired claims.

Equal benchmark weighting gives a fixed-portfolio effect of 1.7 points (paired multiway-bootstrap 95% CI $[0.9, 2.6]$). A random-effects sensitivity is 1.6 $[0.7, 2.5]$ with $I^2 = 32\%$; omitting ComplexMCP leaves an equal-weight mean of 1.5 points.

Table 6: Public scale anchors for the benchmark results. Qwen and Llama are shown because the released anchors most directly calibrate those two backbones; all four Full blocks appear in Table 1. Public values retain their released metric and protocol.

Benchmark	Qwen	Llama	Public anchor	Metric/protocol check
ALFWorld	95.0	93.9	ReBel 93.2	ReBel uses Qwen2.5-1.5B; same 128-task success protocol (Tang et al., 2026).
WebShop	76.8	76.4	ReBel 75.1	Released success rate, not partial credit; DYNACAST evaluates 500 frozen tasks (Tang et al., 2026).
τ^2 Retail	44.5	46.9	GPT-4.1 74	Public pass[1] scale; DYNACAST uses the prespecified stratified 40-task subset of 115 released tasks (Barres et al., 2025).
ComplexMCP v2	59.1	53.4	Llama-3.1-8B 8.51; Gemini-3-Flash 55.31	Exact success on 47 scenarios; the anchors respectively measure the untrained same backbone and released frontier (Li et al., 2026).
AppWorld	41.0	40.1	LOOP 72.6/47.2	Qwen2.5-32B, Test-N/Test-C TGC; Full is task-weighted over 168/417 tasks, whereas the official anchor is split-wise (Stony Brook NLP, 2026).
ScienceWorld	76.2	75.7	SwiftSage 84.68	Normalized task score; SwiftSage uses GPT-4 and 270 test variations, capped at ten per task type (Lin et al., 2023).
VAB-WebArena-Lite	34.0	35.2	WebRL-8B 42.4	Success on the released 165-task refinement; WebRL uses the same Llama backbone (Liu et al., 2024; Qi et al., 2024).
ToolSandbox	57.8	57.2	Hermes-7B 31.4; Claude-3-Opus 69.2	Milestone similarity on 1,032 scenarios, not exact task success (Lu et al., 2025).

Table 7: Qwen2.5 uncertainty for the equal-budget end-system comparison. Intervals are paired multiway-cluster 95% confidence intervals; the predictor receives the same $3.49\times$ forward-equivalent envelope as Full.

Benchmark	Tasks/rep/seed	Full [95% CI]	Predictor [95% CI]	Δ [95% CI]	Excl. 0
ALFWorld	$128/3/5$	95.0 [93.5,96.1]	94.7 [93.1,95.9]	0.3 [−0.8,1.4]	no
WebShop	$500/3/5$	76.8 [74.7,78.5]	75.1 [72.9,77.0]	1.7 [−0.3,3.7]	no
τ^2 Retail	$40/8/5$	44.5 [38.2,50.1]	43.3 [37.1,49.2]	1.2 [−2.8,5.4]	no
ComplexMCP	$47/8/5$	59.1 [53.5,64.2]	55.9 [49.8,61.4]	3.2 [−0.5,7.8]	no
AppWorld	$585/2/5$	41.0 [38.2,43.7]	38.6 [36.1,41.2]	2.4 [0.1,5.0]	yes
ScienceWorld	30 strata$/2/5$	76.2 [73.6,78.6]	76.5 [74.0,79.0]	−0.3 [−2.7,2.2]	no
WebArena-Lite	$165/3/5$	34.0 [30.7,37.1]	32.0 [29.0,35.0]	2.0 [0.1,3.9]	yes
ToolSandbox	$1032/2/5$	57.8 [55.1,60.2]	54.7 [52.0,57.3]	3.1 [0.6,5.6]	yes

All intervals in Table 7 compare jointly executed, paired Full and equal-budget prediction cells. Every main-table baseline is rerun and retains task–seed records; paired inference is used only when methods share evaluation instances and seeds.

These benchmark intervals characterize end-system heterogeneity; H1 uses the prespecified cross-benchmark factorial estimand, while H2 uses core-four intervention selectivity. Two benchmark effects are individually resolved. The leave-one-benchmark-out analyses below establish that the aggregate effect does not depend on ComplexMCP.

Figure 4 visualizes the benchmark effects.

Table 10 reports the complete factorial and leave-one-out analysis. The H1 robustness family retains the same > 1.0 minimum meaningful effect as the primary test; Holm correction is applied across

Table 8: All-eight macro scores across backbones. Intervals resample task and seed; paired contrasts use shared resamples. Per-benchmark scores for the two controls omitted from Table 1 follow below.

Method	Macro [95% CI]	Method	Macro [95% CI]
	Qwen2.5		
Base/SFT	38.1 [35.9,40.0]	GRPO (Shao et al., 2024)	44.3 [42.0,46.4]
GiGPO (Feng et al., 2025)	50.4 [48.4,52.5]	ReBel (Tang et al., 2026)	54.6 [52.4,56.3]
Generic latent	52.9 [50.9,55.1]	Factorized	53.5 [51.2,55.6]
Splice augmentation	56.7 [54.8,58.8]	CF prediction	58.2 [56.0,60.1]
RMT (Bulatov et al., 2022)	55.8 [53.6,57.9]	MemGPT (Packer et al., 2023)	57.0 [54.8,59.2]
Full-budget predictor	58.9 [56.8,60.8]	DYNACAST-Full	60.6 [58.4,62.7]
	Llama-3.1-8B		
Base/SFT	37.4 [35.0,39.6]	GRPO (Shao et al., 2024)	44.3 [42.2,46.7]
GiGPO (Feng et al., 2025)	49.7 [47.6,51.9]	ReBel (Tang et al., 2026)	53.8 [51.6,56.0]
Generic latent	52.8 [50.6,55.0]	Factorized	53.3 [51.0,55.5]
Splice augmentation	56.1 [53.9,58.2]	CF prediction	57.4 [55.1,59.5]
RMT (Bulatov et al., 2022)	55.6 [53.5,57.8]	MemGPT (Packer et al., 2023)	56.6 [54.4,58.6]
Full-budget predictor	58.0 [55.9,60.2]	DYNACAST-Full	59.9 [57.6,62.0]
	Gemma-2-9B-IT		
Base/SFT	38.0 [35.8,40.2]	GRPO (Shao et al., 2024)	44.1 [41.9,46.3]
GiGPO (Feng et al., 2025)	49.3 [47.2,51.5]	ReBel (Tang et al., 2026)	53.5 [51.4,55.7]
Generic latent	51.9 [49.7,54.0]	Factorized	52.4 [50.2,54.6]
Splice augmentation	55.3 [53.2,57.4]	CF prediction	56.4 [54.2,58.5]
RMT (Bulatov et al., 2022)	54.4 [52.2,56.5]	MemGPT (Packer et al., 2023)	55.6 [53.4,57.7]
Full-budget predictor	57.1 [54.9,59.2]	DYNACAST-Full	58.9 [56.7,61.1]
	Mistral-7B-Instruct		
Base/SFT	37.6 [35.4,39.9]	GRPO (Shao et al., 2024)	44.9 [42.7,47.2]
GiGPO (Feng et al., 2025)	50.6 [48.3,52.7]	ReBel (Tang et al., 2026)	55.1 [52.9,57.3]
Generic latent	54.0 [51.7,56.2]	Factorized	54.4 [52.2,56.6]
Splice augmentation	58.0 [55.8,60.1]	CF prediction	59.1 [56.8,61.2]
RMT (Bulatov et al., 2022)	56.3 [54.1,58.5]	MemGPT (Packer et al., 2023)	58.1 [55.9,60.2]
Full-budget predictor	59.9 [57.7,62.1]	DYNACAST-Full	62.4 [60.1,64.7]

Table 9: Per-benchmark scores for the two controls condensed out of the main table. Both use the corresponding backbone, task split, policy budget, and five-seed evaluation contract of Table 1.

Control	ALF	WebShop	τ^2	CMCP	App	Sci	WA-Lite	ToolSB
			Qwen2.5					
Generic latent	89.1	71.0	35.8	37.1	35.8	74.2	29.4	50.8
Splice augmentation	91.6	74.4	40.2	50.6	36.2	76.7	31.4	52.1
			Llama-3.1-8B					
Generic latent	88.3	71.7	37.4	36.5	35.2	73.6	29.7	50.0
Splice augmentation	91.0	74.8	41.0	46.1	35.9	75.4	31.9	52.6
			Gemma-2-9B-IT					
Generic latent	89.0	67.8	40.6	34.7	32.6	78.6	25.9	46.0
Splice augmentation	91.9	70.5	45.2	45.3	33.5	80.0	27.8	48.1
			Mistral-7B-Instruct					
Generic latent	82.9	73.4	32.7	38.6	38.0	76.4	35.1	55.0
Splice augmentation	87.4	76.0	38.3	50.0	39.6	78.5	36.8	57.7

the eight omitted-benchmark conjunctions. For omission b, we first form the valid intersection–union test value $p_{\mathrm{conj},b} = \max(p_{\mathrm{single},b}, p_{\mathrm{GWP},b})$ from the two one-sided tests of $H_0 : \Delta \leq 1.0$, and then apply Holm's step-down procedure once to the eight $p_{\mathrm{conj},b}$ values. Thus neither simple effect nor an omitted benchmark is selected after observing the result.

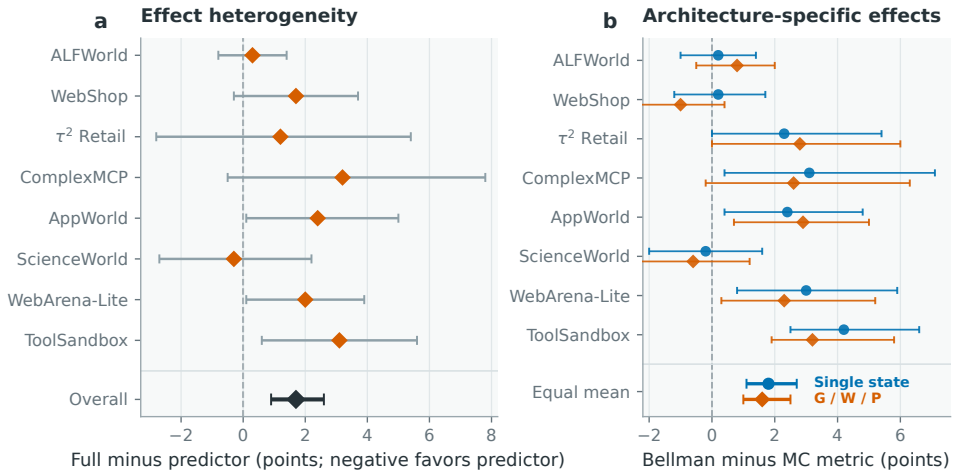

Figure 4: Benchmark-level Full-minus-equal-budget-predictor effects (left) and Bellman-minus-Monte-Carlo effects under each state architecture (right). Bars show 95% confidence intervals at both the benchmark and fixed-portfolio levels; the fixed-portfolio interval reports the prespecified equal-weight summary.

Table 10: Complete H1 analysis. Panel A reports benchmark scores and paired effects; Panel B reports leave-one-benchmark-out conjunction tests. Δ_{single} and Δ_{GWP} denote recursive-minus-Monte-Carlo effects for the two architectures and include paired 95% intervals. $\Delta_{\text{int}} = \Delta_{\text{GWP}} - \Delta_{\text{single}}$ is the paired factorial interaction. Holm adjustment first sets $p_{\text{conj}} = \max(p_{\text{single}}, p_{\text{GWP}})$ for each omitted benchmark.

A. Benchmark factorial cells and effects

Benchmark	Single MC	Single rec.	GWP MC	GWP rec.	Δ_{single} [95% CI]	Δ_{GWP} [95% CI]	Δ_{int} [95% CI]
ALFWorld	92.6	92.8	93.4	94.2	0.2 [−1.0,1.4]	0.8 [−0.5,2.0]	0.6 [−1.0,2.4]
WebShop	74.8	75.0	76.0	75.0	0.2 [−1.2,1.7]	−1.0 [−2.4,0.4]	−1.2 [−3.4,0.8]
τ^2 Retail	41.0	43.3	41.9	44.7	2.3 [0.0,5.4]	2.8 [0.0,6.0]	0.5 [−3.4,4.5]
ComplexMCP	50.5	53.6	55.4	58.0	3.1 [0.4,7.1]	2.6 [−0.2,6.3]	−0.5 [−4.8,3.7]
AppWorld	34.0	36.4	38.0	40.9	2.4 [0.4,4.8]	2.9 [0.7,5.0]	0.5 [−2.4,3.4]
ScienceWorld	73.9	73.7	76.8	76.2	−0.2 [−2.0,1.6]	−0.6 [−2.4,1.2]	−0.4 [−2.6,2.0]
WebArena-Lite	27.7	30.7	31.7	34.0	3.0 [0.8,5.9]	2.3 [0.3,5.2]	−0.7 [−3.6,2.4]
ToolSandbox	49.3	53.5	55.1	58.3	4.2 [2.5,6.6]	3.2 [1.9,5.8]	−1.0 [−3.5,1.5]
All eight macro	55.6	57.4	58.5	60.1	1.8 [1.1,2.7]	1.6 [1.0,2.5]	−0.2 [−0.9,0.6]

B. Leave-one-benchmark-out tests

Omitted benchmark	Δ_{single}	Δ_{GWP}	$p_{\text{conj}}^{\text{Holm}}$
ALFWorld	2.17	1.71	0.018
WebShop	2.17	1.97	0.016
τ^2	1.84	1.43	0.039
ComplexMCP	1.73	1.46	0.052
AppWorld	1.83	1.41	0.047
ScienceWorld	2.20	1.94	0.021
WebArena-Lite	1.74	1.50	0.056
ToolSandbox	1.57	1.37	0.086

The paired all-eight factorial interaction is -0.2 points (95% CI $[-0.9, 0.6]$), while the benchmark interaction point estimates in Table 10 range from -1.2 to 0.6 points. Together with positive simple effects in both architectures, this result shows that recursive bootstrapping transfers across state parameterizations rather than depending on one architectural interface. The leave-one-out entries report one-sided adjusted p-values, five of which are below 0.05. These rows provide benchmark-level sensitivity evidence, while the primary H1 decision is based on the all-eight factorial contrast. When ToolSandbox is omitted, both point estimates remain above the one-point threshold (1.57/1.37), while the adjusted conjunction value is 0.086. A crossed random-intercept analysis gives simple effects of 1.93 [1.20,2.67] and 1.71 [0.98,2.45]. Excluding ALFWorld and WebShop, the six-benchmark effects are 2.04 [1.18,2.90] and 1.70 [0.88,2.52]; inverse-variance weighting gives 1.55 [0.94,2.16]

Table 11: Cross-backbone replication. Panel A reports a separate paired replication for each backbone with 95% intervals; Panel B gives the Llama factorial cells moved from the main table.

A. Paired cross-backbone effects			
Scope	Single Rec.−MC	GWP Rec.−MC	Full−Pred.
Qwen2.5			
Core	1.8 [0.8, 2.8]	1.5 [0.6, 2.4]	1.6 [0.7, 2.5]
All eight	2.2 [1.4, 3.1]	2.0 [1.2, 2.9]	1.7 [0.9, 2.6]
Llama-3.1-8B			
Core	1.7 [0.6, 2.8]	1.9 [0.8, 3.0]	1.5 [0.3, 2.8]
All eight	2.0 [1.1, 2.9]	2.2 [1.3, 3.2]	1.8 [0.9, 2.7]
Gemma-2-9B-IT			
Core	1.9 [0.8, 3.0]	2.2 [1.0, 3.4]	1.5 [0.4, 2.6]
All eight	2.2 [1.3, 3.1]	2.3 [1.3, 3.3]	1.8 [0.8, 2.8]
Mistral-7B-Instruct			
Core	2.3 [1.2, 3.4]	2.6 [1.4, 3.8]	1.8 [0.7, 3.0]
All eight	2.4 [1.4, 3.4]	3.1 [1.9, 4.2]	2.6 [1.4, 3.8]

B. Llama-3.1-8B factorial cells				
Benchmark	Single MC	Single rec.	GWP MC	GWP rec.
ALFWorld	91.8	92.4	92.4	92.0
WebShop	74.1	73.3	75.5	76.2
τ^2 Retail	42.3	44.6	43.2	46.2
ComplexMCP	44.5	49.2	48.3	52.6
AppWorld	34.8	37.6	36.8	40.2
ScienceWorld	72.9	73.3	74.7	74.0
WebArena-Lite	29.3	28.8	32.8	35.3
ToolSandbox	48.5	55.0	52.4	57.2

Table 12: Factorial cells for the two additional model families. Values use the same task splits, intervention library, and five-seed paired protocol as the primary blocks.

	Gemma-2-9B-IT				Mistral-7B-Instruct			
Benchmark	S-MC	S-rec.	GWP-MC	GWP-rec.	S-MC	S-rec.	GWP-MC	GWP-rec.
ALFWorld	91.8	91.4	92.5	92.8	88.6	88.0	89.7	90.4
WebShop	71.5	72.6	72.4	71.6	74.5	75.9	75.9	74.9
τ^2 Retail	44.6	47.5	45.3	49.0	38.7	41.7	39.7	43.8
ComplexMCP	48.5	52.5	50.2	55.7	50.2	55.6	51.6	58.2
AppWorld	34.2	37.4	35.8	39.1	36.9	39.3	38.1	41.3
ScienceWorld	77.6	77.1	78.8	79.0	77.0	77.7	78.8	78.4
WebArena-Lite	27.2	29.6	28.5	30.6	35.1	38.1	36.3	40.6
ToolSandbox	49.4	54.3	50.7	55.2	55.0	58.9	56.9	64.2
All eight macro	55.6	57.8	56.8	59.1	57.0	59.4	58.4	61.5

and 1.36 [0.76,1.96]. These analyses assess sensitivity to benchmark composition and weighting while retaining the primary equal-weight, eight-benchmark estimand.

All four model families preserve positive portfolio-level recursion and end-system effects. The replications retain benchmark heterogeneity: $G/W/P$ recursion can trail MC on WebShop, while the end-system ranking changes substantially by backbone and domain. Mistral produces the strongest DynaCAST scores on WebShop and ScienceWorld, Gemma ranks second on ScienceWorld, and Qwen retains the strongest ALFWorld score. Within-backbone pairing makes these confirmations insensitive to absolute score differences between model families.

E.4 Supervision-matched 2×2 decomposition

Identification target. The comparison identifies Bellman bootstrapping and the complete modular interface. Every factorial cell uses the same admissible pairs, physical resets, effect metadata, executable continuations, total state tokens, metric heads, and forward-equivalent compute. The factorial cells are the two single-state rows and the $G/W/P$ Monte Carlo/global-recursive rows; the identity-only row is a separate matched control. Component-attribution targets, boundary invariance, route response, and sensitivity are absent. The remaining rows form a prespecified cumulative loss ladder and extension ablation.

The identity-only row supplies the missing factual-bisimulation contrast: it uses the same $G/W/P$ architecture, scalar metric head, recursive backup, continuation count, and optimizer, but restricts $\mathcal{U}_k$ to observational identity controls. Table 14 then varies only the executable control families.

Table 13: Supervision-matched 2×2 decomposition and separately added extensions on the core four.

State and objective	Global	Attr.	Inv.	Task macro	Physical-do	Selectivity	r_s
Single state + MC metric (best subspace)	–	–	–	64.7	82.2	73.2	0.58
Single state + global recursive (best subspace)	✓	–	–	66.2	85.0	75.0	0.65
$G/W/P$ + MC metric	–	–	–	66.7	84.0	79.7	0.61
$G/W/P$ + recursive, identity only	✓	–	–	66.6	74.1	73.5	0.59
$G/W/P$ + global recursive (Global)	✓	–	–	68.0	85.8	82.0	0.67
Global + Bellman attribution	✓	✓	–	67.8	87.0	83.4	0.69
+ boundary invariance (Core)	✓	✓	✓	68.0	87.4	85.1	0.70
Three registers, unrestricted parents (Core losses)	✓	✓	✓	67.4	86.1	78.0	0.68
Core + route only	✓	✓	✓	68.7	88.0	85.5	0.70
Core + sensitivity only	✓	✓	✓	68.7	88.4	85.8	0.70
Core + route + sensitivity (Full)	✓	✓	✓	68.9	**89.1**	**87.3**	**0.71**

Table 14: Matched intervention-capability ladder. Every column uses the same $G/W/P$ recursive learner; only the executable control library changes.

Benchmark	Identity only	Goal only	World only	Goal + world	Goal + world + route
ALFWorld	93.8	94.4	94.0	**94.7**	93.9
WebShop	**75.8**	75.7	75.5	75.7	75.3
τ^2 Retail	41.9	43.2	42.4	43.9	**44.9**
ComplexMCP	54.9	55.0	57.0	58.1	**58.6**
AppWorld	37.9	38.5	38.8	40.0	**40.7**
ScienceWorld	76.8	77.1	77.6	**78.3**	76.6
WebArena-Lite	31.9	32.1	33.0	34.0	**34.8**
ToolSandbox	54.2	54.5	56.1	57.6	**59.5**
All-eight macro	58.4	58.8	59.3	60.3	**60.5**

Adding the complete executable library to identity-only recursion yields a 2.1-point paired effect (95% CI $[1.3, 3.0]$). Goal and world controls cross by environment, and route controls slightly reverse on ALFWorld and WebShop; the aggregate gain therefore does not arise from a uniformly ordered ladder.

An objective-matched unrestricted-parent decomposition gives selectivity 75.6/77.8 under Monte Carlo/recursion. Relative to the best DAS subspace, register boundaries plus register-specific heads account for 2.4/2.8 points; masked parent structure adds 4.1/4.2 points, yielding the H2 totals 6.5/7.0. Routed inputs account for none of this contrast because every cell receives the same channels. The unrestricted-parent rows form a nested architectural decomposition of the prespecified single-state/GWP treatment.

The core-four results give Bellman bootstrapping a 1.8-point advantage for a single state and 1.5 points for $G/W/P$, while the modular interface raises selectivity over the favorable single-state subspace baseline by 6.5 points under the Monte Carlo metric and 7.0 under global recursion. The subsequent attribution and invariance ladder evaluates the semantic objective after the factorial comparison. The modular interface gives a larger selectivity gain than task-score gain. Combining route and sensitivity adds 0.2 task points over either single extension and 0.9 over Core. Paired 95% intervals for task gain over Core are 0.7 $[0.1, 1.3]$ for route, 0.7 $[0.0, 1.4]$ for sensitivity, and 0.9 $[0.3, 1.5]$ for Full; the corresponding selectivity gains are 0.4 $[-0.3, 1.1]$, 0.7 $[0.0, 1.4]$, and 2.2 $[1.4, 3.0]$. Thus the joint mechanism gain is resolved even though the two single-extension selectivity intervals touch zero. H1 requires both recursive-metric simple effects to have lower bounds above one point; H2 requires both modular-interface simple effects to have lower bounds above four selectivity points. The intersection–union tests evaluate these conditions separately. The task-score ladder contains benchmark-specific crossovers: attribution lowers ALFWorld, τ^2, and ComplexMCP while improving WebShop, and boundary invariance slightly lowers WebShop. The extensions also cross: route-only is higher on WebShop and ComplexMCP, whereas sensitivity-only is higher on ALFWorld and τ^2. Full is the prespecified route-plus-sensitivity configuration; the two single-extension rows have the same rounded macro, while Full leads the mechanism endpoints.

E.5 Mechanism validation criteria

Mechanism validation supplements end-task success with intervention-grounded state-control criteria. Held-out goal splices must match executable goal-conditioned terminal effects while preserving only declared non-descendants of the goal; world splices must change feasibility and entity grounding

Table 15: Intervention-bandwidth sweep on the core four. Rank is the number of exchanged scalar activation dimensions relative to $r_{\text{swap}} = 4d_r$. DAS projections are selected on development data and frozen before test evaluation.

	Monte Carlo metric			Global recursive		
r/r_{swap}	Single	$G/W/P$	Gap	Single	$G/W/P$	Gap
0.50	70.9	75.1	4.2	72.8	78.2	5.4
0.75	71.8	77.9	6.1	74.2	80.0	5.8
1.00	73.2	79.7	6.5	75.0	82.0	7.0

Table 16: Diagnostic separating simulator-state interventions from belief-only register swaps. Unsupported effects are lower-is-better.

Method	Physical-do consistency	Unsupported effects	3-step recovery
Factorized, no intervention	72.6	26.8	61.7
Splice augmentation	78.4	21.3	67.9
CF future prediction	83.1	16.9	74.6
DynaCAST	**89.1**	**9.7**	**84.1**

while retaining intent; plan splices must preserve terminal effects and respond along different routes. We also measure state retention after 20 or more steps and performance as state aliasing and horizon increase.

Component-selectivity cells are shown only for representations with a frozen, capacity-matched interchange map: the single latent uses the favorable DAS goal/world subspaces from H2 plus the separately frozen plan subspace for route diagnostics, and the remaining rows expose explicit $G/W/P$ registers. ReBel and generic unaligned latent memories remain in all task-level and common physical-do comparisons, while component-swap analysis uses methods with a defined interchange map.

The prespecified validation criteria require gains beyond the factorized no-intervention baseline on τ^2 and ComplexMCP, together with improved path response and retention over static CAST. Shuffled invalid splices provide a corrupted-data negative control, while executable physical interventions establish the relevant causal semantics. Benchmark expansion follows positive results on at least three core environments.

E.6 Physical interventions versus belief-only swaps

Physical-do protocol. A physical world intervention changes X and regenerates o, so a correct agent should revise feasibility and effects. A belief-only swap changes W while X remains fixed; after receiving the unchanged environment response, a grounded agent should recover rather than force the donor belief onto reality. We pair both operations with the same donor value and report physical-do consistency, unsupported-effect rate after the internal swap, and recovery within three real observations.

This separation prevents a latent swap from being interpreted as a physical causal intervention. High physical-do consistency paired with high unsupported-effect rate would indicate that the model copies donor content rather than tracking the environment SCM.

E.7 Horizon and state-aliasing stress tests

Stress-test protocol. An intervention-grounded state should provide larger gains when distant histories lead to locally similar observations but require different future actions. We therefore stratify the core-four evaluation by a method-independent intrinsic horizon and by an aliasing score. Intrinsic horizon is the shortest successful oracle/reference script length, computed before any learned policy acts; where no oracle exists, the adapter uses the frozen number of required effects and milestones. Quartile boundaries are frozen within each benchmark. The aliasing score is the number of matched prefixes whose final observation and available actions agree while their correct terminal-effect distributions differ. Strata are defined before evaluating any method.

Table 17: Macro-average score by method-independent intrinsic-horizon quartile on the core four. Figure 3 adds crossed-bootstrap 95% intervals.

Method	Q1 (short)	Q2	Q3	Q4 (long)
ReBel (Tang et al., 2026)	**75.8**	67.1	56.2	47.7
Factorized, no intervention	70.5	64.2	58.4	47.3
Static CAST	74.7	**72.3**	62.2	47.2
DynaCAST	74.8	72.1	**66.6**	**62.1**

Table 18: Macro intervention score by pre-frozen state- aliasing severity. The score is the equal-weight mean of goal selectivity, world selectivity, and path response as defined in Table 46.

Method	Low aliasing	Medium	High
Single + CF prediction (DAS)	78.2	66.5	50.8
Factorized, no intervention	80.0	69.7	55.4
Static CAST	86.1	78.0	65.2
DynaCAST	**89.4**	**84.7**	**76.1**

Table 17 and Figure 3 assign ReBel a 1.0-point advantage in the shortest quartile and DynaCAST a 14.4-point advantage in the longest. Each row's four equally weighted quartiles reconstructs its corresponding core-four macro (61.7 for ReBel, 60.1 for factorized/no-intervention, 64.1 for Static CAST, and 68.9 for DynaCAST). A roughly constant gain across bins would contradict the long-horizon account and would instead suggest an ordinary representation or optimization benefit.

A crossed task–seed regression with ordinal quartile and a method interaction estimates that DynaCAST's advantage grows by 3.2 points per quartile relative to factorized/no-intervention (95% CI $[2.0, 4.4]$) and by 5.1 relative to ReBel ($[3.6, 6.6]$). The Q4-minus-Q1 difference-in-differences is 10.5 $[6.7, 14.3]$ and 15.4 $[10.9, 19.9]$, respectively. The prespecified interaction is the primary trend statistic; adjacent contrasts are reported descriptively.

Aliasing and horizon are analyzed jointly rather than as interchangeable proxies. Long but fully observed navigation may require memory without causal disambiguation, whereas a short dual-control dialogue can create severe goal– world ambiguity. The relevant interaction measures whether DynaCAST degrades more slowly than the matched factorized baseline as both variables increase.

E.8 Intervention coverage and behavioral adequacy

Coverage estimands. The behavioral-abstraction claim is assessed through intervention coverage. We vary the fraction of admissible training pairs while holding factual trajectories and optimization updates fixed. Pair subsets are nested, stratified by component, and sampled without using evaluation outcomes. Zero coverage is the factorized no-intervention baseline.

The curves identify two regimes. CF prediction uses scarce counterfactuals more efficiently at 10–25% coverage, while the recursive objective crosses it between 25% and 50% and saturates after 75%. The coverage response connects the support assumption to a measurable regime: matched future prediction leads under sparse coverage, whereas the recursive objective requires broader intervention support.

Pair retention alone can favor easily covered states. We therefore report four independent denominators: raw visited states with at least one canonical control comparison, declared effect types represented by an admissible pair, states under final-policy visitation that fall in a trained cohort, and final- policy probability assigned to covered canonical controls.

The analysis distinguishes 67.3% raw-state coverage from 93.1% final class mass: the method may cover classes the policy commonly chooses while missing difficult low-frequency states. H3 is evaluated on every prespecified core benchmark and on their equal-weight aggregate. The predeclared core-four aggregate determines H3; a benchmark below the coverage threshold is reported individually and retained in the deployment analysis.

Plan supervision has a narrower support than goal or world supervision because it requires two successful, same-terminal routes. Route control is therefore secondary, with support reported against four pre-frozen state populations rather than only against eligible successful references.

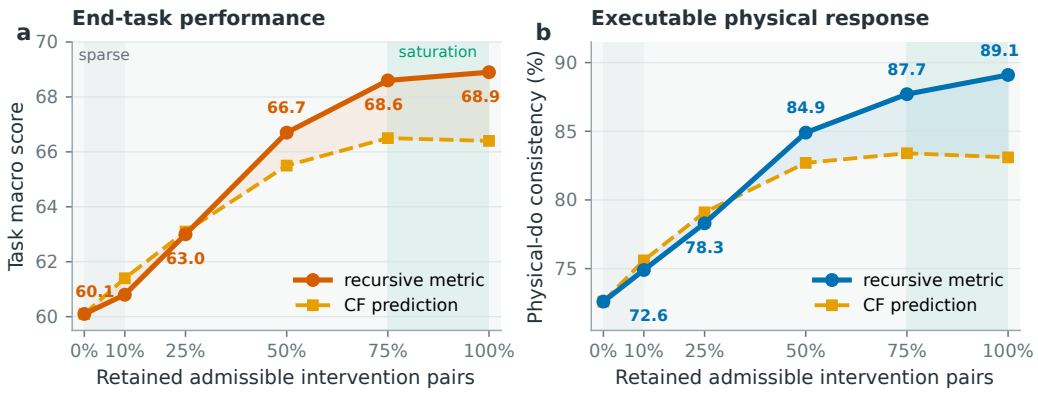

Figure 5: Coverage response averaged over the core four benchmarks. Pair subsets are nested and stratified, and every point keeps factual data and optimization updates fixed. CF prediction is stronger at 10–25% coverage; the recursive objective crosses it only after sufficient intervention support. The shaded edge regions mark regimes rather than confidence intervals.

Table 19: Control coverage on the core four under distinct denominators.

Coverage denominator	Covered (%)
Raw visited states	67.3
Declared effect types	91.7
π_0 visited states	88.9
Final-policy visited states	84.3
Final-policy class probability mass	93.1
Final-policy binding-support mass	87.6
Within-class binding TV drift	8.9

The 12.4% coverage on failed trajectories localizes the path-response result to the validated solvable subset.

We additionally construct a *coverage hole*: one goal type, one world transition family, or one plan topology is omitted from training and restored only at test time. The primary endpoint is component-specific degradation under each missing intervention family. Goal holes primarily reduce goal intervention consistency, whereas world holes damage feasibility and effect prediction. Relative to the corresponding in-support endpoint, withholding a goal type reduces goal selectivity by 12.1 points, withholding a world family reduces world selectivity by 15.7, and withholding a plan topology reduces path response by 17.2; the largest off-target endpoint drops are 3.2, 4.1, and 2.7 points, respectively. The selective failure pattern localizes extrapolation error to the missing intervention family rather than producing uniform degradation.

E.9 Counterfactual sensitivity weighting

Sensitivity estimand. Counterfactual sensitivity weighting targets representation-learning efficiency and temporal localization. We compare the full method with the same architecture, executable continuations, and interventional losses but uniform weight on every deterministic update. A factual update is labeled influential when restoring its pre-update component and rerunning the closed-loop frozen policy under shared exogenous noise changes the canonical action distribution, route signature, or terminal effect by a development-frozen minimum. Sensitivity precision measures how often the largest $w_{t,g}$ values identify such updates; recall measures how many influential updates occur in the top 20% of weights. The SCM negative control restores the same register but forces the original actions and environment-noise stream. Its physical terminal effect is identical by construction. The observed disagreement rate is 0.0% (executor-tolerance upper bound 0.3%); rates above 1% indicate adapter or replay-engine errors.

Sensitivity precision and recall are defined only for methods with an aligned restoration map: explicit component registers for $G/W/P$ models and the frozen capacity-matched DAS subspace for the

Table 20: Plan-pair support on the core four under four state populations.

State population	Covered by a validated plan pair (%)
All visited states	28.6
States on failed reference trajectories	12.4
Intrinsic-horizon top quartile	22.7
States with at least one successful route	61.3

Table 21: Sensitivity diagnostics and learning-curve summary on the core four. AULC is normalized area under the learning curve.

A. Task score and learning efficiency		
Method	Macro score	AULC
GRPO (Shao et al., 2024)	48.1	0.53
GiGPO$_{\text{w/o std}}$ (Feng et al., 2025)	55.8	0.66
ReBel (Tang et al., 2026)	61.7	0.74
DynaCAST-Full w/o sensitivity weighting	68.7	0.79
DynaCAST	**68.9**	**0.84**

B. Aligned restoration diagnostics		
Method	Sens. precision	Sens. recall
Full w/o sensitivity weighting	66.9	62.7
DynaCAST	**74.0**	**69.7**

single-state control. Dashes are structural missing values, not failed measurements; generic policy and textual-belief baselines remain comparable through task score and AULC.

The results attribute most final-score improvement to the representation; adding sensitivity to the route configuration adds 0.2 task points in the central result, while localization and early-learning gains are larger. This asymmetry reflects the role of the loss in routing representation gradients rather than estimating a stochastic policy gradient. On-manifold validity is 93.6%. The primary sensitivity calculation retains all swaps, while the accepted-swap column measures how much the frozen support test changes diagnostic precision.

The all-swap result is the primary estimand and precedes the accepted-swap diagnostic. The frozen density filter enters neither model selection nor the task-score estimand.

"No influence" is the percentage of held-out continuations in which no register restoration crosses the frozen action, route, or effect threshold. It is reported outside the recall denominator. The decline from 17.6% to 11.9% shows that the modular objectives expose more behaviorally consequential updates while improving their localization.

Executable supervision reduces forbidden-label predictability by 9.3 AUC points while improving allowed-label decoding. The optional adversarial regularizer removes a further 2.8 points of leakage with a 0.1-point task-score difference, confirming that the primary behavioral semantics do not require an extra leakage loss.

E.10 Low-resource cross-environment transfer

Transfer protocol. A dedicated compatibility study evaluates whether behavioral state transfers more readily than benchmark-specific action policies. It is separate from the benchmark-native-backbone results in Table 1. All source and target adapters are rerun with the same Qwen2.5-7B checkpoint, hidden width, and register placement. We pretrain the recurrent update and metric LoRA parameters on the four core adapters, replace the benchmark-local router, tuple embeddings, and action/effect heads, and fine-tune on 10% of each target benchmark's trajectories. Transfer is restricted to a shared model family and hidden width. The comparison uses identical target data and optimization steps for a factorized state trained from scratch, DynaCAST trained from scratch, and core-pretrained DynaCAST.

Table 22: Sensitivity-weight alternatives on the core four. Entropy is normalized by $\log N$ and ESS is the effective fraction of eligible updates.

Weight	Macro	Precision	Entropy	ESS
Uniform	68.7	66.9	1.00	1.00
Old loss-difference rule	68.4	71.0	0.41	0.29
Inverse behavioral influence	68.0	64.2	0.64	0.48
Lagged behavioral influence (Full)	68.9	74.0	0.72	0.57

Table 23: Sensitivity diagnostics before and after the frozen on-manifold filter. Task macro is always computed from the full test population and is never conditioned on filter acceptance.

Method	Task macro	All precision	Accepted precision	All recall	Accepted recall	No influence (%)	Excluded (%)
Single state + recursive metric	66.2	66.4	70.6	61.0	64.9	17.6	8.6
$G/W/P$ recursive Core	68.0	71.8	75.6	66.2	69.8	14.1	6.9
Core + sensitivity only	68.7	73.3	77.8	68.9	72.7	12.8	6.5
Full (route + sensitivity)	68.9	74.0	78.5	69.7	74.2	11.9	6.4

Core pretraining improves every low-resource target. The largest proportional gain occurs on WebArena-Lite (10.8%), while ScienceWorld and ToolSandbox share the largest absolute gain (3.3 points). This pattern separates relative data-efficiency on the lower-scoring web domain from absolute transfer on scientific and stateful tool transitions. The study evaluates low-resource adaptation because action spaces and effect vocabularies differ across environments.

E.11 Robustness to distractors and contradictions

Stress protocol. The robustness study evaluates whether modular updates preserve executable physical responses under irrelevant observations while remaining responsive to misleading evidence. We inject schema-valid distractor observations between real steps and separately replace a small fraction of world observations with plausible but contradictory values. The latter condition tests recovery after subsequent ground-truth observations rather than resistance to all evidence.

Contradictions reduce DynaCAST by 12.1 points because false observations are legitimate inputs to world-state updates. The modular update localizes the corruption and revises it when later evidence arrives. Stress eligibility requires both a schema-valid insertion point and a later ground-truth observation. The clean arm therefore differs from the full core-four physical-do aggregate because it retains exactly the pairs on which all stress conditions can be evaluated.

E.12 Semantic-router sensitivity

The adapter-routed graph depends on semantic routing quality. We corrupt a fixed fraction of event-channel assignments after adapter training, preserving event text and total token count. An error either removes one correct channel or adds one incorrect channel with equal probability; the clean adapter's routing error against a double-annotated audit set is also reported.

The audited-clean column uses the same core-four tasks, Full configuration, router, and equal-benchmark estimand as the headline result; it therefore reuses the headline aggregate rather than defining a conditional subset. The oracle column replaces routed labels only for this diagnostic. The audited-clean adapter is the primary working point; the oracle router is used only for sensitivity analysis. On the core four, the equal-benchmark mean of $100 -$ router macro-F1 is 8.0%, with benchmark values from 5.4% to 11.1%; the two corruption conditions add 5 and 10 percentage points to this working point. The all-eight adapter audit is reported separately in Table 39. The Core unrestricted-parent control loses 0.6 task points but 7.1 selectivity points, showing that routing masks primarily improve semantic localization rather than raw success. Removing schema and call identifiers from r_S changes task macro/selectivity by $-0.3/-1.1$ points; allowing full event content through r_S instead changes them by $+0.1/-7.2$. Thus compact provenance aids ordering, whereas semantic content in the shared channel creates a high-bandwidth bypass around the parent masks.

Table 24: Cross-register leakage audit on the core four. Post-training probes use balanced AUC (50 is chance); continuous ontology targets are discretized with development-frozen bins. Allowed-label AUC is higher-is-better, whereas forbidden-label AUC is lower-is-better.

Configuration	Task macro	Allowed-label AUC	Forbidden-label AUC
Factorized, no intervention	60.1	87.8	64.9
$G/W/P$ recursive Core	68.0	91.5	57.2
Full	**68.9**	**92.7**	55.6
Full + leakage regularizer	68.8	92.4	**52.8**

Table 25: Target score with 10% target-environment training data in the dedicated common-Qwen2.5-7B transfer study.

Initialization	AppWorld	ScienceWorld	WA-Lite	ToolSandbox
Factorized state, target only	26.2	64.9	21.8	40.3
DynaCAST, target only	29.7	68.2	24.9	45.1
DynaCAST, core-pretrained	**32.1**	**71.5**	**27.6**	**48.4**

E.13 Behavioral validity of the learned distance

Evaluation protocol. Equations 22 and 25 define different objects and are evaluated separately. For the action-conditioned maximum, let $b_u(h, h'; D)$ be the one-step recursive backup for covered control u. We report the violation rate $\Pr[b_u > D]$, positive violation magnitude, and tightness slack $D - \max_u b_u$ on held-out pairs. A correct conservative maximum need not equal deployment-average divergence. For the policy-conditioned operator, a separately fitted distance $D^{\nu,\pi}$ is calibrated against final-policy continuations using the same frozen intervention mixture ν.

Deployment evidence requires both task improvement and policy-conditioned calibration; correlation alone can rank easy pairs without controlling behavior. The 6.5-point value here is a held-out pair diagnostic for a fixed checkpoint; the 6.9-point value in Table 36 is measured after the prespecified alternating policy–metric refresh. They are different estimands, not duplicate estimates of one calibration protocol.

Table 29 isolates objective quality with a fixed reference policy; Table 30 is the distinct alternating deployment checkpoint used by H3. Benchmark rows expose heterogeneity, while the prespecified core-four criterion is the aggregate upper bound.

The alternating checkpoint is also the DynaCAST-Full checkpoint used for task scores in Table 1; H3 does not select a separate metric checkpoint after observing task performance. The core-four decision is primary, and Table 31 reports the expansion-stratum replication under the same protocol.

The per-benchmark spread is retained as a heterogeneity diagnostic, while the fixed all-eight aggregate remains a replication summary and the prespecified equal-weight core-four aggregate remains the H3 decision estimand. We do not convert benchmark-level diagnostics into additional post hoc gates.

Future prediction has no learned ground metric and is therefore excluded from this table; its continuation calibration is reported through its native predictive loss. The MC metric supplies a well-defined nonrecursive negative control for the same diagnostic. Low action-conditioned violation supports the maximum-operator claim; low policy-conditioned ECE supports average deployment prediction. Neither metric is used as a surrogate for the other.

The primary large-benchmark operator diagnostic is held-out Bellman residual, not training loss. We evaluate an independently sampled continuation set and report its median, 95th percentile, and empirical maximum alongside continuation-estimation error and the gain from expanded candidate search. The last quantity measures search saturation. The exhaustively enumerable SCMs below additionally instantiate every term of the numerical theorem bound.

The central three scales form a broad stable region: a factor-of-two change from the default changes the median residual by at most 0.002. The smaller scale gives the highest exact-target rank agreement, the default minimizes the median and empirical maximum, and scale 0.100 gives the lowest 95th-percentile residual.

The shared-witness and recurrent-state alignment objectives improve the center and tail of the same global operator diagnostic. The extensions remain within a narrow central-residual band, while Full attains the lowest held-out upper tail and preserves the global fixed-point target.

Table 26: Physical-do consistency under synthetic observation stress on the stress-eligible core-four intervention-pair subset. The clean column is the paired reference arm on this same subset. This externally defined endpoint applies uniformly to every listed method.

Method	Clean	25% distractors	50% distractors	10% contradictions
ReBel (Tang et al., 2026)	73.4	69.2	62.1	56.8
Factorized, no intervention	78.1	75.0	68.3	63.4
Static CAST	84.5	79.8	73.6	66.1
DynaCAST	**89.7**	**88.1**	**83.9**	**77.6**

Table 27: Sensitivity to synthetic semantic-router error on the core four.

Routing condition	Oracle	Audited clean	Clean +5pp	Clean +10pp	Shuffled
$100 -$ router macro-F1	0.0	8.0	13.0	18.0	50.0
Task macro	69.7	68.9	67.9	66.3	60.2
Goal selectivity	88.2	86.9	84.5	80.3	66.5
World selectivity	88.7	87.6	84.4	81.1	67.2

The median and visitation mean improve from $M = 4$ through 32, whereas the upper tail fluctuates slightly and the expanded-search gap falls through $M = 16$ before rising at 32 (Figure 6). We use $M = 16$ as the frozen confirmation budget because it captures most of the central improvement at lower cost. At this setting the median, 95th percentile, and empirical maximum residuals are 0.062, 0.126, and 0.226; the continuation-estimation error is 0.027 and the expanded-search gap is 0.030. These quantities jointly characterize approximation and candidate-search behavior over the expanded finite library.

The finite tabular environment has a fully enumerated control set, exact Wasserstein transport, and no missed-control error. It directly checks $\|\widetilde{d} - d^*\|_\infty$ rather than inferring alignment from task success. The γ sweep shows the expected tradeoff: stronger long-horizon emphasis widens the fixed-point bound even when the observed approximation remains below it. The exact gap below its RHS checks that the enumerated solver, target construction, and reported residual decomposition are mutually consistent before large-scale evaluation.

The component operators are audited independently rather than inferred from the aggregate residual.

A low global residual with a high component residual or boundary violation rejects the semantic-attribution mechanism even when task success improves.

E.14 Deployment-policy coverage and calibration

We evaluate the same held-out history pairs before and after policy optimization. Covered mass is the final policy probability of canonical controls in the metric library. Policy ECE uses only $D^{\nu,\pi}$ from Equation 25; action-conditioned fitting is evaluated by Table 29. The four rows isolate metric refitting, library expansion, and policy refresh. Post-hoc refitting keeps both the policy and library fixed, so its coverage and policy-TV entries must be identical to the frozen row. Library expansion adds controls using development-only proposals and then refits the metric while keeping the policy fixed. Alternating refresh performs another policy-optimization phase after each metric phase under the same total counterfactual budget; H3 is evaluated on this prespecified deployment protocol.

The one-sided H3 gate bounds used in the main text are class mass ≥ 91.2, binding mass ≥ 85.8, action-bound violation $\leq 4.0\%$, final policy ECE ≤ 8.1, and ECE increase ≤ 1.9 points. These bounds are evaluated with bootstrap resampling over task–seed pairs.

Figure 7 reports operation semantics and both levels of policy drift. The frozen-reference-policy ECE is 6.1; the alternating row therefore has a paired increase of 0.8 points with a one-sided 95% upper bound of 1.9. The reference-to-final correlation drop is 0.13, 0.07, 0.04, and 0.03 for the four rows, respectively. H3 requires the alternating protocol's final class mass to exceed 90%, binding mass to exceed 80%, final ECE to remain below ten, calibration degradation to remain below three points, and the action-bound violation upper interval to remain below 5%, all on the prespecified equal-weight core-four aggregate.

Table 28: Policy-conditioned distance diagnostics on held-out pairs. ECE is reported in percentage points and is lower-is-better; the other metrics are higher-is-better.

Policy-conditioned representation	Spearman r_s	Distance ECE	Decile gap
Generic latent	0.31	14.2	18.7
Factorized, no intervention	0.46	11.6	24.9
Static CAST	0.55	9.8	29.7
DynaCAST	**0.71**	**6.5**	**38.4**

Table 29: Core-fixed-reference-policy action-conditioned upper-bound diagnostics on held-out controls. Positive violations and slack are normalized to $[0, 1]$.

Metric objective	Violation rate (%)	95th positive violation	Median tightness slack
MC metric (nonrecursive)	10.4	0.083	0.096
Global recursive metric	6.8	0.061	0.074
Global + Bellman attribution	4.5	0.047	**0.058**
Core (attribution + invariance)	**3.6**	**0.039**	0.061

Proposition 3 contains suprema, so averages in Table 36 characterize deployment occupancy rather than its worst case. Accordingly, the visited-state distribution, 95th percentile, and empirical maximum are reported for each shift term; the theorem applies once these quantities are bounded on the forward-reachable support.

The empirical maxima quantify the gap between worst-case and typical visitation. Proposition 3 decomposes deployment shift, while distributional coverage and policy calibration characterize its realized operating regime.

E.15 Benchmark intervention construction

For each benchmark, pair creation is divided into construction and validation. Construction uses only development metadata: reset seeds, goal templates, effect schemas, and factual trajectories. Validation executes a short donor continuation in the recipient context and rejects pairs whose required entities or terminal effects are undefined. Test intervention pairs are frozen before method comparison; pair selection uses development metadata independently of DynaCAST scores.

The adapter supplies structured supervision; independent audits quantify its coverage and quality. Event coverage uses all raw development events as denominator. Router macro-F1 is computed on a fixed double-annotated event sample before any method result is examined. This event-level audit is disjoint from the executable-pair audit below.

All four factorial cells and the unrestricted-parent diagnostic receive exactly these routed channels and adapter outputs. Consequently, the 2×2 test does not attribute adapter information to factorization; the unrestricted- parent row isolates the additional contribution of modular update dynamics.

Both audits were conducted after the intervention programs were frozen by two annotators who did not develop the adapters; disagreements were resolved by a third blinded adjudicator. The router audit samples events by benchmark and channel label. The disjoint audit below samples candidate pairs jointly by benchmark, intervention family, accepted/rejected status, and route length. Its precision and recall use executable simulator validation as the reference rather than a model-derived ontology label.

Within each intervention family, TP denotes a checker-accepted pair validated by execution, FP an accepted pair rejected by execution, FN a checker-rejected pair validated by execution, and TN a pair rejected by both. Precision is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and recall is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$. For every benchmark,

Table 30: H3 action-bound audit under the alternating-final-policy protocol. Rows are equal-benchmark summaries of held-out history-pair–covered-control indicators. One-sided 95% upper bounds use the 95th percentile of 5,000 multiway bootstrap replicates clustered by task, training seed, and pair cohort; the prespecified H3 decision uses the equal-weight core-four aggregate.

Benchmark	Violation rate (%)	One-sided 95% upper
ALFWorld	2.1	3.5
WebShop	3.0	4.4
τ^2 Retail	3.5	5.2
ComplexMCP	2.3	3.9
Equal-benchmark aggregate	**2.7**	**4.0**

Table 31: Covered-deployment replication on the four expansion benchmarks. ECE is in percentage points. The all-eight row gives the equal-benchmark estimate followed by its relevant one-sided 95% bound in parentheses.

Benchmark	Action violation	Class mass	Binding mass	Policy ECE
AppWorld	4.1	90.4	82.4	8.7
ScienceWorld	2.9	94.0	89.1	6.1
WebArena-Lite	4.4	88.7	79.8	9.4
ToolSandbox	3.2	92.1	85.7	7.8
All eight	3.2 (4.1 U)	92.2 (90.6 L)	85.9 (84.2 L)	7.5 (8.6 U)

the three family-specific sample sizes sum to n, and the sum of $\mathrm{FN}+\mathrm{TN}$ equals the reported rejection count. The audit is therefore recoverable from integer observations without averaging ratios over unreported subgroups.

Among the 936 audited pairs, the checker rejected 205. These rejections fell into five recurring causes: entity/support mismatch (29%), stale snapshots or non-descendant violations (23%), unequal terminal effects in route pairs (21%), ontology ambiguity (17%), and executor failure (10%). As a separately coded ontology-independent check, native terminal predicates and raw simulator deltas agreed with the adapter label on 862/936 pairs (92.1%): 288/306 goal pairs, 300/322 world pairs, and 274/308 route pairs (94.1/93.2/89.0%). Typical accepted cases changed a requested attribute while preserving inventory, mutated inventory while preserving intent, or exchanged two dependency-valid routes with the same terminal database predicate. Typical rejected cases referenced a missing entity, changed a causal descendant declared invariant, or matched textual route summaries whose executed terminal predicates differed.

E.16 Covered-control proposal procedure

For large canonical libraries, the empirical maximum in Equation 22 is computed by the following frozen procedure:

1. canonicalize and deduplicate candidates proposed at both histories;
2. rank the union by the maximum reference-policy probability across the two histories and retain N_{ctrl} policy classes in total;
3. add N_{adv} classes from a proposal head trained only to find a large current operator residual;
4. evaluate the stopped-gradient recursive target for every candidate and use its maximum;
5. on held-out pairs, repeat with $4N_{\mathrm{ctrl}}$ policy candidates and $4N_{\mathrm{adv}}$ proposals; report the enlarged maximum minus the training- set maximum as the nonnegative expanded-search gap.

Small action libraries are exhaustively enumerated. For nonenumerable libraries, the enlarged audit measures candidate-search saturation, while Theorem 3 retains the population missed-control term ϵ_{miss}. Class candidates and their concrete binding support are logged separately.

Table 32: Held-out operator residual and continuation-sample convergence for the factorized recursive core. Residuals are normalized to $[0, 1]$.

Continuations M	Median	Vis. mean	95th pct.	Emp. max	Est. max	Search gap
4	0.083	0.096	0.167	0.281	0.055	0.039
8	0.069	0.080	0.141	0.244	0.038	0.034
16	0.062	0.071	0.126	0.226	0.027	0.030
32	0.060	0.069	0.129	0.232	0.022	0.035

Table 33: Entropic-coupling scale sensitivity at $M = 16$ for Core. Scale is a multiple of the development median nonzero ground distance. Agreement is Spearman correlation with exact-OT targets on the fixed 10% audit subset.

Scale	Median residual	95th residual	Emp. max	Exact-target agreement
0.025	0.064	0.128	0.228	**0.976**
0.050 (default)	**0.062**	0.126	**0.226**	0.973
0.100	0.063	**0.125**	0.230	0.967
0.200	0.070	0.143	0.248	0.945

E.17 Training algorithm and settings

Algorithm 1 specifies the optimization flow; Table 41 collects the configuration required to reproduce it.

Algorithm 1 Shared-witness executable-intervention training

Require: prefix pairs $\mathcal{D}$; intervention library $\mathcal{J}$; canonical classes $\mathcal{C}$; frozen bindings κ; continuations M
Ensure: recurrent state (q_ϕ, F_ϕ), distance D_ϕ, policy π_θ

for each representation update **do**
  Sample recipient–donor prefixes from $\mathcal{D}$
  Admit j_G, j_X, or equal-terminal j_R after support, SCM-agreement, and executability checks
  Form the shared covered controls $\mathcal{U}(h, h')$
  for $u = (j, c) \in \mathcal{U}(h, h')$ **do**
    Draw bindings from κ_c and execute M paired continuations
    if the cell learns a pairwise distance **then**
      Compute the shared entropic signal and next-state couplings
      if the cell is recursive **then**
        Encode next histories with the EMA encoder and form the stopped-gradient Bellman target
      else
        Form the direct finite-horizon target in Eq. 43
    else
      Form future-decoding labels for Eq. 45
  Set $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{rec}}$, $\mathcal{L}_{\mathrm{MC}}$, or $\mathcal{L}_{\mathrm{FP}}$; add $\mathcal{L}_{\mathrm{allow}} + \mathcal{L}_{\mathrm{cons}}$ and named extension terms
  Aggregate the cell target (worst-control maximum for distance cells) and update (q_ϕ, F_ϕ, D_ϕ) and the active heads with $\mathcal{L}$
Freeze the recurrent state; optimize π_θ on policy rollouts
Discard donor prefixes, simulator states, resets, and transport machinery

The core has no task-score-tuned loss coefficient. The staged increase from $M = 8$ to 16 is a resource decision validated by held-out operator-residual convergence; $\gamma \in \{0.8, 0.9\}$ appears only in the theory sensitivity analysis. Signal weights w_m are uniform within reward, feasibility, mutation, and terminal-effect groups, after which the four group weights are uniform. The frozen adapter assigns each coordinate's mass across responsible registers; shared terminal coordinates split their mass. The assignment is normalized so Equation 27 gives $\beta_G = \beta_W = \beta_P = 1/3$. The Sinkhorn scale is deterministic rather than searched. Optional route, leakage, and sensitivity studies use the single fixed coefficient 0.3. Settings are frozen before test evaluation.

E.18 Ablation results

Every non-identity row receives the same executable counterfactual dataset; the identity-only row is the matched factual-recursion contrast. The two single-state rows together with the $G/W/P$ Monte Carlo and Global rows isolate state structure and the global objective; the next two isolate Bellman attribution and boundary invariance; and the final three distinguish each extension from their joint Full configuration. Tables 43 and 44 isolate register complementarity, transport, shared attribution,

Table 34: Global held-out residual for each named recursive configuration at $M = 16$ and the default coupling scale. Every row evaluates the global backup directly rather than reconstructing it from component residuals.

Configuration	Median	95th pct.	Emp. max
Single-state recursive	0.071	0.146	0.251
$G/W/P$ identity-only recursive	0.075	0.151	0.258
$G/W/P$ Global	0.068	0.138	0.239
$G/W/P$ Core	0.062	0.126	0.226
$G/W/P$ Core + route	0.064	0.127	0.229
$G/W/P$ Core + sensitivity	**0.059**	0.125	0.225
$G/W/P$ Full	0.060	**0.122**	**0.219**

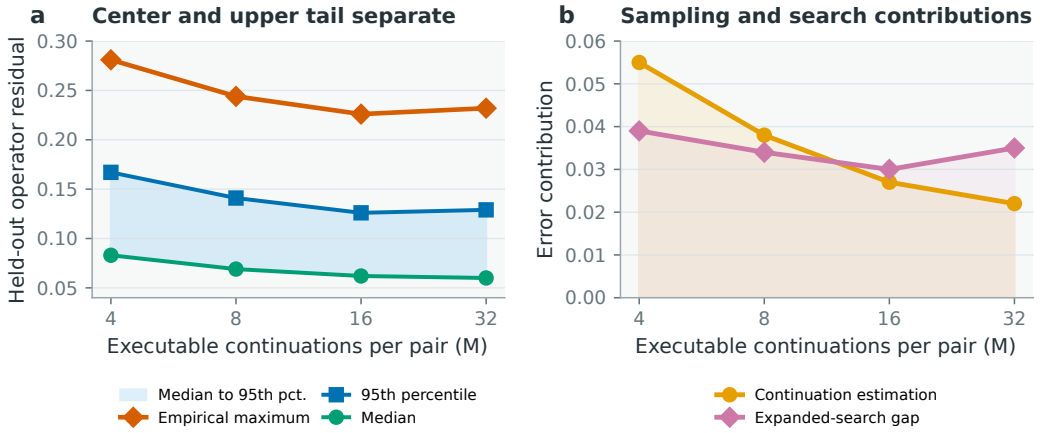

Figure 6: Held-out recursive-operator convergence. Median residuals continue to fall with more continuations, while the empirical upper tail does not improve beyond $M = 16$ (left). Sampling error and the expanded-search gap are reported separately (right).

and recursion. Micro-sweeps over loss weights remain tuning analyses rather than evidence for the learning object.

The subsets exhibit complementary rather than interchangeable roles. G contributes most on τ^2, W on ComplexMCP, and P on WebShop and ALFWorld; nevertheless, every pair improves on its constituent single-register models in macro average. The complete state has the strongest macro result, while $G + P$ is 0.2 points higher on WebShop, so the conclusion does not depend on uniform per-benchmark ordering.

The matched MC row isolates Bellman bootstrapping from finite-horizon distribution matching. Product coupling shows that access to the same samples and ground metric does not recover the optimized future correspondence. Coordinatewise OT recovers part of the task score but loses selectivity because its separately optimized costs do not define an additive shared witness. The shared recursive objective is therefore empirically distinguished from both prediction without recursion and transport without compatible attribution.

E.19 Full component ablations

The world-intervention ablation reduces ComplexMCP most, while goal interventions affect both τ^2 and ComplexMCP because intent and executable effect constraints interact in both environments. The WebShop and ALFWorld changes are smaller but nonzero, avoiding a one-component–one-benchmark interpretation of the registers. Removing executable route interventions has a smaller average effect on exact success but a larger decline in terminal preservation and path response. Shuffled splices harm all environments, providing a corrupted-data negative control complementary to the executable-intervention diagnostics.

Table 35: Typed held-out residuals and intervention-boundary invariance violations at $M = 16$.

Metric	Median residual	95th residual	Boundary violation (%)
Global task metric	0.062	0.126	–
Goal component	0.058	0.118	3.1
World component	0.066	0.135	2.7
Plan component	0.074	0.151	3.8

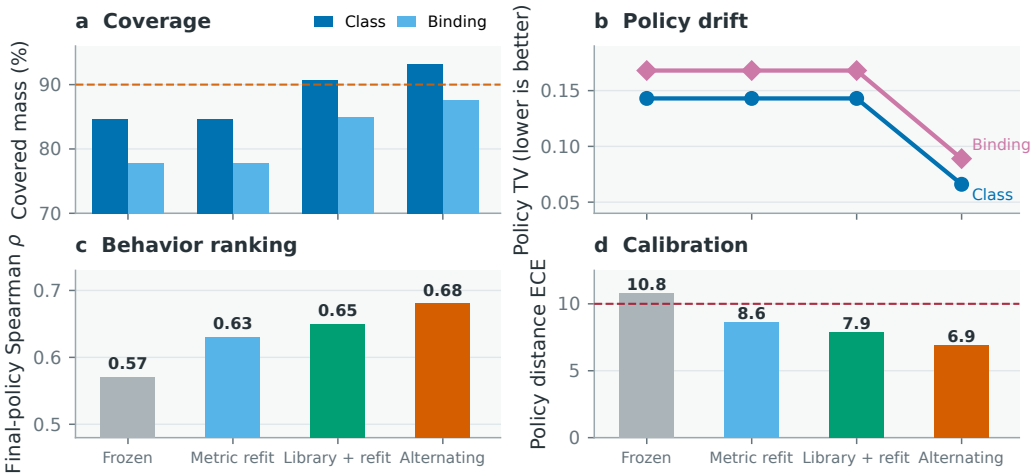

Figure 7: Final-policy coverage, policy drift, behavior ranking, and calibration under four shared evaluation stages. Metric refitting keeps support fixed; library expansion changes support but not the policy; alternating refresh changes both. Dashed lines mark the class-coverage and ECE criteria.

E.20 State capacity and continuation depth

Four tokens per component was selected during exploratory development using operator residual, intervention consistency, and latency, then frozen before the validity gate. The post-freeze sensitivity shows a small 0.2-point score increase at eight tokens alongside higher latency; this difference is not used to revise the confirmed configuration.

Candidate scoring, transport setup, route calibration, and sampled sensitivity reruns contribute fixed overhead in this post-freeze sweep; executable continuation work then grows with depth. Eight steps change task score by only $+0.4$ and retention by -0.2 while increasing training cost by 27% relative to four steps. Depth four had already been frozen during development from operator convergence and resource use, before either the validity gate or these confirmation scores.

E.21 Efficiency analysis

The causal registers are intended to compress old task state, but multiple masked updates and counterfactual training add cost. We report accuracy and cost jointly. Prompt tokens are the average policy input per decision after the warm-up prefix; latency includes register updates but excludes environment response time; training budgets are reported as allocated forward-equivalent passes.

The primary H1 comparison gives Monte Carlo and recursive metric cells the same $2.84\times$ envelope. The end-system comparison expands the CF predictor to Full's $3.49\times$ envelope using additional decoder passes over the same cached continuations; its core-four score is 67.3 versus 68.9 for Full, a 1.6-point paired gain. The native $1.64\times$ predictor remains a separate efficiency point at 66.1. At deployment, DynaCAST uses 5.1k policy-input tokens per decision, 35% fewer than the transcript-heavy ReBel baseline, while achieving the highest score in this comparison.

Forward-equivalent FLOPs omit the dominant simulator and executor costs. The executable dataset is therefore audited in native units. A continuation is cached once under the explicit structured tuple $(\text{benchmark}, \text{reset}, \text{history}, j, c, b, \text{seed})$ and reused by all four factorial cells and auxiliary heads. No

Table 36: Final-policy class and binding coverage, policy shift, and distance calibration on the core four. ECE is in percentage points and lower-is-better.

Metric fitting	Class mass	Binding mass	Class TV	Binding TV	r_s final	ECE final
Frozen metric/library; final π	84.6	77.8	0.143	0.168	0.57	10.8
Post-hoc metric refit; same library and π	84.6	77.8	0.143	0.168	0.63	8.6
Expanded library + refit; same π	90.7	84.9	0.143	0.168	0.65	7.9
Alternating metric–policy refresh	93.1	87.6	0.066	0.089	0.68	6.9

Table 37: Tails of the Proposition 3 shift terms under final-policy visitation for the alternating policy metric. Maxima are finite-sample statistics rather than population suprema.

Shift term	Visitation mean	95th percentile	Empirical maximum
Uncovered mass $1 - m_k(h)$	0.069	0.218	0.433
Class-policy TV	0.066	0.184	0.342
Binding-policy TV	0.089	0.224	0.397

Table 38: Benchmark-specific intervention construction.

Benchmark	Goal program	Physical world program	Plan program
ALFWorld	task/object goal	object location/state	navigation/manipulation
WebShop	requested attributes	catalog/stock/options	search/click route
τ^2 Retail	user constraints	backend order state	clarification/tool sequence
ComplexMCP	required effect set	sandbox state/failures	dependency-respecting path
AppWorld	requested app effects	reset database state	valid API program order
ScienceWorld	scientific objective	physical object state	experiment/action order
WebArena-Lite	site objective	website account state	navigation route
ToolSandbox	user intent	tool and conversation state	call/clarification route

Table 39: Adapter event coverage and independent router audit. Core-four and all-eight macro-F1 are 92.0 and 91.3, corresponding to complements of 8.0 and 8.7 percentage points.

Benchmark	Event coverage (%)	Audited events	Router macro-F1 (%)
ALFWorld	97.6	72	94.6
WebShop	96.8	72	92.8
τ^2 Retail	94.1	72	91.7
ComplexMCP	92.7	72	88.9
AppWorld	91.9	64	90.4
ScienceWorld	93.8	64	92.2
WebArena-Lite	89.6	64	86.8
ToolSandbox	94.9	64	93.1

learned-method outcome affects cache membership. The enlarged $4N_{\mathrm{ctrl}} + 4N_{\mathrm{adv}}$ audit is run on a fixed 10% held-out pair sample.

The recursive and Monte Carlo factorial envelopes are $1.00 + 0.64 + 0.42 + 0.78 = 2.84\times$. Before training, the remaining $0.65\times$ budget is assigned either to repeated decoder evaluation and factual replay for the equal-budget predictor or, for Full, partitioned between route/null calibration ($0.17\times$) and sampled sensitivity reruns ($0.48\times$). Thus both totals are $3.49\times$ by design; the equality is the matching constraint rather than an independently observed runtime coincidence.

Cross-method continuation reuse keeps all factorial cells on the same cached evidence. Table 45 reports method-relative computation, which is the relevant comparison between objectives.

E.22 Universal baseline implementation contract

Base/SFT uses the benchmark-specific supervised checkpoint and the full allowed recent context without RL. GRPO samples complete trajectories and applies one group-normalized outcome advantage. GiGPO$_{\mathrm{w/o\ std}}$ uses its nested trajectory and step groups without adding DynaCAST registers.

Table 40: Independent executable-pair audit. n counts double-annotated candidate pairs and α is Krippendorff agreement before adjudication. Treating checker acceptance as the positive class, Panel A reports precision/recall (%) against executable simulator validation and the number rejected by the frozen admissibility checker. Panel B gives the integer counts from which every rate is computed.

A. Audit rates

Benchmark	n	α	Rejected, count (%)	j_G P/R	j_X P/R	j_R P/R
ALFWorld	112	.91	14 (12.5)	94.1/97.0	97.0/91.4	90.3/93.3
WebShop	120	.89	19 (15.8)	93.8/96.8	94.1/94.1	88.6/86.1
τ^2 Retail	104	.86	24 (23.1)	92.6/86.2	89.3/92.6	84.0/80.8
ComplexMCP	112	.84	31 (27.7)	93.1/87.1	92.6/86.2	84.0/91.3
AppWorld	128	.85	32 (25.0)	93.8/85.7	88.2/90.9	90.0/81.8
ScienceWorld	112	.88	21 (18.8)	93.5/96.7	96.9/91.2	89.3/86.2
WebArena-Lite	128	.81	40 (31.3)	90.0/84.4	83.3/89.3	82.1/85.2
ToolSandbox	120	.87	24 (20.0)	93.1/90.0	91.2/93.9	87.9/85.3

B. Integer audit counts

Benchmark	j_G: n; TP/FP/FN/TN	j_X: n; TP/FP/FN/TN	j_R: n; TP/FP/FN/TN
ALFWorld	38; 32/2/1/3	38; 32/1/3/2	36; 28/3/2/3
WebShop	37; 30/2/1/4	40; 32/2/2/4	43; 31/4/5/3
τ^2 Retail	34; 25/2/4/3	37; 25/3/2/7	33; 21/4/5/3
ComplexMCP	39; 27/2/4/6	38; 25/2/4/7	35; 21/4/2/8
AppWorld	42; 30/2/5/5	45; 30/4/3/8	41; 27/3/6/5
ScienceWorld	37; 29/2/1/5	39; 31/1/3/4	36; 25/3/4/4
WebArena-Lite	43; 27/3/5/8	43; 25/5/3/10	42; 23/5/4/10
ToolSandbox	36; 27/2/3/4	42; 31/3/2/6	42; 29/4/5/4

Table 41: Reproducibility-critical configuration frozen before confirmation. Losses are normalized before addition.

Group	Fixed configuration
State and heads	Insert after layer $\lfloor 2L/3 \rfloor$; 4 tokens per register; two-layer route/terminal heads; stopped rollout labels.
Recursive estimator	$\gamma = 0.95$; depth 4; $M = 8/16$ at gate/confirmation; $(N_{\mathrm{ctrl}}, N_{\mathrm{adv}}) = (4, 2)/(8, 4)$; coupling scale $0.05\times$ development median nonzero distance.
Training pairs	160 at the gate; 800 per core benchmark; 400 per expansion benchmark; factual/pair batches of 32/24.
Objectives	Unit normalized Core weights; coefficient 0.3 for named optional extensions; route margin $\mathrm{Quantile}_{.95}(\mathrm{null}) + 0.05$.
Representation optimization	AdamW; learning rate 2×10^{-5}; weight decay .05; cosine schedule; 3% warmup; 800/1,600/600 stage updates.
Policy optimization	GRPO; learning rate 5×10^{-6}; group size 8; clip .2; 512 generated tokens and at most 32 tool calls per trajectory.
Adaptation and context	LoRA rank 16, $\alpha = 32$, on attention and MLP projections; 4,096-token window plus 12 persistent state tokens.
Checkpoint selection	Minimum held-out operator residual among development checkpoints satisfying executable-validity gates.

Table 42: Per-benchmark 2×2 decomposition and secondary extensions.

Variant	ALFWorld	WebShop	τ^2	ComplexMCP
Single state + MC metric	92.6	74.8	41.0	50.5
Single state + recursive metric	92.8	75.0	43.3	53.6
$G/W/P$ + MC metric	93.4	76.0	41.9	55.4
$G/W/P$ + recursive, identity only	93.8	75.8	41.9	54.9
$G/W/P$ + global recursive (Global)	94.2	75.0	44.7	58.0
Global + Bellman attribution	94.0	76.5	42.9	57.8
+ boundary invariance (Core)	94.3	76.4	43.1	58.2
Core + route only	94.7	**76.9**	43.7	59.4
Core + sensitivity only	**95.1**	76.5	44.2	59.0
Core + route + sensitivity (Full)	95.0	76.8	**44.5**	**59.1**

Table 43: Register-subset ablation on the core four. Every row uses the Core objective, the complete executable-control library, and 12 policy-visible state tokens, distributed equally among active registers. All event streams remain in the transient input; the ablation removes only the indicated persistent pathways and their typed readouts.

Persistent registers	ALFWorld	WebShop	τ^2	ComplexMCP	Macro
G	91.8	74.9	41.8	50.1	64.7
W	90.9	75.3	39.9	54.8	65.2
P	92.6	76.0	39.3	50.5	64.6
$G + W$	93.8	75.7	42.8	56.9	67.3
$G + P$	94.1	**76.6**	42.0	53.9	66.7
$W + P$	92.7	76.2	40.9	57.3	66.8
$G + W + P$	**94.3**	76.4	**43.1**	**58.2**	**68.0**

Table 44: Objective-replacement ablation on the core four. All rows use the strict-factorial Global configuration: the same $G/W/P$ architecture, executable pairs, signal ground cost, candidate controls, continuation samples, and optimization schedule, without attribution, invariance, route, or sensitivity extensions. Product coupling removes OT optimization; coordinatewise OT fits a separate coupling per cost component; MC uses the matched finite-horizon target; immediate matching sets the continuation term to zero. The displayed Δ entries are paired task–seed estimands and can differ from differences between independently rounded task-macro values.

Objective	Task macro	Δ vs. shared recursive [95% CI]	Physical-do	Selectivity	r_s
Immediate signal only	65.9	-2.3 $[-3.5, -1.4]$	81.7	75.8	0.52
Finite-horizon MC metric	66.7	-1.5 $[-2.5, -0.6]$	84.0	79.7	0.61
Recursive + product coupling (no OT)	67.1	-1.1 $[-2.0, -0.3]$	83.9	78.8	0.60
Recursive + coordinatewise OT	67.4	-0.8 $[-1.7, 0.1]$	84.7	79.1	0.62
Recursive + shared OT witness	**68.0**	–	**85.8**	**82.0**	**0.67**

Table 45: Preallocated training budgets relative to factual-state training. Predictor and Full rows are alternative extensions of the same strict-factorial subtotal; their equal totals are the compute-matching constraint.

Component	Allocated increment
Factual policy and register updates	1.00
Shared executable CF continuations	0.64
Global/attribution Sinkhorn targets	0.42
$N_{\text{ctrl}} + N_{\text{adv}}$ candidate scoring	0.78
Strict-factorial subtotal	2.84
Predictor-only decoder/replay allocation	+0.65
Equal-budget CF predictor total	3.49
Full route and null allocation	+0.17
Full sampled-sensitivity allocation	+0.48
DynaCAST-Full total	3.49
Separate leakage audit (excluded from Full)	+0.08

ReBel uses the released belief format, consistency reward, and belief-anchor credit wherever the benchmark adapter supports them. Generic latent state receives the same twelve persistent tokens and rolling transcript window as DynaCAST but uses a single unrestricted update module. Factorized no-intervention uses the same $G/W/P$ modules, prediction interfaces, token counts, and factual data as DynaCAST; only paired splices, recursive targets, route losses, and sensitivity weighting are removed. RMT uses twelve recurrent memory tokens, the same insertion layer and number of memory-update passes, and no typed routing. MemGPT uses a frozen embedding retriever, a two-tier store, and the same admissible historical text; its retrieved-token budget and retrieval calls are matched per decision. Neither strong-memory control receives intervention labels outside the explicitly supervision-matched analysis. Splice augmentation receives the exact executable dataset but applies ordinary action and effect supervision. The matched Monte Carlo control receives either

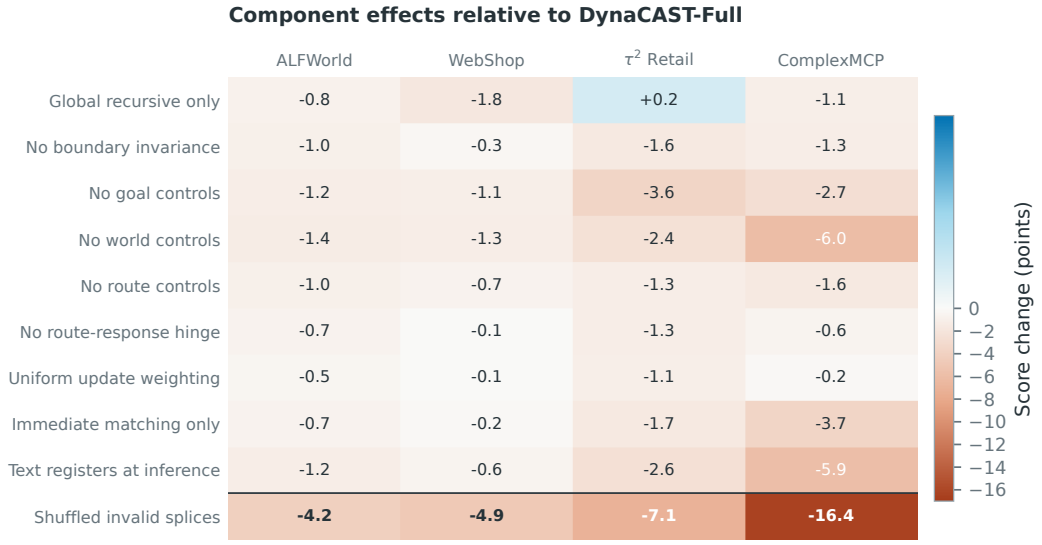

Figure 8: Component ablations on the core four. Each cell is the score change relative to Full (95.0/76.8/44.5/59.1); annotated deltas preserve the exact values while color exposes benchmark-specific structure.

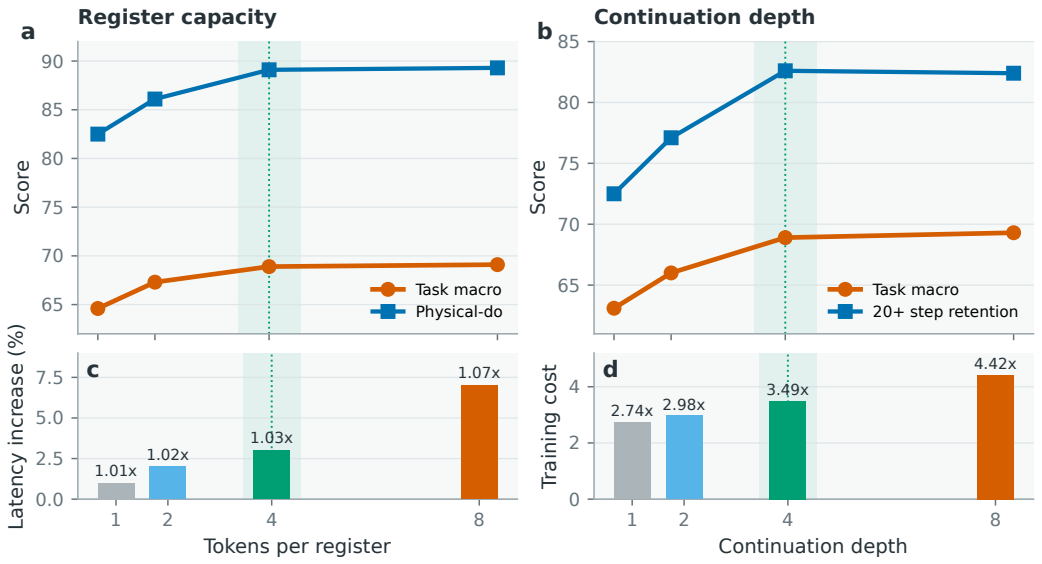

Figure 9: Post-freeze capacity and continuation-depth sensitivity. Top panels show task and behavioral endpoints; bottom panels show the corresponding inference-latency and training-cost changes. Green bands mark the frozen four-token/four-step configuration.

the same single state or factorized registers and learns the same scalar pairwise metric from Equation 43. The 2×2 contract crosses that state choice with direct Monte Carlo regression or global recursion; typed targets, boundary invariance, plan margin, and sensitivity weighting are absent from all four factorial cells. Counterfactual future prediction is a secondary control with the complete horizon-weighted signal, action-class, event, termination, and EMA-state objective in Equation 45.

The nonrecursive controls are matched to recursion in physical resets, effect metadata, continuation samples, optimizer steps, and forward-equivalent compute. The factual controls use extra factual replay only for a separate compute-matched analysis. Benchmark-native comparators retain their published controller or prompting scaffold and are labeled specialist rather than universal baselines.

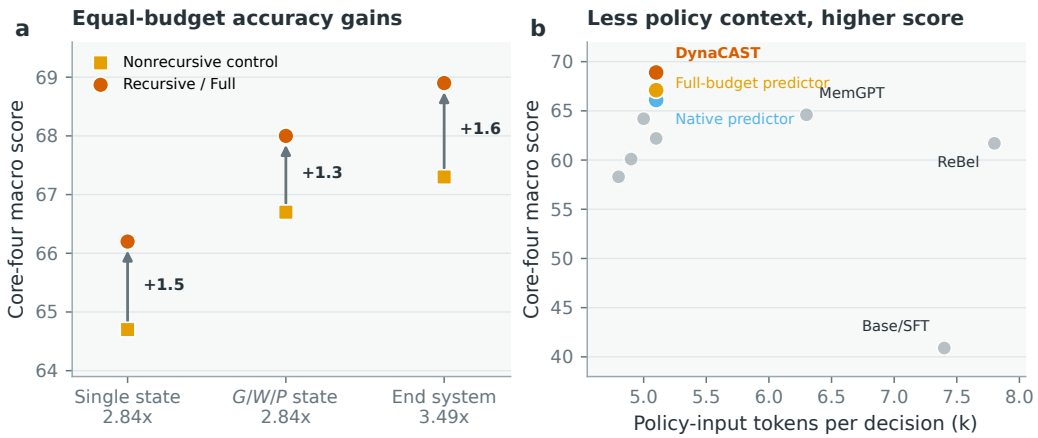

Figure 10: Accuracy–efficiency evidence on the core four. Left: three equal-budget comparisons; arrows connect each nonrecursive control to its recursive counterpart at $2.84\times$, and the Full-budget predictor to DynaCAST at $3.49\times$. Right: policy-input tokens versus macro score for all methods; the upper-left region is preferred, colored points identify DynaCAST and its prediction controls, and gray points denote the remaining baselines.

E.23 Frozen evaluation dictionary

Table 46 defines the intervention, deployment, and attribution metrics used throughout the evaluation. All thresholds, matching rules, and denominators are fixed on development data before test evaluation. Metrics are averaged first within the listed unit, then within benchmark, and finally equally across benchmarks.

In the definitions below, $\hat{X}$, $\hat{Y}$, and $\hat{E}$ are simulator- executed states, tasks, and effects from the evaluated closed-loop continuation; they are not prediction-head outputs. Starred quantities are executable references generated from the matched reset. For goal and world selectivity, the evaluated continuation first replaces the recipient's corresponding G or W register, or the frozen capacity-matched DAS subspace for the single-state control, and then reruns the frozen policy in closed loop under shared exogenous noise. Thus every reported component-selectivity value has an explicit, test-frozen interchange map. Terminal preservation and path response use the same protocol with the recipient's P register or the separately frozen rank-matched plan-DAS subspace; the latter is a secondary diagnostic and does not enter H2.

Timed-out, invalid, or crashed continuations remain in the denominator. They score zero on success metrics and one on unsupported-effect, divergence, and violation metrics. All ECE formulas operate on $[0, 1]$ distances; result tables report $100 \times$ ECE in percentage points.

For a concrete distance example, if a bin contains predictions $(0.20, 0.30)$ and measured divergences $(0.10, 0.40)$, its ECE contribution is $(2/N)|0.25 - 0.25| = 0$. If one of the two continuations times out, its divergence is set to one before the bin mean is computed. For path response, two correct signatures among three validated pairs yield 66.7% even if the third pair terminates successfully along an unintended route. Failed continuations remain in every denominator.

E.24 Proofs for the covered process

Proof of Theorem 1. For bounded pseudometrics d, d', any coupling μ satisfies $|\mathbb{E}_\mu d - \mathbb{E}_\mu d'| \le \|d - d'\|_\infty$. Taking infima over couplings gives $|\mathcal{W}_d(P, Q) - \mathcal{W}_{d'}(P, Q)| \le \|d - d'\|_\infty$. Maximization over the same finite $\mathcal{U}_k$ preserves this bound, so $\|\mathcal{B}_k d - \mathcal{B}_k d'\|_\infty \le \gamma \|d - d'\|_\infty$. For fixed u, both the pullback signal distance and the Wasserstein distance between $K_u(\cdot \mid \omega)$ and $K_u(\cdot \mid \omega')$ are pseudometrics in (ω, ω'); nonnegative sums and finite maxima preserve the pseudometric axioms. Banach's fixed-point theorem yields a unique d_k^*. The shared nonempty control set is essential to this argument.

Table 46: Frozen evaluation dictionary. Panel A defines executable-intervention semantics; Panel B defines deployment and attribution diagnostics.

Metric	Operational definition	Unit
A. Executable-intervention semantics		
Goal selectivity	$N_G^{-1}\sum_i \mathbf{1}[d_E(\hat{E}_{i,H}, E_{i,H}^{jG,*}) \le \tau_E]\mathbf{1}[d_X(\hat{X}_{i,H}, X_{i,H}^{jG,*}; \mathcal{N}_G) \le \tau_X]$ over admissible goal-do pairs. The reference uses the same reset under the changed goal; $\mathcal{N}_G$ contains only adapter-declared non-descendants of Y, while action-mediated descendants may change.	intervention pair
World selectivity	$N_W^{-1}\sum_i \mathbf{1}[\Delta\hat{X}_i^S = \Delta X_i^{*S}]\mathbf{1}[\hat{Y}_i = Y_i]$ over physical world-do pairs.	intervention pair
Physical-do consistency	$N_X^{-1}\sum_i \mathbf{1}[d_E(\hat{E}_i, E_i^*) \le 0.05]$ after simulator intervention and observation regeneration.	reset–pair
Terminal preservation	$N_P^{-1}\sum_i \mathbf{1}[d_E(\hat{E}_i^p, \hat{E}_i^{p'}) \le 0.05]$ over validated same-terminal plan pairs.	plan pair
Path response	$N_P^{-1}\sum_i \mathbf{1}[\hat{a}_{\mathrm{div}} = a_{\mathrm{div}}^*]\mathbf{1}[\hat{m}_{1:L} = m_{1:L}^*]$ over the same validated plan pairs.	plan pair
Unsupported-effect rate	$N_B^{-1}\sum_i \mathbf{1}[\hat{E}_i \notin \mathcal{E}(X_i, Y_i)]$ after a belief-only swap, where $\mathcal{E}$ is the adapter enumeration.	belief swap
B. Deployment and attribution diagnostics		
Policy distance ECE	$\sum_{b=1}^{10} \lvert B_b\rvert N^{-1} \lvert \overline{D}_b^{\nu,\pi} - \overline{\delta}_b^{\pi}\rvert$ with equal-frequency bins fixed on development data. Only the policy-conditioned metric is used.	history pair
Policy continuation divergence	$\delta^{\pi} = 0.5H^{-1}\sum_{t=1}^{H}[0.7\,\mathrm{TV}(p(c_t), p'(c_t)) + 0.3\,\mathbb{E}_c\mathrm{TV}(p(\bar{b}_t \mid c), p'(\bar{b}_t \mid c))] + 0.5\mathcal{W}_{d_E}(p(E_H), p'(E_H))$, normalized to $[0,1]$; c_t is the canonical class and $\bar{b}_t$ the frozen binding bin.	pair and seed
Action-bound violation	$N_U^{-1}\sum_{i,u} \mathbf{1}[b_u(h_i, h_i'; D) > D(h_i, h_i')]$ over held-out covered controls; violation magnitude is the positive part.	pair and control
Sensitivity precision	$\lvert T_{\cdot 2}\rvert^{-1}\sum_{(t,g)\in T_{\cdot 2}} I_{t,g}$, where $T_{\cdot 2}$ contains the top 20% of weights and $I_{t,g} = 1$ when shared-noise closed-loop register restoration exceeds the frozen action/route/effect divergence threshold.	update component
Sensitivity recall	$(\sum_{t,g} I_{t,g})^{-1}\sum_{(t,g)\in T_{\cdot 2}} I_{t,g}$; continuations without an influential update are excluded from recall and reported in Table 23.	update component

Proof of Proposition 1. For each augmented state ω, the fixed ν_k and normalized $\pi_0(\cdot \mid h(\omega), j)$ induce the probability measure $\sum_{j,c} \nu_k(j)\pi_0(c \mid h(\omega), j)K_{j,c}(\cdot \mid \omega)$ and the analogous signal measure. Each Wasserstein term in Equation 25 is therefore the pullback of a metric between two values of one state-indexed kernel, and is a pseudometric in (ω, ω'). The contraction argument in Theorem 1 applies without the maximization step. Banach's theorem gives the unique fixed point. Pair-dependent ν would not define this state-indexed map and is precisely why that construction is ruled out.

Proof of Lemma 1. Pushing any coupling of (P_0, P_1) through (Q_q, Q_q) gives a coupling of $(Q_{q\#}P_0, Q_{q\#}P_1)$ with identical expected cost, proving the right-hand side of Equation 48 is no larger. Conversely, standard Borel spaces admit regular conditional distributions $P_0(d\omega \mid Q_q(\omega) = z)$ and $P_1(d\omega' \mid Q_q(\omega') = z')$. Any coupling of the push-forward distributions can be lifted by these conditionals to a coupling of (P_0, P_1), again with identical cost $D(z, z')$. Taking infima in both directions proves equality.

Proof of Theorem 2. Let $d^{(0)} = 0$ and $d^{(n+1)} = \mathcal{B}_k d^{(n)}$. Induction on an n-step finite-horizon Bellman recursion shows $\lvert V_{k,n}^*(\omega) - V_{k,n}^*(\omega')\rvert \le d^{(n)}(\omega, \omega')$: the immediate difference is bounded by the first term of $\mathcal{B}_k$, and the Kantorovich coupling of next histories bounds the continuation difference. Contraction gives uniform convergence of $d^{(n)}$ to d_k^* and of normalized finite-horizon values to V_k^*, proving Equation 49. Replacing maximization with the fixed cohort-level ν_k and π_0 mixture gives the policy-conditioned statement; because the mixture is independent of the compared pair, it defines one Markov kernel on the cohort. This proof applies to exact kernels and Wasserstein distance, not finite-sample Sinkhorn targets. For Equation 50, repeat the finite-horizon induction with controls restricted to $\mathcal{U}_k^{\mathrm{id}}$. At every step its Bellman backup is bounded by the corresponding maximum over $\mathcal{U}_k$, so the iterates are bounded by $d^{(n)}$ and the limiting result follows.

Proof of Theorem 3. The triangle inequality for the three operator errors gives $\lVert\widetilde{d}_\phi - \mathcal{T}\widetilde{d}_\phi\rVert_\infty \le \epsilon_{\mathrm{res}} + \epsilon_{\mathrm{est}} + \epsilon_{\mathrm{miss}}$. Since $d_{\mathcal{T}} = \mathcal{T}d_{\mathcal{T}}$ and $\mathcal{T}$ is a γ-contraction,

$$\begin{aligned}\lVert\widetilde{d}_\phi - d_{\mathcal{T}}\rVert_\infty &\le \lVert\widetilde{d}_\phi - \mathcal{T}\widetilde{d}_\phi\rVert_\infty + \lVert\mathcal{T}\widetilde{d}_\phi - \mathcal{T}d_{\mathcal{T}}\rVert_\infty \\ &\le \epsilon_{\mathrm{res}} + \epsilon_{\mathrm{est}} + \epsilon_{\mathrm{miss}} + \gamma\lVert\widetilde{d}_\phi - d_{\mathcal{T}}\rVert_\infty.\end{aligned}$$

Rearranging proves Equation 54.

Proof of Proposition 2. For the exact theoretical signal witness, Equations 27– 28 give $\Delta_\rho = \sum_g \beta_g \Delta_g$. For the sampled practical witness, define $\widehat{\Delta}_\rho^{\varepsilon,M} = \mathbb{E}_{\Lambda_\varepsilon^\star} c_\rho$ and $\widehat{\Delta}_g^{\varepsilon,M} = \mathbb{E}_{\Lambda_\varepsilon^\star} c_g$. The same cost partition gives $\widehat{\Delta}_\rho^{\varepsilon,M} = \sum_g \beta_g \widehat{\Delta}_g^{\varepsilon,M}$ under the shared entropic signal coupling. Linearity of expectation under that coupling and the shared next-state coupling $\Gamma_\varepsilon^\star$ then gives $\sum_g \beta_g T_g = \widehat{\mathcal{B}}_k^{\varepsilon,M} D_{\bar{\phi}}$. Subtracting this identity from $D_\phi = \sum_g \beta_g d_\phi^g$ gives the equality in Equation 58; its inequality is the triangle inequality. No interchange of a maximum and a convex combination is used.

Proof of Theorem 4. Equation 55 ensures that every coupling of either pair of next-state kernels—in particular, an optimizer for the exact cost and an optimizer for the learned cost—is supported on $\mathcal{D}_{t+1}$. Using the learned-cost optimizer for one direction and the exact-cost optimizer for the reverse direction gives

$$\left| \mathcal{W}_{\widehat{d}_{t+1}}(K_u(\cdot \mid \omega), K_u(\cdot \mid \omega')) - \mathcal{W}_{d_{t+1}}(K_u(\cdot \mid \omega), K_u(\cdot \mid \omega')) \right| \le \|\widehat{d}_{t+1} - d_{t+1}\|_{\infty, \mathcal{D}_{t+1}}.$$

Taking the maximum over the common control set preserves this bound. The terminal discrepancy is at most $\varepsilon_{H-1}^{\sup}$; applying the one-step residual inequality backward therefore yields $\sum_{t=0}^{H-1} \gamma^t \varepsilon_t^{\sup}$, proving Equation 56. For the occupancy statement, fix the witness rule w_t before evaluation and apply its non-maximizing backup $\mathcal{T}_t^w$. Integrating the corresponding one-step recursion under η_t, the coupling selected by w_t pushes η_t to η_{t+1}, so the continuation expectation is exactly the next occupancy term. Iteration proves Equation 57. Freezing w_t is essential: an occupancy average for one witness rule does not upper-bound a different statewise maximizing control.

Proof of Proposition 3. The total-variation discrepancy between the deployed concrete-action policy and the reference class-plus-binding policy is at most $\epsilon_{\text{class}} + \epsilon_{\text{bind}} + \epsilon_{\text{out}}$ by conditioning on covered classes and applying the triangle inequality. Because $\mathcal{S}$ is forward invariant under both kernels, this statewise bound holds at every continuation state reached from either compared initial state. The discounted simulation lemma on $\mathcal{S}$ for the normalized return defined in Theorem 2 bounds $|V^\pi(\omega) - V^{\pi_0}(\omega)|$ by this quantity divided by $1-\gamma$. Apply this at ω and ω', insert Theorem 2, and replace d_k^{ν,π_0} by $\widehat{d}_\phi$ using Theorem 3. This yields Equation 62.

Proof of Lemma 2. If the route predictor is independent of its swapped P input, then $\widehat{q}_\psi(p; z) = \widehat{q}_\psi(p'; z)$ and their population Jensen–Shannon divergence is zero. Its hinge is therefore $[\delta_{\text{path}} - 0]_+ = \delta_{\text{path}}$. The finite-sample null calibration does not alter this population identity. Hence a predictor that ignores P cannot minimize the route objective on a feasible pair with distinct route labels. Closed-loop route swaps separately test whether the frozen deployment policy consumes the learned plan interface.

E.25 Execution protocol

Appendix E defines the validity gate, confirmatory estimands, crossed resampling, frozen strata, and failure denominators. The execution contract adds only system-level controls: every newly executed method within a benchmark shares the backbone checkpoint, SFT data, generated-token and tool-call ceilings, environment revision, and task split; full adapters and multi-seed runs begin only after the validity gate.

The task-level record is the link between execution and statistical analysis. Benchmark, task, replicate, training seed, backbone, and configuration identify the comparison unit; the metric and protocol fields specify how its outcome enters an aggregate. Paired contrasts join records on their shared evaluation units before computing differences, preserving the correspondence between methods throughout resampling. Intervention-pair diagnostics and update-level measurements retain their own identifiers and denominators. This organization allows a reported effect to be traced from its aggregate through the paired observations to the corresponding execution configuration, while keeping task success, intervention consistency, and operator diagnostics distinct.

E.26 Reproducibility and artifact commitment

The supplementary artifact provides a runnable DeliverySCM reference implementation of recurrent goal, world, and plan registers, executable intervention controls, and shared-witness recursive targets.

The implementation exposes the entropic transport coupling and its unregularized ground-cost evaluation, EMA target construction, component attribution, and recurrent consistency objectives as separate code components. Configuration files specify the model and optimizer settings; command-line entry points support training, checkpoint loading, and evaluation.

A verification script runs the unit tests, a short CPU training example, checkpoint evaluation, and control-library export without privileged credentials. The tests cover transport calculations, intervention behavior, shared-witness additivity, gradient propagation, and checkpoint loading. An exhaustive finite-state check enumerates the DeliverySCM physical states and controls, solves the Bellman metric, and checks its residual, metric properties, and fixed-policy value bound. The artifact also provides execution instructions and a Slurm entry script for running the reference configuration on a compute cluster.

The verification workflow follows the interfaces used by the reference learner. It first checks the transport and intervention primitives, then exercises the training stages and writes a checkpoint that can be loaded by the standalone evaluation entry point. Control export makes the intervention/action combinations available for inspection, and the enumerated operator check supplies a numerical reference for the recursive target. These outputs serve complementary purposes: unit tests check local invariants, checkpoint evaluation checks persistence of the learned state, and finite-state diagnostics check the operator induced by the environment and control library. The configuration file and execution instructions connect these checks into a repeatable command-line workflow.

The protocol specification records the experimental settings and H1–H3 contrasts; the task-level schema defines the units for subsequent experimental records. Upon publication, we will release the DynaCAST implementation, training and evaluation scripts, configuration files, and accompanying documentation in a public repository under an open-source license.

Table 47: Exact finite operator check across 24 enumerated delivery SCMs.

γ	Fixed-point gap	Infinite RHS	$H = 4$ gap	$H = 4$ RHS	Informative
0.80	0.061	0.090	0.032	0.053	both
0.90	0.128	0.210	0.049	0.072	both
0.95	0.284	0.580	0.067	0.108	both

Table 48: Learning-object comparison. Closure makes controls comparable across histories with different literal action sets; typed target describes component attribution.

Method family	State criterion	Compared controls	Open-action closure	Typed target
Structured memory (Bulatov et al., 2022; Packer et al., 2023)	Retention / reconstruction	Observed trajectory	None in objective	None
World models / DeepMDP (Ha & Schmidhuber, 2018; Gelada et al., 2019)	Predictive latent dynamics	Factual actions	Fixed interface	None
Bisimulation / DBC (Ferns et al., 2011; Zhang et al., 2020)	Reward–transition similarity	Shared factual actions	Common set assumed	Global
Causal interchange (Geiger et al., 2022)	Supplied-abstraction agreement	Internal swaps	Supplied map	Swap-local
DynaCAST	Executable-future equivalence	Interventions × canonical actions	Frozen bindings	Shared recursive witness

Together, Tables 47 and 48 complete the finite-state and learning-object validation record. Table 47 enumerates every admissible intervention and canonical action in a small state space, so the fixed point, truncated backup, and fixed-policy value bound are evaluated without sampling assumptions; varying the discount factor checks continuity as future effects receive more weight. Table 48 records the factual or intervention-indexed controls, frozen open-action closure, and shared witness that define comparability and component attribution. Together these checks link implementation regression tests to the learning criterion and close the reproducibility protocol used in the main experiments.

The finite-state check is intentionally separate from benchmark performance: it tests transport, control enumeration, and recursive target construction on a fully enumerable system. The learning-object comparison then specifies which controls and targets make two histories comparable, so the ablations retain a fixed semantic reference while sharing the same recurrent backbone. This separation also clarifies how to investigate a discrepancy: inspect the control export for admissibility, the transport tests for coupling behavior, and the operator residual for recursive consistency before comparing the resulting policy outcomes.