FROM PREDICTION TO SELECTION: OUTCOME SUPERVISION FOR CLOSED-LOOP BEST-OF-K

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate outcome prediction across states need not imply accurate action ranking within a state. We study this distinction in repeated best-of-five control with a frozen actor. Alongside a theoretical construction, we compare four critic-training conditions: Base, ordinary-rollout augmentation, same-state counterfactual (CF) outcome supervision, and a control that replaces candidate targets with their within-state mean. Across 40 LIBERO tasks, CF yields absolute success-rate gains over Base of 4.44% and 3.75% for IQL and CQL, respectively, with confidence intervals across tasks and initializations spanning zero. In the PointMaze UMaze Fast64 configuration, CF improves normalized return by 5.20 points over Base and 2.89 points over the mean-target control, with positive confidence intervals for both gains. Candidate-specific gains remain uncertain in LIBERO, and CF effects vary across UMaze configurations. Local LIBERO ranking differences have confidence intervals spanning zero. The study separates the overall effect of outcome supervision from the effect of retaining candidate-specific targets.

1 INTRODUCTION

> *"When a measure becomes a target, it ceases to be a good measure."*
> Marilyn Strathern (1997, p. 308)

This formulation of Goodhart's law motivates a concrete question for learned critics: what changes when an outcome predictor is used to select actions? A score can separate easy states from difficult ones while ordering the actions available at one state poorly. Its state-level component helps prediction but is shared by every candidate in an action comparison. When a frozen robot policy proposes several actions, a critic must choose among them; each choice also changes the states encountered later. The relevant outcome is therefore the utility of repeated selection over a complete episode.

Value-guided policies and learned verifiers already select among policy proposals (Nakamoto et al., 2024; Hansen-Estruch et al., 2023; Kwok et al., 2025; Dai et al., 2025). Simulation also has a long history in rollout-based policy improvement (Dimitrakakis & Lagoudakis, 2008). We study the *supervision of the deployed critic*, holding the actor and selection rule fixed. Ordinary trajectories supply the outcome of the action taken. Restoring a saved simulator state (an *anchor*) supplies outcomes of alternative actions from that same state. We average repeated continuations into auxiliary targets and train the critic to use these data at deployment, without online branching.

Our four-arm study asks three questions. Does adding same-state outcome supervision improve closed-loop utility over Base training? How does it compare with adding ordinary trajectories to temporal-difference (TD) training? What does retaining each candidate's target add beyond assigning every candidate the anchor's mean target? The last comparison keeps the candidate inputs and training draws, separating the overall supervision effect from the effect of the auxiliary target structure (Figure 1; Table 1).

CF has the highest mean success in LIBERO, with task/root intervals for its gains over Base spanning zero. In the UMaze configuration using the original TD dataset (Fast64), candidate-specific targets improve return over anchor mean. Local LIBERO ranking differences have intervals spanning zero. We report both complete-episode utility and local scoring measurements.

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.

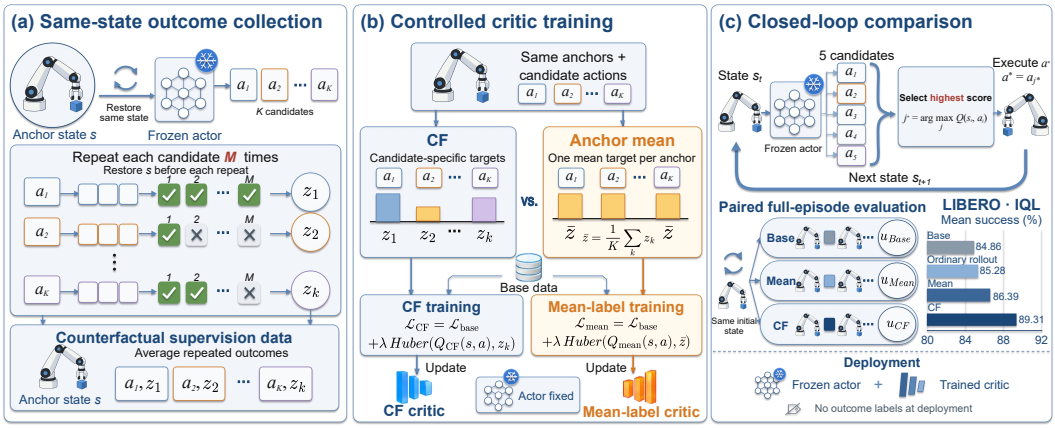


Figure 1: Same-state outcome supervision and closed-loop evaluation with a frozen actor. (a) Restore an anchor before each candidate continuation and average its outcomes. (b) CF retains candidate-specific targets; anchor mean assigns the same mean target to every candidate. The training expressions are schematic; both arms compute Huber loss per Q head before averaging, as in Eq. 2, using z_{uk} for CF and $\bar{z}_u$ for anchor mean. (c) Repeated best-of-five deployment uses no outcome labels or simulator branching. Base and ordinary-rollout provide the TD-training comparisons. The inset reports IQL mean success on the current 240-root LIBERO panel; Table 2 gives the full four-arm results.

Contributions. First, a construction and local response analysis explain how state-averaged prediction and candidate selection can differ. Second, four-arm comparisons in LIBERO and PointMaze distinguish adding outcome supervision from retaining candidate-specific auxiliary targets. Third, local diagnostics compare within-state ordering with closed-loop utility, and configuration comparisons assess how the observed effects vary across data and training settings.

2 What an evaluation must distinguish

At decision h, a fixed proposal mechanism built from a frozen actor π produces $C_h = (a_h^{(1)}, \dots, a_h^{(K)})$ in context x_h. Context includes the observation and, when used, the remaining horizon. A selector σ chooses an index and executes the corresponding action or action-chunk prefix. Its complete-episode utility is $J(\sigma) = \mathbb{E}_\sigma[U]$: episode success for LIBERO, normalized undiscounted return for PointMaze. These endpoints are reported separately. Uniform index selection is σ_0; its action distribution equals the original actor's only when the candidate-generation mechanism has that marginal.

Prediction does not determine candidate preference. For a reference continuation σ_0, define

$$q_h^0(x, C, k) = \mathbb{E}[U \mid x_h = x, C_h = C, \text{ execute } k, \text{ then continue with } \sigma_0].$$

Changing only the current decision has conditional effect

$$\sum_{k=1}^{K} \big(\sigma(k \mid x, C) - 1/K\big) q_h^0(x, C, k). \tag{1}$$

Adding the same state-only score to every candidate leaves the argmax unchanged. Equation 1 likewise depends on differences between continuation values at one context, which cross-state prediction need not recover.

Proposition 1 gives a motivating construction. Let states be $X \in \{-b, b\}$, candidates be Gaussian A_k, and success after executing A have probability $\ell(X + A)$, where ℓ is logistic. The two noisy scorers $V_\pm(X, A) = X \pm \gamma A + \eta$ have the identical state-only predictor $\mathbb{E}_{A,\eta}[V_\pm \mid X] = X$. Nevertheless, maximizing V_+ improves success over uniform selection and maximizing V_- reduces it. The common state-only AUC approaches one as b grows, while the selection effects shrink toward

zero. At each finite b, the *averaged state predictors* agree and the selection effects have opposite signs.

The multi-step simulation in Appendix D.2 (Figure 7(a)) gives AUCs of 0.91 and 0.90, yet the two scorers change success by +55.74 and -39.65 pp relative to uniform selection. Here AUC uses candidate-mean scores averaged over a reference trajectory; the proposition uses a one-state predictor. Theorem 1 describes the local selection effect under Gaussian finite-horizon assumptions: a small perturbation improves utility to first order when its score direction aligns with a smoothed continuation-value gradient. The construction and local analysis motivate direct measurements of selection for the learned critics.

Separating supervision components. Counterfactual targets can be written as an anchor mean plus candidate differences, $z_{uk} = \bar{z}_u + (z_{uk} - \bar{z}_u)$. Comparing CF to Base changes both components of the auxiliary target. Comparing CF to an anchor-mean control instead keeps the anchors and candidate inputs while removing the auxiliary target differences. Section 3 specifies this control and the training components it retains.

This decomposition connects the prediction–selection distinction to the training design. Ordinary trajectories record the outcome of the action taken at each visited state. Branching instead holds the initial context fixed and observes several candidate continuations, supplying both an estimate of the anchor's average outcome and differences among its candidates. Regressing these outcomes can change both the shared level and the relative values of critic scores. Although a shared score shift leaves the current argmax unchanged, fitting shared targets through an action-conditioned network can also change its parameters and subsequent predictions. The effect of training on anchor means therefore requires an empirical control.

In particular, the anchor-mean critic retains action inputs and the original TD objective, so it can still assign different scores to different actions. The CF-minus-mean comparison asks what candidate-specific auxiliary targets add to that learner. Using the same retained anchors, candidates, and training draws keeps the comparison focused on the target intervention. The mathematical construction motivates this question; whether outcome regression improves the trained selector is determined by the deployment comparison.

Three distinct evaluation objects. *Cross-state prediction* uses state-averaged scores in the construction and reference-trajectory averages in the controlled simulation. *Within-state ordering* is measured for the LIBERO critics by pairwise AUC and selected-candidate success after one intervention followed by the frozen actor (K1). *Closed-loop utility* measures complete episodes with repeated critic-guided selection (K5). The last two differ in continuation policy and state occupancy. Appendix A.3 separately states what selection logs identify under different coverage assumptions.

3 Supervision sources and deployment

All comparisons freeze the actor and deploy the trained critic with repeated best-of-five selection (K5). Base trains on the original transition data. Ordinary-rollout adds actor trajectories to the TD pool. CF and anchor mean add auxiliary outcome targets to the Base objective. Simulator resets supply training labels; deployment uses only the frozen actor and trained critic. A *root* denotes a task-specific reset initialization, and an evaluation *panel* is a fixed set of roots shared across conditions.

Collecting same-state candidate outcomes. At anchor x_u, save the simulator state and generate $K = 5$ candidates. Restore the state before each candidate's M continuations, execute that candidate, and follow the collection continuation to the episode limit. With branch target G_{ukm}, the auxiliary label is $z_{uk} = M^{-1} \sum_m G_{ukm}$. LIBERO uses $M = 3$ binary eventual-success outcomes with discount one. The actor proposes ten-action chunks (H10); each intervention executes the first five actions (H5). The continuation uses one frozen-actor proposal at each decision (K1). PointMaze uses $M = 8$ cumulative-return continuations with discount 0.99 and the frozen K1 actor. At each anchor the candidate set, horizon, and continuation are shared. Labels estimate outcomes under this collection continuation. Both caches are outcome-selected: LIBERO retains anchors with a positive candidate success-count range; the main UMaze configuration retains failed-parent anchors

Table 1: Controlled comparisons and the questions they address.

Comparison	Matched or retained	Interpretation
CF − Base	Actor, deployment rule, base recipe, evaluation panel	Overall effect of adding the auxiliary supervision package
CF − ordinary-rollout	Actor and deployment rule; ordinary trajectories augment TD data	Changes support, target type and sampling; budgets differ in UMaze
CF − anchor mean	Within each configuration: anchors, candidate inputs, paired draws	Effect of replacing mean targets, including changed variance and gradients

with a candidate-return range of at least one. Anchor mean reuses these retained anchors without reapplying the filter.

Critic training and deployment. Across the critic implementations, the CF auxiliary term is

$$\mathcal{L}_{\text{aux}} = \mathbb{E}_{u,k}\left[\frac{1}{2}\sum_{j=1}^{2}\text{Huber}_1\big(Q_j(x_u, a_{uk}) - z_{uk}\big)\right], \qquad \mathcal{L} = \mathcal{L}_{\text{base}} + 0.1\mathcal{L}_{\text{aux}}. \tag{2}$$

Huber errors are computed separately for each Q head and then averaged. The auxiliary objective is value regression. IQL uses Q TD regression and V expectile regression (Kostrikov et al., 2022). For LIBERO, we implement a finite-candidate variant of CQL (Kumar et al., 2020) with twin Q heads and no V head. Our implementation uses a data-action-anchored conservative penalty and an additional soft Q-range penalty. The objectives and implementation details are in Appendices B.2 and B.3. PointMaze uses the separate IQL update specified in Appendix C.1. At each deployment decision the critic selects $k^* = \arg\max_k \min_j Q_j(x, a^{(k)})$.

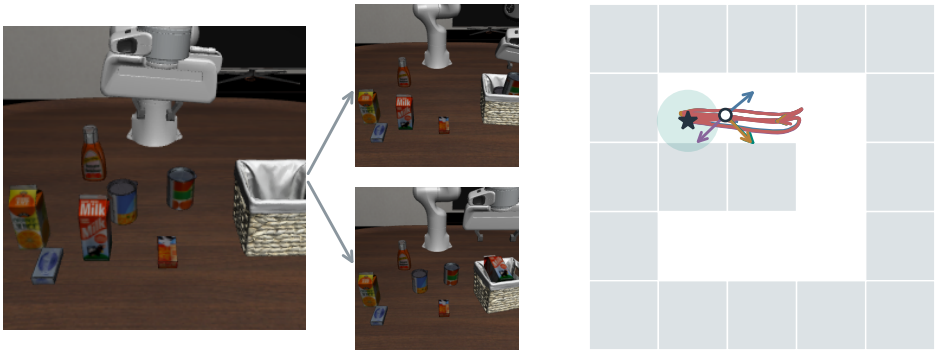
Anchor-mean control. At each retained anchor, this control replaces the five candidate targets by their arithmetic mean, without reapplying the outcome filter. It retains the action-conditioned Q heads, candidate inputs, Base TD supervision, initialization, and paired training draws. Averaging changes target variance and gradients along with candidate differences; the comparison measures the combined effect of this change in auxiliary targets.

LIBERO data and comparison. The fixed Base set contains 80 successful and 320 failed trajectories. CF uses 237 anchors and five candidates per anchor. Ordinary-rollout expands the TD pool with collected full trajectories and retains the Base sampling quotas; after pool expansion, its row draws are not identical to Base draws. In LIBERO, ordinary-rollout matches the 631,871 retained CF branch policy actions; this count excludes historical filtering, setup, and settling. Within each algorithm, CF and Base share initialization and Base draws.

PointMaze UMaze data and comparison. The main UMaze configuration, denoted Fast64, uses a frozen offline IQL actor (checkpoint S1), 181 retained branch anchors, and absolute-return auxiliary targets. Base and CF train on the original 1,000,000-row TD pool. Ordinary-rollout adds 153,600 rows from 512 actor trajectories, bringing its pool to 1,153,600 rows. CF instead uses branch-return targets in the auxiliary loss; anchor mean uses the same branch pool. We evaluate each critic on 500 common reset initializations for 300 steps, selecting among five proposals from the fixed actor at each step. The score is normalized undiscounted cumulative return. Success does not terminate an episode, so return rewards sustained goal occupancy. CF branch collection uses 21.78 million environment actions, while the ordinary-rollout increment uses 153,600; this configuration compares supervision sources at different acquisition budgets. Appendix C.1 gives the complete protocol.

4 Experiments

We evaluate the three supervision questions through complete-episode LIBERO success and PointMaze UMaze return. Figure 2 illustrates the recorded branches; all comparisons freeze the actor.



(a) LIBERO branches (b) UMaze branches

Figure 2: Recorded collection examples. (a) LIBERO: an anchor (left) branches into selected successful (upper right) and failed (lower right) continuations from five candidates with three repeats each. (b) PointMaze UMaze Fast64: colors distinguish five candidates; arrows from the open circle show candidate actions, and paths show one recorded continuation per candidate, selected from eight repeats. The star and shaded disk mark the goal and its region.

Paired evaluation. We evaluate 40 tasks across four LIBERO suites (Liu et al., 2023) with a frozen $\pi_{0.5}$-based actor (Physical Intelligence et al., 2025) at update 3,000 (checkpoint BC84). At each decision it proposes five H10 chunks; the critic selects and executes H5 without fallback. Each condition has three training seeds, evaluated on 240 common roots (six per task) using final 15,000-update EMA critics. The panel excludes known critic-supervision roots; 86 roots appeared in earlier evaluation, and earlier actor/expert-bank exposure is unknown. Fast64 uses three training seeds and 500 common roots per model, with 300 steps and K5 selection at every step. Both studies reuse the original three-arm observations when adding anchor mean. Protocols are in Appendices B.4 and C.1; the historical 800-root LIBERO matrix remains separate (Appendix B.5).

Uncertainty. Paired seed-t intervals describe variation over trained seeds. LIBERO's task/root bootstrap holds these seeds fixed; UMaze's crossed bootstrap resamples seeds and roots. LIBERO intervals are 95%; Fast64 uses 97.50% for the original CF-minus-Base/rollout comparisons and 95% for the added CF-minus-mean comparison. The separate comparison families and descriptive contrasts are retained (Table 5); these intervals do not account for selection among configurations.

4.1 Overall closed-loop effects

For LIBERO IQL, CF raises mean success from 84.86% to 89.31%, a 4.44 pp difference (seed-t 95% CI $[2.86, 6.03]$; task/root CI $[-0.42, 9.59]$). CQL rises from 80.00% to 83.75%, a 3.75 pp difference (respectively $[-5.27, 12.77]$ and $[-1.25, 9.03]$). Both task/root intervals span zero. These are within-algorithm comparisons; IQL and CQL use different objectives and architectures.

Fast64 uses the original million-row TD pool and absolute-return branch targets. CF scores 59.67, compared with 54.47 for Base: a +5.20-point difference (seed-t 97.50% CI $[+1.54, +8.86]$; crossed-bootstrap CI $[+4.05, +6.34]$). Both interval lower bounds are positive. Return rewards earlier arrival and sustained goal occupancy because success does not terminate the episode. Fast64 was selected for primary presentation during development after observing comparative results. Section 4.4 places this within-configuration result alongside the other UMaze configurations.

4.2 What do candidate-specific targets add?

Replacing anchor means with the recorded candidate-specific targets raises Fast64 mean return from 56.78 to 59.67, a 2.89-point difference (seed-t 95% CI $[1.86, 3.91]$; crossed-bootstrap CI $[2.48, 3.36]$). The same branch inputs and acquisition costs underlie both arms. This comparison changes the auxiliary targets, including their variance and gradients. The final-step success contrast is 3.73 pp with both intervals spanning zero (Appendix C.1, final-step success intervals).

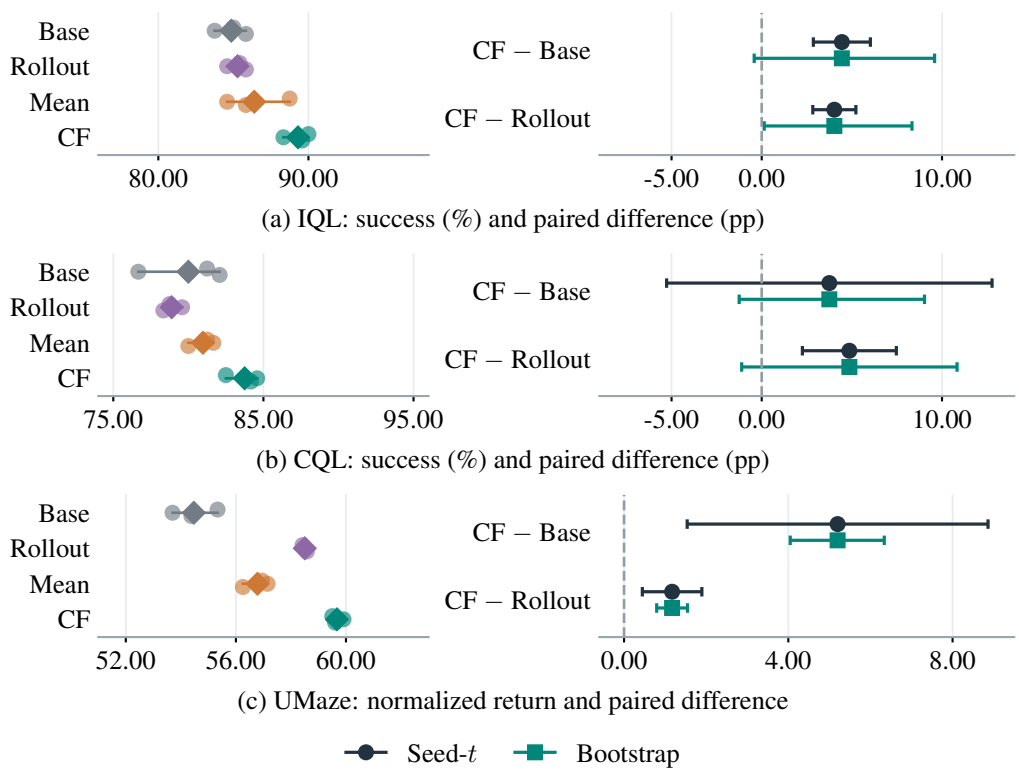


Figure 3: Closed-loop results. (a,b) LIBERO success (%) for IQL and CQL ($\alpha = 0.1$); (c) normalized cumulative return for UMaze Fast64. Left: results for four supervision arms, with dots for all three training seeds and diamonds for means. Mean denotes action-conditioned anchor-mean supervision. Right: paired CF-minus-Base and CF-minus-rollout differences; circles show seed-t intervals and squares show bootstrap intervals. LIBERO uses the 240-root panel and task/root hierarchical bootstrap with trained seeds fixed (95% intervals). UMaze uses the original 500-root panel and crossed seed/root bootstrap (97.50% intervals for its two prespecified comparisons).

On LIBERO, anchor mean reaches 86.39% and 80.97% SR for IQL and CQL. CF's increments are 2.92 and 2.78 pp, respectively, with both interval types spanning zero (Figure 4). Full paired contrasts and per-seed results are in Appendices B.4 and C.1.

4.3 Ordinary trajectories and acquisition cost

With LIBERO IQL, ordinary-rollout reaches 85.28% success. CF exceeds it by 4.03 pp (seed-t 95% CI $[2.83, 5.22]$; task/root CI $[0.14, 8.33]$). Both intervals for this descriptive contrast are positive. For CQL, ordinary-rollout reaches 78.89%; the CF difference is 4.86 pp (respectively $[2.26, 7.47]$ and $[-1.11, 10.83]$), with the task/root interval spanning zero. The retained-action budgets match, excluding historical CF filtering and setup.

Fast64 ordinary-rollout scores 58.50, leaving a +1.17-point CF increment (seed-t 97.50% CI $[+0.45, +1.89]$; crossed-bootstrap CI $[+0.79, +1.54]$). Here CF branch collection uses 21.78 million actions against 153,600 ordinary increment actions, about 142 times as many. Across both domains the arms also differ in support, target type and sampling (Table 1). The CF-minus-mean comparison avoids this acquisition mismatch by reusing the same branch pool.

4.4 Local diagnostics and configuration dependence

Local ordering and closed-loop utility. Table 2 places local scores alongside closed-loop success. The local diagnostic samples one state from each root's frozen-K1 trajectory and evaluates five candidates with six K1 continuations each. Of the 240 states, 77 supply outcome-discordant pairs for

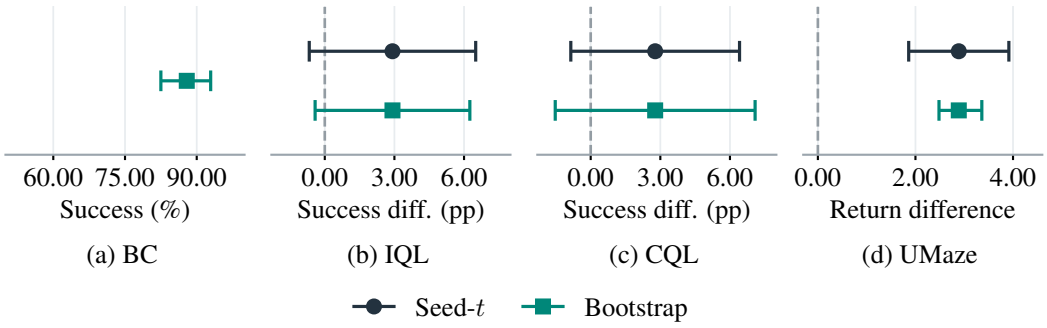


Figure 4: BC baseline and CF minus action-conditioned anchor mean. (a) Frozen BC actor: 87.92% absolute LIBERO success (211/240), from one existing K1 parent rollout per root on the same 240-root panel; bootstrap only. (b,c) IQL/CQL success differences: 2.92/2.78 pp; (d) UMaze Fast64 normalized-return difference: 2.89 points. Circles show seed-t intervals; squares show task/root bootstrap for LIBERO (saved weights) and crossed seed/root bootstrap for Fast64. All intervals are 95%; Figure 3(c) retains the original Fast64 two-comparison 97.50% family.

Table 2: LIBERO four-arm deployment and local diagnostics on the 240-root panel. SR is full-episode K5 success. Local AUC is the within-state ordering probability (chance: 50.00); selected-candidate success (SCS) follows the chosen intervention with K1. All values are percentages; local uniform selection gives 84.93% SCS. Intervals and paired contrasts are in Appendix B.4.

Algorithm	Arm	SR	Local AUC	SCS
IQL	Base	84.86	48.81	84.77
IQL	Ordinary-rollout	85.28	49.52	84.72
IQL	CF	89.31	49.63	84.88
IQL	Anchor mean	86.39	50.59	84.84
CQL	Base	80.00	51.30	85.14
CQL	Ordinary-rollout	78.89	54.68	85.76
CQL	CF	83.75	49.59	84.88
CQL	Anchor mean	80.97	51.04	84.98

AUC. CF-minus-mean AUC differences are -0.97 pp for IQL and -1.45 pp for CQL, with both 95% interval types spanning zero. Selected-candidate success is also close to the local uniform reference. An earlier disjoint diagnostic gives similarly uncertain preference changes (Appendix D.1).

K1 labels evaluate one intervention followed by the actor; K5 deployment changes later decisions and state occupancy. The continuation policy and the number of informative states affect how the local diagnostic relates to closed-loop utility. Fast64 was evaluated through closed-loop outcomes, without a corresponding local AUC measurement.

Effects vary across configurations. Table 3 summarizes the other UMaze comparisons. The expanded-TD configuration has near-zero CF-minus-Base and negative CF-minus-rollout estimates. The success/failure-mixture configuration matches 1,484,400 branch actions to the ordinary increment, yet its absolute- and relative-target CF estimates are below Base. Their corrected seed-t intervals cross zero. The quarter-rollout study's CF contrasts are descriptive; CF is below the full ordinary control under both intervals. Since support, targets, budgets and panels vary together, their individual contributions remain unresolved. Appendix C.2 maps the configuration names and gives complete results.

4.5 Interpreting the supervision comparisons

The four arms separate two questions that a Base-versus-CF comparison alone cannot resolve. Adding branch outcomes changes the supervision available to the critic, while replacing anchor means with candidate targets changes how that supervision distinguishes actions at each anchor. A

Table 3: UMaze configuration comparisons in normalized-return points. Counts denote Base TD transitions and added ordinary-rollout actions. R25 uses one quarter of the full increment (R); both rows share the CF-minus-Base contrast. Targets are absolute except rel. (relative); the two mixture rows share data pools and budgets. Each configuration retains its evaluation panel; full intervals and comparison families are in Table 27.

Configuration	TD rows	Added actions	CF − Base	CF − ordinary
Original TD (Fast64)	1,000,000	153,600	+5.20	+1.17
Expanded TD	1,153,600	1,261,200	+0.18	-1.27
Rollout control (R25)	1,000,000	315,300	+4.90	+0.37
Rollout control (R)	1,000,000	1,261,200	+4.90	-0.80
Success/failure mix	120,000	1,484,400	-0.27	-0.35
Success/failure mix (rel.)	120,000	1,484,400	-0.03	-0.10

gain over Base can include both effects. Fast64 provides evidence for an additional return benefit from candidate targets within its configuration; the LIBERO comparisons leave this increment unresolved. The anchor-mean arm is consequently useful even when CF has the highest mean: it tests how much of the observed gain requires retaining the recorded differences among candidates.

The local and complete-episode measurements address different decisions. Local scoring evaluates candidate choices at states visited by the frozen actor and under that actor's subsequent continuation. Repeated selection also changes later choices and the states at which the critic is queried. An ordering measured under the first process need not characterize the second. Moreover, only states with observed outcome-discordant pairs contribute to local AUC, whereas selected-candidate success includes the full sampled panel. These differences help interpret the diagnostic's scope without attributing the deployment gains to an unmeasured mechanism. A direct test of that mechanism would require local comparisons under states and continuations matched to repeated deployment.

The acquisition comparison answers a further question. Ordinary-rollout adds transitions to TD training, whereas CF compresses branch continuations into auxiliary targets. Their difference therefore combines the data support, learning objective, and collection allocation. Matching retained actions in LIBERO constrains one resource measure; reusing the branch pool for CF and anchor mean gives a closer comparison of target structure. Together, the results show why supervision source, candidate targets, and complete-episode utility should be examined separately when assessing a critic for repeated selection.

5 Related work

Value guidance and verification. V-GPS reranks generalist robot actions with an offline value function while keeping the policy fixed (Nakamoto et al., 2024). IDQL weights samples from a diffusion behavior model using a critic (Hansen-Estruch et al., 2023). RoboMonkey studies test-time action sampling and verification for VLA models (Kwok et al., 2025), and RoVer uses a robot reward model as a test-time verifier (Dai et al., 2025). Our Base uses value-guided reranking. We hold the actor and deployment rule fixed and study the critic's supervision source, using ordinary trajectories and same-state branch outcomes as alternative training signals.

Rollout improvement and outcome supervision. Rollout sampling approximate policy iteration allocates simulation effort to compare actions for policy improvement (Dimitrakakis & Lagoudakis, 2008). Here simulation supplies auxiliary targets for a learned critic that is reused at every deployment decision. Our comparisons examine two design choices: adding trajectories to TD training versus regressing branch outcomes, and retaining candidate-specific targets versus their anchor means.

Offline critics. BCQ, CQL and IQL learn offline policies using different constraints and value objectives (Fujimoto et al., 2019; Kumar et al., 2020; Kostrikov et al., 2022). We keep the actor frozen and use IQL and a finite-candidate, data-anchored CQL variant to train the critic. Simulator branching supplies outcomes for several actions at each anchor. The PointMaze experiments also compare filtered and unfiltered supervision from the same branch pool.

Predictive scores and decisions. Reward-model overoptimization (Gao et al., 2022), studies of reward-model accuracy and optimization effectiveness (Razin et al., 2025), and process-verifier analyses (Rohatgi et al., 2025) examine how predictive scores behave when used to make decisions. Our construction isolates the distinction between state-averaged prediction and within-state action preferences. In the robot experiments, repeated selection changes subsequent states, so we measure both complete-episode utility and local scores. The supplementary logging analysis draws on logged-propensity evaluation and partial identification (Bottou et al., 2013; Jiang & Li, 2016; Liu et al., 2020; Manski, 2003).

6 LIMITATIONS

The main four-arm comparisons use one frozen actor per domain and three training seeds per condition. The reference-panel exposure and configuration selection in Section 4 limit independent confirmation and transfer claims. The reported 240-root CF-minus-Base differences compare trained selectors; the uniform-selection reference uses the historical 800-root panel.

CF requires restorable simulator states and outcome-selected caches; retained action counts exclude some filtering and compute costs. The target-structure contrast measures changes in variance and gradients as well as candidate labels. The theory concerns specified score distributions and local perturbations. Critic calibration and cross-state prediction quality remain unmeasured. The mechanism behind the deployment gains remains open.

7 CONCLUSION

Moving from prediction to selection requires evaluating a critic through the decisions it induces. Our four-arm study distinguishes adding branch-outcome supervision from retaining candidate-specific auxiliary targets. Their effects depend on the configuration, and the role of local ranking in the observed deployment gains remains unresolved. Supervision source and target structure are distinct design choices that we evaluate through local comparisons and complete closed-loop episodes.

REPRODUCIBILITY STATEMENT

Appendices B.1, B.2, B.4 and C.1 specify the critic objectives, data composition, sampling, optimization, evaluation pairing, and uncertainty calculations, with per-seed and per-suite results. Appendix D.2 specifies the controlled simulator and score aggregation. The theoretical appendices state assumptions and provide proofs.

AI USE STATEMENT

Generative AI tools assisted with writing refinement, code generation, and experiments. The authors take responsibility for the accuracy of the sources, computations, and final content, including material produced with AI assistance.

REFERENCES

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(101):3207–3260, 2013. URL https://www.jmlr.org/papers/v14/bottou13a.html.

Mingtong Dai, Lingbo Liu, Yongjie Bai, Yang Liu, Zhouxia Wang, Rui SU, Chunjie Chen, Liang Lin, and Xinyu Wu. RoVer: Robot reward model as test-time verifier for vision-language-action model. arXiv:2510.10975, 2025. URL https://arxiv.org/abs/2510.10975.

Christos Dimitrakakis and Michail G. Lagoudakis. Rollout sampling approximate policy iteration. arXiv:0805.2027, 2008. URL https://arxiv.org/abs/0805.2027.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2052–2062, 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html. arXiv:1812.02900.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. arXiv:2210.10760, 2022. URL https://arxiv.org/abs/2210.10760.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. IDQL: Implicit Q-learning as an actor-critic method with diffusion policies. arXiv:2304.10573, 2023. URL https://arxiv.org/abs/2304.10573.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022. URL https://iclr.cc/virtual/2022/poster/5941. arXiv:2110.06169.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html. arXiv:2006.04779.

Jacky Kwok, Christopher Agia, Rohan Sinha, Matt Foutter, Shulu Li, Ion Stoica, Azalia Mirhoseini, and Marco Pavone. RoboMonkey: Scaling test-time sampling and verification for vision-language-action models. arXiv:2506.17811, 2025. URL https://arxiv.org/abs/2506.17811.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. arXiv:2306.03310, 2023. URL https://arxiv.org/abs/2306.03310.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6184–6193, 2020. URL https://proceedings.mlr.press/v119/liu20a.html.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003. doi: 10.1007/b97478. URL https://link.springer.com/book/10.1007/b97478.

Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. arXiv:2410.13816, 2024. URL https://arxiv.org/abs/2410.13816.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv:2504.16054, 2025. URL https://arxiv.org/abs/2504.16054.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. What makes a reward model a good teacher? An optimization perspective. arXiv:2503.15477, 2025. URL https://arxiv.org/abs/2503.15477.

Dhruv Rohatgi, Abhishek Shetty, Donya Saless, Yuchen Li, Ankur Moitra, Andrej Risteski, and Dylan J. Foster. Taming imperfect process verifiers: A sampling perspective on backtracking. arXiv:2510.03149, 2025. URL https://arxiv.org/abs/2510.03149.

Marilyn Strathern. 'Improving ratings': audit in the British University system. *European Review*, 5 (3):305–321, 1997.

Appendix

A Mathematical statements and proofs **12**

A.1 Gaussian selection and finite-horizon local response 12

A.2 A state-evaluation and selection dissociation construction 13

A.3 What deterministic selection logs identify 13

B LIBERO protocols and complete results **14**

B.1 Shared LIBERO data, architecture and training 14

B.2 IQL objective 15

B.3 Finite-candidate CQL objective in the fixed-data matrix 15

B.4 Source-isolated 240-root LIBERO four-arm evaluation 16

B.5 Historical 800-root LIBERO matrix: complete comparison 22

B.6 CQL conservative-weight sensitivity 24

B.7 Detailed LIBERO outcomes 24

C UMaze protocols and configuration comparisons **27**

C.1 Fast64 UMaze: original 500-root panel and four-arm comparison 27

C.2 UMaze configuration comparison and earlier PointMaze results 29

C.3 Earlier UMaze four-arm study 30

D Diagnostics and visual evidence **34**

D.1 Earlier 240-state local G/U diagnostic 34

D.2 Controlled-MDP measurements 34

D.3 Visual evidence and deterministic example selection 35

E Supplementary logging analysis and historical scoring **37**

E.1 Observed-fork sibling-ranking measurements 37

E.2 Randomized index logging and estimation cost 37

E.3 Higher-threshold importance-sampling recovery 39

A MATHEMATICAL STATEMENTS AND PROOFS

A.1 GAUSSIAN SELECTION AND FINITE-HORIZON LOCAL RESPONSE

Consider a finite-horizon Markov process with terminal payoff $Y = g(X_{H+1}) \in [0,1]$, fixed initial distribution, and policy-independent transition kernels $P_h(dx' \mid x, a)$. At each step, independently conditional on x, draw $a^{(k)} = \mu_h(x) + \sigma_a Z_k$, $Z_k \sim \mathcal{N}(0, I_d)$, and $\eta_k \sim \mathcal{N}(0, \sigma_\eta^2)$ with $\sigma_a, \sigma_\eta > 0$. The score is $\phi_h(x) + \gamma \psi_h(x)^\top (a^{(k)} - \mu_h(x)) + \eta_k$, where $\gamma \geq 0$ and $\sup_{h,x} \|\psi_h(x)\| \leq B < \infty$. All draws are fresh and independent of subsequent environment randomness. Write J_γ for argmax selection and J_0 for uniform selection. Let $Q_h^0(x,a)$ be success probability after executing a and then following the uniform reference. Define the smoothed action-value gradient

$$G_h(x) = \nabla_b \mathbb{E}_Z[Q_h^0(x, b + \sigma_a Z)]\big|_{b=\mu_h(x)} = \sigma_a^{-1} \mathbb{E}_Z[Z Q_h^0(x, \mu_h(x) + \sigma_a Z)].$$

Gaussian smoothing makes this expression well-defined for bounded measurable Q_h^0. Let $m_K = \mathbb{E}[\max_{k \leq K} Z_k^{(1)}]$.

Theorem 1 (Selected mean and local response)**.** *For the model above the selected action has exact mean shift*

$$\delta_{\gamma,h}(x) = \frac{m_K \gamma \sigma_a^2 \psi_h(x)}{\sqrt{\sigma_\eta^2 + \gamma^2 \sigma_a^2 \|\psi_h(x)\|^2}}.$$

For fixed $H, K, d, B, \sigma_a, \sigma_\eta$, as $\gamma \to 0$,

$$J_\gamma - J_0 = \gamma \frac{m_K \sigma_a^2}{\sigma_\eta} \sum_{h=1}^{H} \mathbb{E}_0[\psi_h(X_h)^\top G_h(X_h)] + O(\gamma^2) \tag{3}$$

$$= \sum_{h=1}^{H} \mathbb{E}_0[\delta_{\gamma,h}(X_h)^\top G_h(X_h)] + O(\gamma^2). \tag{4}$$

If the coefficient of γ in Equation 3 is nonzero, it determines the sign for all sufficiently small positive γ.

Proof. Fix h, x with $\psi \neq 0$ and put $u = \psi/\|\psi\|$, $s^2 = \sigma_\eta^2 + \gamma^2 \sigma_a^2 \|\psi\|^2$, and $r = \gamma \sigma_a \|\psi\|/s$. The standardized score T_k is standard normal and $u^\top Z_k = r T_k + \sqrt{1 - r^2} E_k$, where E_k is standard normal independent of all standardized scores. Components perpendicular to u are also independent of the scores. Consequently, for the winning index the parallel component has law

$$U_r = r M_K + \sqrt{1 - r^2} E, \qquad M_K = \max_{k \leq K} T_k,$$

with E independent of M_K. This proves the mean-shift formula.

For completeness, the selected-action law has a uniform second-order expansion. Let φ_v denote a centered normal density of variance v and write $f_r(z) = \mathbb{E}[\varphi_{1-r^2}(z - r M_K)]$. Then $f_0 = \varphi_1$ and $\partial_r f_r|_{r=0} = m_K z \varphi_1(z)$. For any fixed $r_0 < 1$, differentiating a normal density twice gives

$$\sup_{|r| \leq r_0} \|\partial_r^2 f_r\|_1 \leq C(r_0) \mathbb{E}[1 + M_K^2] < \infty.$$

With $v = 1 - r^2$ and $y = z - r M_K$, the second derivative of $\varphi_v(y)$ is $M_K^2 \partial_{yy} \varphi_v + 4 r M_K \partial_{yv} \varphi_v - 2 \partial_v \varphi_v + 4 r^2 \partial_{vv} \varphi_v$. The L^1 norms of these normal derivatives are uniformly bounded for $v \in [1 - r_0^2, 1]$, giving the displayed inequality and justifying differentiation under the expectation. Taylor's integral remainder therefore implies

$$\|f_r - \varphi_1 - r m_K(\cdot) \varphi_1\|_1 \leq C_K r^2.$$

Tensoring with the independent perpendicular normal coordinates preserves this bound. Since $r = \gamma \sigma_a \|\psi\|/\sigma_\eta + O(\gamma^3)$ uniformly, for every bounded measurable F,

$$\mathbb{E}_\gamma[F(a)] - \mathbb{E}_0[F(a)] = \gamma \frac{m_K \sigma_a}{\sigma_\eta} \mathbb{E}_Z[\psi^\top Z F(\mu + \sigma_a Z)] + O(\gamma^2) \|F\|_\infty.$$

When $\psi = 0$, the selected action has the reference law exactly, so the same expansion applies. The uniform bound on ψ makes the remainder uniform in h, x.

Let $T_{\gamma,h}$ be the state transition operator after integrating out the selected action. The preceding density bound implies, in the operator norm on bounded functions, $T_{\gamma,h} = T_{0,h} + \gamma L_h + O(\gamma^2)$. Expanding the finite product $T_{\gamma,1} \cdots T_{\gamma,H}$ gives one first-order insertion L_h per step; every term with two insertions or a remainder is $O(\gamma^2)$ for fixed H. Applying the product to g and the initial law gives

$$\gamma \sum_h \mathbb{E}_0 \left[\frac{m_K \sigma_a}{\sigma_\eta} \mathbb{E}_Z [\psi_h(X_h)^\top Z Q_h^0(X_h, \mu_h(X_h) + \sigma_a Z)] \right] + O(\gamma^2),$$

which is Equation 3. The gradient identity follows by differentiating the Gaussian convolution of bounded Q_h^0. Replacing the linear coefficient by $\delta_{\gamma,h}$ changes the expression by $O(\gamma^3)$, and the sign statement follows by dividing by γ. □

A.2 A STATE-EVALUATION AND SELECTION DISSOCIATION CONSTRUCTION

Proposition 1 (Identical state evaluation, opposite selection effects)**.** *For every $K \geq 2$ there is a one-decision model with two scorers having the same state-only outcome-prediction statistics on the uniform reference, where the state-only predictor averages the score over reference actions and score noise, and respective success probabilities strictly above and below uniform selection. The common state-only ROC AUC can be made arbitrarily close to one.*

Proof. Let $X \in \{-b, b\}$ be equiprobable, $b > 0$, and independently draw scalar candidates $A_k \sim \mathcal{N}(0, \sigma_a^2)$ and score noises $\eta_k \sim \mathcal{N}(0, \sigma_\eta^2)$. After executing A, draw terminal success as a Bernoulli variable of mean $\ell(X + A)$, where $\ell(t) = (1 + e^{-t})^{-1}$. Use scorers $V_\pm(X, A_k) = X \pm \gamma A_k + \eta_k$ with $\gamma > 0$. The state-only scores satisfy $\bar{V}_\pm(X) = \mathbb{E}_{A,\eta}[V_\pm(X, A) \mid X] = X$. Uniform selection ignores the scores, so the joint law of this predictor and outcome Y is exactly the same for the two scorers.

Writing $r = \gamma\sigma_a / \sqrt{\sigma_\eta^2 + \gamma^2 \sigma_a^2} > 0$, the action selected by V_+ has law $\sigma_a(rM_K + \sqrt{1 - r^2}E)$. Couple $M_K = \max(T_1, \ldots, T_K)$ to T_1. Then $M_K \geq T_1$, with strict inequality on an event of positive probability. The coupled variable $\sigma_a(rT_1 + \sqrt{1 - r^2}E)$ is a reference normal action. Since ℓ is strictly increasing, the plus selector has strictly greater success at each X. By symmetry the minus selector has the reflected selected-action law, yielding strictly smaller success than the reference. Averaging over X preserves both strict inequalities.

For the reference evaluation, let $p_b = \mathbb{E}_A[\ell(b + A)]$. Symmetry gives $\mathbb{E}_A[\ell(-b + A)] = 1 - p_b$ and $\Pr(Y = 1) = 1/2$. With half credit for tied predictions, the ROC AUC of predictor X is $p_b^2 + \frac{1}{2}[2p_b(1 - p_b)] = p_b$. Dominated convergence gives $p_b \to 1$ as $b \to \infty$. Thus state-level predictive quality and action-selection direction can be separated within this construction. The controlled multi-step experiments report their own measured cross-state score metrics in Appendix D.2. □

A.3 WHAT DETERMINISTIC SELECTION LOGS IDENTIFY

The following statements distinguish conditional index coverage from state–action coverage. They use separate statistical models.

Proposition 2 (Sharp interval with an unrestricted missing response)**.** *Consider one decision context C (including candidates and scores), with a fixed context distribution, deterministic logging index $b(C)$, and bounded potential outcomes $Y(k) \in [0, 1]$. The log contains $(C, b(C), Y(b(C)))$. Assume an unrestricted conditional response law for each unexecuted index. For a known target selector σ', let*

$$A = \mathbb{E}[\sigma'(b(C) \mid C)\mu_b(C)], \quad q = \mathbb{E}[1 - \sigma'(b(C) \mid C)], \quad \mu_b(C) = \mathbb{E}[Y \mid C].$$

The sharp identified set for target success is $[A, A + q]$. Here each potential outcome can include an entire continuation under a fixed rule.

Proof. Iterated expectation gives $J(\sigma') = A + \mathbb{E}[\sum_{k \neq b(C)} \sigma'(k \mid C)\mu_k(C)]$. Each missing response mean lies in $[0, 1]$, which proves the bounds. Keeping the observed conditional law fixed,

assign every missing outcome mean $t \in [0, 1]$ using independent Bernoulli potential outcomes. This changes no logged observation and yields target value $A + tq$, attaining the entire interval. For example, with two fixed candidates, $b = 1$, observed mean $1/2$, and target index 2, the assignments $\mu_2 = 0$ and $\mu_2 = 1$ produce identical logs and selection effects $-1/2$ and $+1/2$. This construction identifies the role of missing sibling responses under the stipulated unrestricted model. □

Proposition 3 (Finite Markov identification with state–action coverage)**.** *Consider a finite-state, finite-action, time-inhomogeneous Markov model with bounded terminal outcome drawn from a kernel depending only on the terminal state, observed states and executed actions, and transition and outcome kernels invariant to the selector. Suppose every state–action pair having positive target occupancy at step h also has positive logging occupancy at that step, and every target terminal state has positive logging occupancy. The initial law and the Markov target's induced action probabilities are known. The population logging law then identifies the target value.*

Proof. At every covered pair, the logging conditional distribution of the next state identifies $P_h(x' \mid x, a)$ by the Markov and selector-invariance assumptions. At covered terminal states it identifies $g(x) = \mathbb{E}[Y \mid X_{H+1} = x]$. The target value follows from backward recursion

$$V_{H+1}(x) = g(x), \qquad V_h(x) = \sum_a \pi'_h(a \mid x) \sum_{x'} P_h(x' \mid x, a) V_{h+1}(x'),$$

followed by integration over the initial law. Target-reachable successors are covered by hypothesis, so every required conditional is identified. The action probabilities π'_h integrate the known candidate-generation and selection rules. This argument operates on state–action occupancy; the previous proposition operates on unrestricted context–index responses. □

B LIBERO PROTOCOLS AND COMPLETE RESULTS

B.1 SHARED LIBERO DATA, ARCHITECTURE AND TRAINING

The current 240-root and historical 800-root comparisons reuse the same critic training recipes. Each root is a task-specific reset initialization; each panel is a fixed set of these roots. Their uncertainty calculations are specified separately below.

Identity and optimization. The frozen actor is a $\pi_{0.5}$-based actor-only checkpoint at update 3,000 (BC84); its visual features and normalization are frozen with it. The H5 IQL critic and finite-candidate CQL critic are separate implementations. The CQL Q/state-projection architecture has 8,796,162 trainable parameters and no V head. The IQL critic has twin FiLM Q heads and a V head. A zero-initialized 32-to-2048 state projection is added to the frozen pooled visual feature; each Q head projects the 2048-dimensional feature to width 512 and embeds the flattened 5-by-32 action through widths 1024 and 512 to generate FiLM scale and bias. The head MLP uses width 2048 internally, GELU activations and a scalar output. The V head uses the same feature width without action conditioning. CQL retains the Q/state components and omits V. The IQL equations are given in Appendix B.2; the finite-candidate CQL equations in Appendix B.3 define its training objective. The common Base dataset S80/F320 contains 80 successful and 320 failed trajectories (28,164 H5 transitions). The added failure subset comes from four LIBERO-10 tasks and excludes the evaluation roots for that subset. A separate E100 expert bank supplies candidate/action support, not TD targets. CQL uses $\alpha = 0.1$, conservative temperature one, full-H5 conservative rows, and a coefficient-one soft range penalty on $[-1, 1]$. Its offline actor proposals and next-state indices are aligned to the common Base data.

Within each algorithm, the initialization is shared by seed; Base and CF replay the same Base draws. The expanded ordinary-rollout pool uses the same sampling quotas, not the same indices. IQL and CQL parameters are not paired across their different architectures.

Auxiliary labels and sampling. The fixed CF237 cache contains 237 points from 217 roots, with five candidates and three continuation outcomes per candidate. It combines 144 earlier points and 93 added points. Both collections use the same BC84 actor-only checkpoint, five H10 proposals with H5 intervention, and frozen K1 continuation without a critic. The outcome fractions describe

Table 4: Shared LIBERO critic training settings. IQL and CQL use their respective objectives in Appendices B.2 and B.3; the frozen actor and visual backbone are not optimized.

Setting	Value
Training	Seeds 69–71; 15,000 updates; final EMA checkpoint
Optimizer	AdamW, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$
Learning rate	300-update warmup to 10^{-4}; cosine decay to 10^{-5}
Regularization	Kernel weight decay 0.05; gradient norm clipped at one
EMA	Decay 0.995
Base batch	64 rows: four successes and twelve failures per suite
CF batch and weight	16 anchors, all five candidates, two Q heads; weight 0.1

this continuation. The first 144 points originate from failure trajectories; the 93 added points include 77 from successful and 16 from failed parent episodes. All retained points have a positive range of candidate success counts: 139 have range one, 74 range two and 24 range three out of three repeats. Thus the LIBERO cache is outcome-selected.

CF samples 16 anchors hierarchically by suite, task, root and point, evaluating all five candidates at each anchor. Its auxiliary loss averages 80 labels over two Q heads at weight 0.1. The Base batch has four successful and twelve failed rows per suite; the dataset S80/F320 ratio is distinct from this batch ratio.

B.2 IQL OBJECTIVE

All four conditions share the frozen actor and original S80/F320 data described in Appendix B.1; ordinary-rollout expands the TD pool. Their objective is

$$\mathcal{L}_{\mathrm{IQL}} = \mathcal{L}_Q + \mathcal{L}_V + \lambda \mathcal{L}_{\mathrm{aux}}, \qquad \lambda_{\mathrm{Base}} = \lambda_{\mathrm{rollout}} = 0, \quad \lambda_{\mathrm{CF}} = \lambda_{\mathrm{mean}} = 0.1.$$

Here $\mathcal{L}_{\mathrm{aux}}$ is Eq. 2, using candidate targets z_{uk} for CF and anchor means $\bar{z}_u$ for the mean-label control. Writing $\bar{Q}, \bar{V}$ for EMA networks and d for the terminal indicator, the implemented objectives are

$$\begin{aligned} y_{\mathrm{IQL}} &= \mathrm{sg}\Big[\mathrm{clip}_{[0,1]}\{r + (1-d)\bar{V}(s')\}\Big], \\ \mathcal{L}_Q &= \mathbb{E}\left[\tfrac{1}{2}\sum_{j=1}^{2} \mathrm{Huber}(Q_j(s,a) - y_{\mathrm{IQL}})\right], \\ \delta_V &= \mathrm{sg}[\min_j \bar{Q}_j(s,a)] - V(s), \qquad \mathcal{L}_V = \mathbb{E}[|0.7 - \mathbf{1}\{\delta_V < 0\}|\delta_V^2]. \end{aligned}$$

Here sg stops gradients and Huber uses threshold one. The discount is one. Q and V are optimized jointly; the optimizer and final-EMA evaluation convention are shared with CQL (Table 4). CF targets are empirical success fractions from three continuations for each of five candidates. Ordinary-rollout expands the TD pool without an auxiliary loss. In the historical 800-root matrix, the original six Base/CF critics contribute 4,800 episodes; those observations are reused, not reevaluated. Their per-seed and suite-level results appear once in Tables 13 and 15.

B.3 FINITE-CANDIDATE CQL OBJECTIVE IN THE FIXED-DATA MATRIX

The objective below applies to the S80/F320 matrix at seeds 69–71.

Critic and Bellman target. The actor and feature backbone are frozen. The critic has two Q heads and a state projection (8,796,162 trainable parameters), with no V head. Let a_b denote the recorded H5 action, $a_1, \ldots, a_5$ the actor proposals, d the terminal indicator, and $\bar{Q}$ the EMA critic. The target

and TD loss are

$$y_{\mathrm{CQL}} = \mathrm{sg}\left[\mathrm{clip}_{[0,1]}\left\{r + (1-d)\max_{k\le 5}\min_{j\le 2}\bar{Q}_j(s', a'_k)\right\}\right],$$

$$\mathcal{L}_{\mathrm{TD}} = \mathbb{E}\left[\tfrac{1}{2}\sum_{j=1}^{2}\mathrm{Huber}(Q_j(s,a_b) - y_{\mathrm{CQL}})\right].$$

Discount is one; terminal rows have no bootstrap. Huber uses threshold one. The proposals are cached from the frozen actor at the current and next H5 states.

Conservative term and range regularization. For each head, write $q_{jb} = Q_j(s, a_b)$ and $q_{jk} = Q_j(s, a_k)$. At temperature one, the data-anchored gap is

$$g_j(s) = \log\left(1 + \sum_{k=1}^{5}\exp(q_{jk} - q_{jb})\right) - \log 6.$$

The data-action value is differentiated through this term. Let m_i indicate that all five actions of recorded transition i are valid. For a batch of size B, define

$$\mathcal{L}_{\mathrm{cons}} = \frac{1}{2}\sum_{j=1}^{2}\frac{\sum_{i=1}^{B} m_i g_j(s_i)}{\max(1, \sum_i m_i)}, \qquad R(q) = \max(-1-q, 0)^2 + \max(q-1, 0)^2.$$

Only full-H5 rows contribute to the conservative term, while all rows retain TD supervision. The action mask is used in the loss, not as a critic input. The range penalty averages $R(q_{jk})$ over both heads and all current-state Base proposals, including states whose recorded action has fewer than five valid steps. The complete objective is

$$\mathcal{L}_{\mathrm{CQL}} = \mathcal{L}_{\mathrm{TD}} + 0.1\mathcal{L}_{\mathrm{cons}} + \mathbb{E}_{i,j,k}R(q_{ijk}) + \lambda\mathcal{L}_{\mathrm{aux}},$$
$$\lambda_{\mathrm{Base}} = \lambda_{\mathrm{rollout}} = 0, \qquad \lambda_{\mathrm{CF}} = \lambda_{\mathrm{mean}} = 0.1.$$

Here $\mathcal{L}_{\mathrm{aux}}$ is Eq. 2, using z_{uk} for CF and $\bar{z}_u$ for anchor mean. The range term is a soft penalty, not a clipped Q output or a probability label. Deployment ranks the original five actor candidates by minimum twin-Q, without Q clipping or adding the data action as a sixth candidate.

Shared protocol. Data, sampling, optimizer and final-EMA evaluation follow Appendix B.1 and Table 4. Panel-specific paired uncertainty is in Appendices B.4 and B.5. The auxiliary value loss averages Huber errors over both Q heads and all five candidates at each of 16 sampled anchors. Current- and next-state proposal caches are aligned to the corresponding TD pool, including the ordinary-rollout extension.

B.4 Source-isolated 240-root LIBERO four-arm evaluation

Panel identities and reuse. The main LIBERO results use 240 roots (six per task), distinct from both the historical 800-root matrix in Appendix B.5 and the earlier 240-state G/U diagnostic in Appendix D.1. These are different panels and estimands; their episodes and uncertainty calculations are not pooled. All 240 roots come from a 600-root pool isolated from known critic supervision. After excluding the earlier 240-state roster, a frozen hash ordering selects ranks 1–6 within each task; rank 0 is reserved for eight pilot checks. With normalized task/init identities, the new roster intersects neither the earlier 240 roster nor the known Base, ordinary-rollout, CF, or mean-label supervision sources. For IQL these sources contain 400, 1,400, 466, and 466 unique nonexpert initializations, respectively. Of the 240 roots, 86 occurred in the 800-root evaluation; earlier actor/E100 exposure is unknown. The historical results and their prior development use are reported in the 800-root comparison.

For each algorithm the four-arm comparison has 2,880 complete episodes: 2,160 reused Base/ordinary-rollout/CF episodes and 720 additional mean-label episodes. IQL reuses all twelve 15,000-update EMA models from the 800-root work, including the mean-label checkpoints. CQL reuses its nine reference models and trains three additional mean-label models for 15,000 updates each. The CQL mean arm matches CF initialization, Base and auxiliary draws, optimizer, $\alpha = 0.1$,

auxiliary weight 0.1, and final EMA selection. At each of the same 237 anchors it broadcasts the arithmetic mean of the five labels (computed in float64 and cast to float32); no candidate is removed and no new label is collected. All mean-label critics remain action-conditioned, with the original TD objective.

Closed-loop protocol and local scoring. The actor, success definition, suite horizons, settling, K5/H10 proposals, H5 execution and minimum-twin-Q selection match the historical protocol. Models share each root's full initial input and candidate random addresses; later trajectories can diverge. The 240-root random namespace differs from the 800-root namespace, so shared initializations are not identical trials. Offline scoring uses CPU and deployment uses GPU with the recorded highest critic precision. Pilot cross-device checks found numerical discrepancies in 16 comparisons of the 18 reference models but no Top-1 or pair-order change; the added mean models had neither numerical nor ranking discrepancies.

A separate local experiment samples one nonterminal H5 decision uniformly by reservoir sampling from each root's frozen-K1 parent trajectory, without selecting on Q or outcomes. Five candidate H10 proposals are intervened on for H5, each with six frozen-K1 continuations to success or the remaining absolute horizon. Thus 7,200 branch outcomes are shared by all scorers. For state s and repeat r, only candidate pairs with different binary outcomes enter the local AUC. Let D_s count these pairs and $C_s(Q)$ award one for a correct ordering, one half for an exact Q tie, and zero for an incorrect ordering. We report

$$\mathrm{AUC}(Q) = \frac{\sum_s w_s C_s(Q)}{\sum_s w_s D_s}, \qquad w_s = 1/240.$$

This is a pair-count-weighted ordering probability, not a mean of per-state AUCs. There are 77 informative states, supplying 898 eligible pairs; neither pairs nor repeats are independent training replicates. Selected-candidate success (SCS, denoted Top-1 in the detailed tables) averages the chosen candidate's six outcomes over all 240 states, with the first candidate winning exact score ties. Uniform candidate selection gives 84.93% on this local reference. Local SCS is not closed-loop SR: the former uses K1 continuations from sampled states, whereas the latter repeatedly selects K5 from initial roots. No corresponding new UMaze AUC was measured.

Table 5: Comparison families for the main four-arm panels. Confidence levels and testing roles follow the source analyses, with separate families for each listed set of comparisons.

Panel	Comparison / endpoints	Interval and testing rule
LIBERO240, original	CF − Base: IQL/CQL, AUC/SR	95%; one four-test Holm family
LIBERO240, mean extension	CF − mean: AUC/SR	95%; separate two-test Holm family per algorithm
LIBERO240, other	CF − rollout; Top-1 and other contrasts	95%; descriptive
Fast64, original	CF − Base and CF − rollout return	97.50%; original two-comparison family
Fast64, mean extension	CF − mean return	95%; one added primary contrast

Uncertainty and comparison families. All new LIBERO intervals below are 95%. Seed-t uses three paired training seeds and two degrees of freedom. The original 20,000 task-to-root bootstrap draws (seed 20260925) carry methods and repeats together within fixed suites, holding trained seeds fixed. The original CF-minus-Base family contains four tests: IQL and CQL, each with AUC and SR. The subsequent mean-label extensions each have their own two-test Holm family (AUC and SR for CF minus mean). Each family is analyzed separately; Top-1 and other contrasts are descriptive. A positive-effect designation requires all three seed differences positive, both interval lower bounds positive, and the applicable Holm $p \leq 0.05$.

The four-arm means are in Table 2 in the main text. The following tables give full intervals, contrasts and seed/task outcomes.

For CF minus mean, IQL and CQL SR estimates are positive at all three seeds, but both interval types cross zero (Holm $p = 0.15$ and 0.16). Their AUC differences are negative, -0.97 and -1.45

percentage points, with both interval types crossing zero (Holm $p = 0.45$ and 0.47). IQL CF minus Base has a positive seed-t interval and four-test Holm $p = 0.03$, but its task/root interval crosses zero. CF minus rollout is positive under both intervals for IQL, but was a descriptive comparison. CQL CF minus rollout has a positive seed interval but a task/root interval crossing zero.

Table 6: 240-root absolute metrics and both 95% intervals. All metrics are displayed in percent; AUC values are multiplied by 100.

Alg.	Arm	Metric	Mean	Seed-t CI	Task/root CI
IQL	Base	SR	84.86	$[82.26, 87.47]$	$[78.19, 90.97]$
IQL	Base	AUC	48.81	$[46.49, 51.13]$	$[41.68, 55.84]$
IQL	Base	Top-1	84.77	$[84.51, 85.03]$	$[78.94, 90.25]$
IQL	Ordinary-rollout	SR	85.28	$[83.70, 86.86]$	$[79.44, 90.42]$
IQL	Ordinary-rollout	AUC	49.52	$[44.32, 54.71]$	$[42.98, 55.54]$
IQL	Ordinary-rollout	Top-1	84.72	$[84.27, 85.18]$	$[79.00, 90.14]$
IQL	CF	SR	89.31	$[87.15, 91.46]$	$[85.28, 92.92]$
IQL	CF	AUC	49.63	$[46.58, 52.68]$	$[43.23, 55.53]$
IQL	CF	Top-1	84.88	$[83.61, 86.16]$	$[79.07, 90.30]$
IQL	Anchor mean	SR	86.39	$[81.08, 91.70]$	$[81.67, 90.69]$
IQL	Anchor mean	AUC	50.59	$[48.63, 52.56]$	$[43.87, 57.21]$
IQL	Anchor mean	Top-1	84.84	$[83.89, 85.79]$	$[79.00, 90.25]$
CQL	Base	SR	80.00	$[72.75, 87.25]$	$[73.19, 86.25]$
CQL	Base	AUC	51.30	$[48.79, 53.81]$	$[45.10, 58.21]$
CQL	Base	Top-1	85.14	$[84.79, 85.48]$	$[79.38, 90.56]$
CQL	Ordinary-rollout	SR	78.89	$[77.31, 80.47]$	$[70.97, 85.97]$
CQL	Ordinary-rollout	AUC	54.68	$[51.56, 57.79]$	$[47.24, 62.49]$
CQL	Ordinary-rollout	Top-1	85.76	$[85.14, 86.39]$	$[79.86, 91.23]$
CQL	CF	SR	83.75	$[81.01, 86.49]$	$[77.64, 89.31]$
CQL	CF	AUC	49.59	$[47.83, 51.35]$	$[43.71, 56.21]$
CQL	CF	Top-1	84.88	$[83.78, 85.99]$	$[79.07, 90.32]$
CQL	Anchor mean	SR	80.97	$[78.82, 83.13]$	$[75.00, 86.67]$
CQL	Anchor mean	AUC	51.04	$[44.08, 58.00]$	$[44.93, 57.80]$
CQL	Anchor mean	Top-1	84.98	$[84.37, 85.58]$	$[79.14, 90.49]$

Table 7: 240-root CF-minus-Base/rollout differences (pp). Holm values belong only to the original four-test CF-minus-Base AUC/SR family.

Alg.	Contrast	Metric	Difference	Seed-t 95%	Task/root 95%	Holm p
IQL	CF − Base	SR	4.44	$[2.86, 6.03]$	$[-0.42, 9.59]$	0.03
IQL	CF − Base	AUC	0.82	$[-0.76, 2.39]$	$[-5.06, 6.25]$	0.31
IQL	CF − Base	Top-1	0.12	$[-1.28, 1.51]$	$[-0.79, 1.09]$	–
IQL	CF − rollout	SR	4.03	$[2.83, 5.22]$	$[0.14, 8.33]$	–
IQL	CF − rollout	AUC	0.11	$[-6.87, 7.09]$	$[-5.93, 6.22]$	–
IQL	CF − rollout	Top-1	0.16	$[-1.39, 1.72]$	$[-0.81, 1.16]$	–
CQL	CF − Base	SR	3.75	$[-5.27, 12.77]$	$[-1.25, 9.03]$	0.31
CQL	CF − Base	AUC	-1.71	$[-4.22, 0.80]$	$[-5.84, 2.39]$	0.30
CQL	CF − Base	Top-1	-0.25	$[-1.12, 0.61]$	$[-1.16, 0.65]$	–
CQL	CF − rollout	SR	4.86	$[2.26, 7.47]$	$[-1.11, 10.83]$	–
CQL	CF − rollout	AUC	-5.09	$[-7.87, -2.30]$	$[-11.68, 1.42]$	–
CQL	CF − rollout	Top-1	-0.88	$[-1.76, 0.01]$	$[-1.92, 0.14]$	–

Table 8: 240-root mean-label contrasts (pp). Holm values apply to the two CF-minus-mean AUC/SR tests separately within each algorithm. Other entries are descriptive; dashes mark tests outside this family.

Alg.	Contrast	Metric	Difference	Seed-t 95%	Task/root 95%	Holm p
IQL	CF − mean	SR	2.92	$[-0.67, 6.50]$	$[-0.42, 6.25]$	0.15
IQL	CF − mean	AUC	-0.97	$[-5.46, 3.53]$	$[-6.29, 4.25]$	0.45

Alg.	Contrast	Metric	Difference	Seed-t 95%	Task/root 95%	Holm p
IQL	CF − mean	Top-1	0.05	$[-2.15, 2.24]$	$[-0.76, 0.86]$	–
IQL	Mean − Base	SR	1.53	$[-3.36, 6.42]$	$[-3.47, 6.67]$	–
IQL	Mean − Base	AUC	1.78	$[-1.41, 4.97]$	$[-3.04, 6.69]$	–
IQL	Mean − Base	Top-1	0.07	$[-0.89, 1.03]$	$[-0.67, 0.83]$	–
IQL	Mean − rollout	SR	1.11	$[-3.67, 5.89]$	$[-2.92, 5.42]$	–
IQL	Mean − rollout	AUC	1.08	$[-5.34, 7.49]$	$[-5.24, 7.51]$	–
IQL	Mean − rollout	Top-1	0.12	$[-0.83, 1.07]$	$[-0.72, 0.93]$	–
CQL	CF − mean	SR	2.78	$[-0.86, 6.41]$	$[-1.53, 7.08]$	0.16
CQL	CF − mean	AUC	-1.45	$[-8.43, 5.53]$	$[-5.40, 1.83]$	0.47
CQL	CF − mean	Top-1	-0.09	$[-1.70, 1.51]$	$[-0.86, 0.67]$	–
CQL	Mean − Base	SR	0.97	$[-8.07, 10.02]$	$[-3.47, 5.56]$	–
CQL	Mean − Base	AUC	-0.26	$[-4.82, 4.31]$	$[-4.83, 4.67]$	–
CQL	Mean − Base	Top-1	-0.16	$[-1.11, 0.79]$	$[-0.81, 0.42]$	–
CQL	Mean − rollout	SR	2.08	$[1.05, 3.12]$	$[-3.89, 8.06]$	–
CQL	Mean − rollout	AUC	-3.64	$[-13.34, 6.07]$	$[-9.73, 2.53]$	–
CQL	Mean − rollout	Top-1	-0.79	$[-2.01, 0.44]$	$[-1.78, 0.19]$	–

Table 9: 240-root per-seed results; each SR cell has 240 complete episodes. All displayed metrics are percentages.

Algorithm	Arm	Seed	SR	AUC	Top-1
IQL	Base	69	85.83	48.22	84.86
IQL	Base	70	83.75	48.33	84.79
IQL	Base	71	85.00	49.89	84.65
IQL	Ordinary-rollout	69	85.83	51.78	84.86
IQL	Ordinary-rollout	70	84.58	49.11	84.79
IQL	Ordinary-rollout	71	85.42	47.66	84.51
IQL	CF	69	89.58	49.55	85.07
IQL	CF	70	88.33	48.44	84.31
IQL	CF	71	90.00	50.89	85.28
IQL	Anchor mean	69	85.83	49.89	84.58
IQL	Anchor mean	70	84.58	51.45	85.28
IQL	Anchor mean	71	88.75	50.45	84.65
CQL	Base	69	82.08	50.56	85.28
CQL	Base	70	76.67	50.89	85.00
CQL	Base	71	81.25	52.45	85.14
CQL	Ordinary-rollout	69	78.33	56.12	86.04
CQL	Ordinary-rollout	70	79.58	54.01	85.56
CQL	Ordinary-rollout	71	78.75	53.90	85.69
CQL	CF	69	84.17	50.00	85.07
CQL	CF	70	84.58	48.78	84.38
CQL	CF	71	82.50	50.00	85.21
CQL	Anchor mean	69	80.00	48.44	84.72
CQL	Anchor mean	70	81.67	50.67	85.21
CQL	Anchor mean	71	81.25	54.01	85.00

Table 10: 240-root suite and task SR (%), averaged over all three training seeds. B/R/C/M denote Base, ordinary-rollout, CF, and anchor mean. Task IDs are global LIBERO IDs. Each suite has 60 roots and each task six.

Algorithm	Suite or task	B	R	C	M
IQL	libero_10	70.56	68.89	72.78	64.44
IQL	libero_goal	84.44	86.11	90.56	92.22
IQL	libero_object	95.00	93.33	96.67	95.56

Algorithm	Suite or task	B	R	C	M
IQL	libero_spatial	89.44	92.78	97.22	93.33
CQL	libero_10	60.56	62.22	67.78	62.78
CQL	libero_goal	78.89	81.11	84.44	80.56
CQL	libero_object	90.00	82.78	88.89	91.11
CQL	libero_spatial	90.56	89.44	93.89	89.44
IQL	libero_10/t30	77.78	83.33	83.33	72.22
IQL	libero_10/t31	100.00	100.00	94.44	88.89
IQL	libero_10/t32	94.44	83.33	88.89	83.33
IQL	libero_10/t33	100.00	88.89	77.78	66.67
IQL	libero_10/t34	83.33	72.22	77.78	72.22
IQL	libero_10/t35	72.22	83.33	83.33	83.33
IQL	libero_10/t36	38.89	66.67	55.56	50.00
IQL	libero_10/t37	66.67	66.67	77.78	66.67
IQL	libero_10/t38	22.22	11.11	38.89	22.22
IQL	libero_10/t39	50.00	33.33	50.00	38.89
IQL	libero_goal/t20	100.00	72.22	77.78	88.89
IQL	libero_goal/t21	100.00	94.44	100.00	100.00
IQL	libero_goal/t22	100.00	100.00	100.00	100.00
IQL	libero_goal/t23	61.11	61.11	66.67	83.33
IQL	libero_goal/t24	100.00	94.44	88.89	94.44
IQL	libero_goal/t25	44.44	66.67	83.33	66.67
IQL	libero_goal/t26	100.00	83.33	100.00	100.00
IQL	libero_goal/t27	100.00	100.00	100.00	100.00
IQL	libero_goal/t28	100.00	100.00	100.00	100.00
IQL	libero_goal/t29	38.89	88.89	88.89	88.89
IQL	libero_object/t10	72.22	88.89	88.89	100.00
IQL	libero_object/t11	100.00	88.89	100.00	100.00
IQL	libero_object/t12	100.00	100.00	100.00	100.00
IQL	libero_object/t13	100.00	88.89	100.00	100.00
IQL	libero_object/t14	94.44	100.00	100.00	88.89
IQL	libero_object/t15	100.00	83.33	88.89	77.78
IQL	libero_object/t16	100.00	100.00	94.44	100.00
IQL	libero_object/t17	100.00	100.00	100.00	94.44
IQL	libero_object/t18	100.00	100.00	94.44	94.44
IQL	libero_object/t19	83.33	83.33	100.00	100.00
IQL	libero_spatial/t0	88.89	88.89	100.00	100.00
IQL	libero_spatial/t1	83.33	100.00	100.00	100.00
IQL	libero_spatial/t2	100.00	100.00	100.00	100.00
IQL	libero_spatial/t3	100.00	100.00	94.44	100.00
IQL	libero_spatial/t4	55.56	61.11	88.89	83.33
IQL	libero_spatial/t5	94.44	100.00	100.00	94.44
IQL	libero_spatial/t6	100.00	100.00	100.00	94.44
IQL	libero_spatial/t7	94.44	94.44	94.44	83.33
IQL	libero_spatial/t8	94.44	88.89	94.44	94.44
IQL	libero_spatial/t9	83.33	94.44	100.00	83.33
CQL	libero_10/t30	61.11	66.67	77.78	66.67
CQL	libero_10/t31	88.89	88.89	100.00	88.89
CQL	libero_10/t32	83.33	88.89	83.33	72.22
CQL	libero_10/t33	83.33	72.22	94.44	83.33
CQL	libero_10/t34	55.56	88.89	55.56	50.00
CQL	libero_10/t35	88.89	61.11	88.89	88.89
CQL	libero_10/t36	38.89	77.78	38.89	61.11
CQL	libero_10/t37	38.89	50.00	88.89	66.67
CQL	libero_10/t38	27.78	11.11	16.67	16.67
CQL	libero_10/t39	38.89	16.67	33.33	33.33
CQL	libero_goal/t20	38.89	27.78	61.11	38.89
CQL	libero_goal/t21	94.44	94.44	100.00	100.00

Algorithm	Suite or task	B	R	C	M
CQL	libero_goal/t22	88.89	100.00	88.89	72.22
CQL	libero_goal/t23	33.33	66.67	66.67	55.56
CQL	libero_goal/t24	100.00	100.00	94.44	94.44
CQL	libero_goal/t25	72.22	83.33	94.44	77.78
CQL	libero_goal/t26	72.22	88.89	72.22	88.89
CQL	libero_goal/t27	100.00	100.00	100.00	100.00
CQL	libero_goal/t28	100.00	100.00	100.00	100.00
CQL	libero_goal/t29	88.89	50.00	66.67	77.78
CQL	libero_object/t10	44.44	27.78	72.22	77.78
CQL	libero_object/t11	100.00	100.00	94.44	100.00
CQL	libero_object/t12	94.44	100.00	88.89	94.44
CQL	libero_object/t13	100.00	66.67	83.33	100.00
CQL	libero_object/t14	88.89	94.44	83.33	77.78
CQL	libero_object/t15	72.22	50.00	66.67	66.67
CQL	libero_object/t16	100.00	100.00	100.00	94.44
CQL	libero_object/t17	100.00	100.00	100.00	100.00
CQL	libero_object/t18	100.00	94.44	100.00	100.00
CQL	libero_object/t19	100.00	94.44	100.00	100.00
CQL	libero_spatial/t0	100.00	94.44	88.89	77.78
CQL	libero_spatial/t1	88.89	88.89	94.44	94.44
CQL	libero_spatial/t2	100.00	100.00	100.00	94.44
CQL	libero_spatial/t3	100.00	100.00	100.00	100.00
CQL	libero_spatial/t4	77.78	72.22	94.44	88.89
CQL	libero_spatial/t5	100.00	100.00	94.44	100.00
CQL	libero_spatial/t6	100.00	100.00	100.00	94.44
CQL	libero_spatial/t7	88.89	83.33	94.44	88.89
CQL	libero_spatial/t8	94.44	100.00	94.44	88.89
CQL	libero_spatial/t9	55.56	55.56	77.78	66.67

Table 11: 240-root mean-label paired rescue/harm counts. Each row contains 240 paired episodes. Rescue means left succeeds/right fails; harm reverses it.

Algorithm	Contrast	Seed	Rescue	Harm
IQL	CF − mean	69	21	12
IQL	CF − mean	70	26	17
IQL	CF − mean	71	17	14
IQL	Mean − Base	69	19	19
IQL	Mean − Base	70	21	19
IQL	Mean − Base	71	28	19
IQL	Mean − rollout	69	21	21
IQL	Mean − rollout	70	22	22
IQL	Mean − rollout	71	24	16
CQL	CF − mean	69	30	20
CQL	CF − mean	70	27	20
CQL	CF − mean	71	24	21
CQL	Mean − Base	69	23	28
CQL	Mean − Base	70	31	19
CQL	Mean − Base	71	24	24
CQL	Mean − rollout	69	34	30
CQL	Mean − rollout	70	30	25
CQL	Mean − rollout	71	32	26

Table 12: Historical 800-root LIBERO complete-episode success. Each critic condition has three training seeds; SD is across their success rates. Uniform selection is one shared stream. Anchor mean retains CF candidate inputs but removes within-anchor target differences.

Condition	Successes/episodes	Success (%)	Seed SD
IQL Base	2,012/2,400	83.83	0.51
IQL +ordinary-rollout	2,056/2,400	85.67	0.38
IQL +CF	2,139/2,400	89.13	1.27
IQL +anchor mean	2,099/2,400	87.46	1.16
CQL Base	1,914/2,400	79.75	0.76
CQL +ordinary-rollout	1,908/2,400	79.50	1.32
CQL +CF	1,995/2,400	83.13	1.72
Shared random	673/800	84.13	–

B.5 Historical 800-root LIBERO matrix: complete comparison

The shared data, architecture, training settings and auxiliary-label sampling are specified in Appendix B.1. The following evaluation and uncertainty protocol describes the historical 800-root panel.

Evaluation and statistics. All methods use the same 40 tasks, official initial indices 0–19, environment seed 20260723 and fixed decision-0 candidate panel. The candidate request namespace is shared across all conditions; later states and candidate arrays may diverge. The Spatial/Object/Goal/LIBERO-10 horizons are 220/280/300/520, with ten settling steps, H10 proposals, H5 scoring and execution, and min-twin-Q selection without fallback. Random has an independent selection stream with seed 2026092201, uses candidate-only inference and consumes no critic outputs.

Seed-t intervals use the three paired training-seed differences and two degrees of freedom. The 20,000-draw bootstrap (seed 2026092203) samples tasks then initial roots within each fixed suite, carrying all methods and trained seeds together. It measures panel variability conditional on the trained seeds. For random comparisons, critic outcomes are first averaged across three seeds at each root, then differenced against the single shared random outcome. The two sources of uncertainty are complementary: seed-t varies training seeds, whereas this bootstrap holds them fixed. The positive-effect rule used for the reported IQL gain requires positive paired differences at all three seeds and positive lower bounds for both intervals. CF-minus-ordinary-rollout is the acquisition-control comparison; the mean-label control separately asks whether candidate-specific targets improve over anchor averages. Secondary contrasts are exploratory, and the panel has prior development exposure. Data collection used the full official set of 50 initial configurations per task; this evaluation uses the 20 indices specified above.

Cost accounting. By LIBERO-10/Goal/Object/Spatial, the frozen branch-action budget is 299,698/115,374/123,909/92,890 (631,871 total). The ordinary increment reuses 153,796/84,066/97,051/64,172 actions and adds 145,902/31,308/26,858/28,718. Collection uses the frozen actor's candidate0 policy, H10 proposals with H5 execution, and seed 2026092204. Complete successes and normal-horizon failures enter TD training; interrupted and budget-truncated tails do not. All consumed actions count toward acquisition, and collection is independent of the new evaluation outcomes.

The full historical comparison includes additional training conditions. It reuses six IQL models and specifies twelve additional models, followed by six for the rollout extension, all at 15,000 updates. The corresponding budget is 270,000 critic updates and 15,200 evaluation episodes beyond the reused models. Offline proposal inference, feature extraction, simulator setup, wall time and historical filtering are additional costs outside the action budget. The measured ordinary increment comprises 399,085 reused and 232,786 new policy actions; eight budget-truncated trajectories contribute 633 cost-only actions.

Table 13: Fixed-data matrix per-seed results. Each critic cell contains 800 episodes; random is a single shared stream.

Condition	Seed	Successes/episodes	Success (%)
IQL Base	69	670/800	83.75
IQL Base	70	675/800	84.38
IQL Base	71	667/800	83.38
IQL +CF	69	715/800	89.38
IQL +CF	70	722/800	90.25
IQL +CF	71	702/800	87.75
IQL +ordinary-rollout	69	688/800	86.00
IQL +ordinary-rollout	70	682/800	85.25
IQL +ordinary-rollout	71	686/800	85.75
CQL Base	69	634/800	79.25
CQL Base	70	635/800	79.38
CQL Base	71	645/800	80.63
CQL +CF	69	662/800	82.75
CQL +CF	70	680/800	85.00
CQL +CF	71	653/800	81.63
CQL +ordinary-rollout	69	625/800	78.13
CQL +ordinary-rollout	70	637/800	79.63
CQL +ordinary-rollout	71	646/800	80.75
Shared random	–	673/800	84.13

Table 14: Completed fixed-data contrasts (pp); seed-t and task/root bootstrap 95% intervals. A dash denotes an inapplicable seed interval for a shared random reference. Anchor-mean contrasts use the same paired analysis.

Contrast	Difference	Seed-t CI	Bootstrap CI
IQL Base − random	-0.29	–	[-4.13, +3.37]
IQL +CF − Base	+5.29	[+3.30, +7.29]	[+2.29, +8.50]
IQL +CF − +ordinary-rollout	+3.46	[-0.27, +7.19]	[+0.13, +7.04]
IQL +CF − random	+5.00	–	[+1.29, +8.67]
IQL +ordinary-rollout − Base	+1.83	[-0.23, +3.90]	[-0.96, +4.62]
CQL Base − random	-4.38	–	[-8.13, -0.88]
CQL +CF − Base	+3.38	[-2.38, +9.13]	[+0.33, +6.54]
CQL +CF − +ordinary-rollout	+3.63	[-2.36, +9.61]	[-1.29, +8.75]
CQL +CF − random	-1.00	–	[-4.54, +2.46]
CQL +ordinary-rollout − Base	-0.25	[-2.14, +1.64]	[-4.83, +4.13]
IQL CF − anchor mean	1.67	[-3.16, 6.49]	[-0.83, 4.37]
IQL anchor mean − Base	3.63	[0.52, 6.73]	[0.46, 6.87]

Table 15: LIBERO suite-level results (anchor-mean totals are in Table 12): complete-episode success (%). Overall values are three-training-seed means; SD is across those seeds. Suite entries average the same three seeds. Random is one shared selection stream (SD not applicable).

Condition	LIBERO-10	Goal	Object	Spatial	Overall	SD
IQL Base	62.50	87.17	93.83	91.83	83.83	0.51
IQL +CF	71.83	92.83	96.50	95.33	89.13	1.27
IQL +ordinary-rollout	68.67	89.83	91.67	92.50	85.67	0.38
CQL Base	56.00	87.83	89.00	86.17	79.75	0.76
CQL +CF	63.00	90.00	90.67	88.83	83.13	1.72
CQL +ordinary-rollout	63.00	84.00	85.00	86.00	79.50	1.32
Shared random	62.00	91.50	91.00	92.00	84.13	–

Historical 800-root within-anchor mean-label ablation. At each of the 237 CF anchors we replace all five auxiliary targets by their arithmetic mean, retaining the five candidate inputs. The paired seeds are 69–71; initialization, 15,000 updates, Base and auxiliary draws, loss weight, optimizer and final EMA selection match the IQL CF arm. The three models add 2,400 evaluation episodes; the Base and CF models are reused rather than reevaluated. Across the main three-arm matrix and this control there are 16,800 unique LIBERO critic evaluation episodes, before shared-reference episodes. CF-minus-mean and mean-minus-Base use the same seed-t and 20,000-draw hierarchical bootstrap definitions above. Averaging changes target variance and gradients as well as candidate differences. On this historical panel, IQL CF improves over Base by +5.29 pp and meets the original positive-effect rule. Mean labels improve over Base by 3.63 pp, while CF minus mean is 1.67 pp with both intervals crossing zero. The historical and 240-root panels are analyzed separately.

Table 16: IQL anchor-mean success by training seed. The evaluation roots are paired with those of the reused CF and Base models.

Seed	Successes/episodes	Success (%)
69	689/800	86.13
70	704/800	88.00
71	706/800	88.25

B.6 CQL CONSERVATIVE-WEIGHT SENSITIVITY

The sensitivity changes the CQL conservative weight from $\alpha = 0.1$ to 0.01 for Base and CF, retaining the actor, data, architecture, optimizer, evaluation panel, and three seeds.

Table 17: CQL sensitivity at $\alpha = 0.01$: complete-episode success and the paired change from $\alpha = 0.1$.

Condition	Success (%)	Change (pp)
Base	83.58	+3.83
CF	85.50	+2.38

CF-minus-Base is $+1.92$ pp (seed-t 95% CI $[-0.29, +4.12]$; bootstrap 95% CI $[-1.17, +4.92]$). Both intervals include zero. The mean difference is smaller than at $\alpha = 0.1$, but the across-weight contrast is itself uncertain.

B.7 DETAILED LIBERO OUTCOMES

The following tables report paired rescue/harm counts and results for all 40 tasks.

Table 18: Paired rescue/harm counts for completed contrasts. Rescue means the left condition succeeds and the right fails; harm means the reverse.

Contrast	Seed	Rescue	Harm	Episodes
IQL Base − random	69	72	75	800
IQL Base − random	70	74	72	800
IQL Base − random	71	75	81	800
IQL CF − Base	69	89	44	800
IQL CF − Base	70	82	35	800
IQL CF − Base	71	92	57	800
IQL CF − ordinary-rollout	69	75	48	800
IQL CF − ordinary-rollout	70	85	45	800
IQL CF − ordinary-rollout	71	69	53	800
IQL CF − random	69	88	46	800
IQL CF − random	70	88	39	800
IQL CF − random	71	85	56	800
IQL ordinary-rollout − Base	69	76	58	800

Contrast	Seed	Rescue	Harm	Episodes
IQL ordinary-rollout − Base	70	68	61	800
IQL ordinary-rollout − Base	71	77	58	800
CQL Base − random	69	58	97	800
CQL Base − random	70	58	96	800
CQL Base − random	71	56	84	800
CQL CF − Base	69	89	61	800
CQL CF − Base	70	102	57	800
CQL CF − Base	71	78	70	800
CQL CF − ordinary-rollout	69	112	75	800
CQL CF − ordinary-rollout	70	108	65	800
CQL CF − ordinary-rollout	71	92	85	800
CQL CF − random	69	58	69	800
CQL CF − random	70	70	63	800
CQL CF − random	71	67	87	800
CQL ordinary-rollout − Base	69	89	98	800
CQL ordinary-rollout − Base	70	86	84	800
CQL ordinary-rollout − Base	71	91	90	800

Table 19: Per-task success rates (%), averaged over trained seeds where applicable. I/Q denote IQL/CQL; B/C/R denote Base/CF/ordinary-rollout; U is shared random. Task indices are local to each suite; all 40 tasks are retained.

Task	I-B	I-C	I-R	Q-B	Q-C	Q-R	U
libero_10/0	65.00	66.67	83.33	68.33	75.00	63.33	85.00
libero_10/1	88.33	83.33	86.67	76.67	91.67	91.67	70.00
libero_10/2	71.67	83.33	93.33	83.33	66.67	85.00	80.00
libero_10/3	95.00	93.33	96.67	88.33	96.67	95.00	95.00
libero_10/4	78.33	85.00	83.33	48.33	66.67	81.67	70.00
libero_10/5	76.67	90.00	66.67	73.33	81.67	43.33	85.00
libero_10/6	38.33	55.00	61.67	31.67	25.00	66.67	50.00
libero_10/7	65.00	70.00	78.33	56.67	73.33	63.33	60.00
libero_10/8	8.33	31.67	1.67	3.33	11.67	6.67	5.00
libero_10/9	38.33	60.00	35.00	30.00	41.67	33.33	20.00
libero_goal/0	95.00	90.00	88.33	81.67	85.00	58.33	95.00
libero_goal/1	98.33	100.00	95.00	91.67	98.33	95.00	90.00
libero_goal/2	88.33	98.33	91.67	83.33	91.67	95.00	90.00
libero_goal/3	70.00	68.33	71.67	66.67	71.67	65.00	70.00
libero_goal/4	88.33	98.33	98.33	96.67	91.67	100.00	95.00
libero_goal/5	75.00	86.67	85.00	86.67	85.00	91.67	100.00
libero_goal/6	96.67	96.67	95.00	96.67	95.00	88.33	90.00
libero_goal/7	100.00	100.00	100.00	98.33	100.00	100.00	100.00
libero_goal/8	100.00	98.33	98.33	98.33	100.00	98.33	100.00
libero_goal/9	60.00	91.67	75.00	78.33	81.67	48.33	85.00
libero_object/0	91.67	93.33	86.67	45.00	73.33	38.33	75.00
libero_object/1	100.00	100.00	100.00	100.00	100.00	98.33	100.00
libero_object/2	98.33	100.00	96.67	98.33	93.33	95.00	90.00
libero_object/3	88.33	96.67	81.67	86.67	80.00	76.67	80.00
libero_object/4	98.33	100.00	96.67	100.00	100.00	96.67	100.00
libero_object/5	68.33	81.67	75.00	66.67	68.33	58.33	75.00
libero_object/6	100.00	96.67	100.00	98.33	95.00	90.00	95.00
libero_object/7	100.00	98.33	100.00	100.00	100.00	100.00	100.00
libero_object/8	100.00	100.00	100.00	96.67	96.67	98.33	100.00
libero_object/9	93.33	98.33	80.00	98.33	100.00	98.33	95.00
libero_spatial/0	90.00	91.67	91.67	88.33	78.33	80.00	85.00
libero_spatial/1	95.00	100.00	100.00	75.00	95.00	93.33	90.00
libero_spatial/2	100.00	98.33	98.33	100.00	98.33	100.00	100.00

Task	I-B	I-C	I-R	Q-B	Q-C	Q-R	U
libero_spatial/3	95.00	100.00	100.00	96.67	98.33	95.00	95.00
libero_spatial/4	68.33	91.67	65.00	73.33	85.00	38.33	80.00
libero_spatial/5	90.00	96.67	98.33	96.67	91.67	98.33	95.00
libero_spatial/6	98.33	98.33	96.67	96.67	91.67	91.67	100.00
libero_spatial/7	95.00	86.67	86.67	68.33	81.67	85.00	90.00
libero_spatial/8	96.67	95.00	95.00	95.00	93.33	95.00	100.00
libero_spatial/9	90.00	95.00	93.33	71.67	75.00	83.33	85.00

Table 20: Anchor-mean per-task results summed over three seeds. Task indices are local to each suite; each entry has 60 episodes.

Task	Successes	%	Task	Successes	%
libero_10/0	44/60	73.33	libero_object/0	51/60	85.00
libero_10/1	51/60	85.00	libero_object/1	60/60	100.00
libero_10/2	54/60	90.00	libero_object/2	59/60	98.33
libero_10/3	55/60	91.67	libero_object/3	58/60	96.67
libero_10/4	41/60	68.33	libero_object/4	59/60	98.33
libero_10/5	41/60	68.33	libero_object/5	42/60	70.00
libero_10/6	38/60	63.33	libero_object/6	59/60	98.33
libero_10/7	46/60	76.67	libero_object/7	60/60	100.00
libero_10/8	16/60	26.67	libero_object/8	58/60	96.67
libero_10/9	21/60	35.00	libero_object/9	60/60	100.00
libero_goal/0	55/60	91.67	libero_spatial/0	58/60	96.67
libero_goal/1	58/60	96.67	libero_spatial/1	58/60	96.67
libero_goal/2	60/60	100.00	libero_spatial/2	59/60	98.33
libero_goal/3	41/60	68.33	libero_spatial/3	60/60	100.00
libero_goal/4	60/60	100.00	libero_spatial/4	50/60	83.33
libero_goal/5	51/60	85.00	libero_spatial/5	60/60	100.00
libero_goal/6	59/60	98.33	libero_spatial/6	60/60	100.00
libero_goal/7	60/60	100.00	libero_spatial/7	50/60	83.33
libero_goal/8	60/60	100.00	libero_spatial/8	59/60	98.33
libero_goal/9	55/60	91.67	libero_spatial/9	53/60	88.33

C UMAZE PROTOCOLS AND CONFIGURATION COMPARISONS

C.1 FAST64 UMAZE: ORIGINAL 500-ROOT PANEL AND FOUR-ARM COMPARISON

Fast64 denotes the main UMaze configuration using the original TD dataset. The comparison uses the same frozen offline IQL actor (S1, defined below) and normalization for Base, ordinary-rollout, CF, and anchor mean. Base and CF retain the original 1,000,000-row TD pool. Ordinary-rollout adds 153,600 transitions from 512 actor trajectories, for 1,153,600 rows, and trains critics from the common S1 initialization without an auxiliary loss. CF uses branch-return targets alongside the original TD data.

Actor and initialization. S1 denotes the seed-0 offline IQL actor–critic bundle trained for 100,000 updates on the one-million-transition Minari `D4RL/pointmaze/umaze-v2` dataset. The actor is frozen throughout all four arms. Its input is a seven-dimensional time-augmented state, and its two-dimensional action follows a squashed Gaussian with state-dependent log standard deviation bounded to $[-5, 2]$. Actor, twin Q and V use two width-256 ReLU hidden layers; Q and V have linear outputs. All arms initialize Q, V and target Q from this same bundle and use fresh Adam state for their 20,000-update continuation. S1 itself uses batch 256, learning rate 3×10^{-4}, $\gamma = 0.99$, expectile 0.7, and advantage-weighted actor regression with inverse temperature 3 and weights capped at 100.

Branch targets. The CF labels come from the frozen actor's branch pool. The collector recorded 512 parent trajectories and saved ten restorable states per parent at steps $0, 30, \ldots, 270$. For each failed parent it evaluated five candidates with eight continuations per candidate, then retained anchors with a candidate-return range of $G \geq 1$. This yielded 181 anchors from 144 roots. A CF update samples 16 retained anchors and regresses all five candidate returns with Huber loss of weight 0.1; branch transitions do not enter the TD pool. All four conditions use 20,000-update critics at seeds 11–13.

CF branch collection uses 21.78 million environment actions. The ordinary-rollout increment uses the 153,600 parent-trajectory actions described above.

The environment has a fixed goal, sparse continuing rewards, horizon 300, and success radius 0.45. Success does not terminate an episode. The score is an unclipped normalized undiscounted cumulative return, not an episode success rate:

$$\text{Score} = 100 \frac{\sum_{t=0}^{T-1} r_t - R^{\text{ref}}_{\min}}{R^{\text{ref}}_{\max} - R^{\text{ref}}_{\min}}. \tag{5}$$

The reference constants come from the original dataset metadata: $R^{\text{ref}}_{\min} = 13.49$ and $R^{\text{ref}}_{\max} = 218.70$ (displayed rounded). All score calculations use the unrounded stored constants.

Base and CF reuse their seed-matched checkpoints. Ordinary-rollout trains with fresh Adam state, drawing TD rows uniformly from the expanded pool and independently sampling the remaining horizon uniformly from 1 to 300. Its batch size is 256, $\gamma = 0.99$, expectile 0.7, Adam learning rate 10^{-4} with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, no weight decay, and target EMA rate 0.005. It has no auxiliary loss. Base and CF share paired TD-row and horizon draws within each seed. Ordinary-rollout uses the same sampling seed but an expanded pool, so its TD-row indices are not paired with Base.

Critic objective and evaluation checkpoint. For remaining horizon h, let s include the normalized time coordinate and let s' use $h - 1$. With $\bar{Q}$ the target network, training uses

$$\delta_V = \text{sg}[\min_j \bar{Q}_j(s, a)] - V(s), \qquad \mathcal{L}_V = \mathbb{E}[|0.7 - \mathbf{1}\{\delta_V < 0\}|\delta_V^2],$$

$$y = \text{sg}[r + 0.99\mathbf{1}\{h > 1\}V(s')], \qquad \mathcal{L}_Q = \mathbb{E}\left[\sum_{j=1}^{2}(Q_j(s, a) - y)^2\right],$$

$$\mathcal{L}_{\text{aux}} = \mathbb{E}_{u,k}\left[\tfrac{1}{2}\sum_{j=1}^{2}\text{Huber}_1(Q_j(s_u, a_{uk}) - z_{uk})\right].$$

Table 21: Fast64 UMaze four-arm mean normalized cumulative return on the original 500-root panel.

Arm	Mean score
Base	54.47
Ordinary-rollout	58.50
CF	59.67
Anchor mean	56.78

Table 22: Fast64 UMaze four-arm scores by training seed; each cell has 500 complete episodes.

Seed	Base	Ordinary-rollout	CF	Anchor mean
11	54.38	58.58	59.60	56.25
12	53.70	58.51	59.91	57.15
13	55.33	58.42	59.50	56.95

Each update optimizes V, then Q with $\mathcal{L}_Q + \lambda\mathcal{L}_{\text{aux}}$, then updates $\bar{Q} \leftarrow 0.995\bar{Q} + 0.005Q$. Here $\lambda = 0.1$ for CF and anchor mean and zero for Base and ordinary-rollout. The target z_{uk} is the mean discounted native return across eight frozen-K1 continuations, including the intervention reward. CF samples retained roots uniformly and then retained anchors within each root. The final 20,000-update online Q is evaluated; the EMA network supplies training targets only. No checkpoint selection or early stopping is used in these continuations.

Paired evaluation and uncertainty. Each model is evaluated on the same 500 reset roots, with three training seeds and 300 steps per episode. At every step the frozen actor proposes five candidates; the critic selects by minimum twin-Q, using the established uniform tie rule. Reset seeds are 24,001,000–24,001,499. The original three-arm evaluation contains 4,500 complete episodes and 1,350,000 environment steps. The primary comparisons are CF-minus-Base and CF-minus-ordinary-rollout. The seed-level interval uses a paired t distribution with two degrees of freedom; the 20,000-draw crossed bootstrap (seed 2026092501) resamples paired roots and training seeds, using identical draws across arms. For the two primary contrasts, both comparisons use the prespecified 97.50% interval level. A stable positive effect requires positive differences at all three seeds and positive lower bounds for both corrected intervals. This within-configuration correction does not adjust for the development choice among configurations.

Anchor-mean control. For the additional comparison in Figure 5, we replace the five candidate targets at each of the same 181 retained anchors by their arithmetic mean, without reapplying the filter. Candidate inputs, initialization, Base and auxiliary training draws, loss weight, 20,000 updates, and seeds 11–13 match the CF arm. The control is evaluated on the same 500 roots per seed; the original three arms are reused. The candidate namespace and random streams are retained, and the mean arm maps its tie-breaking stream to that of the same-seed CF arm. Its mean normalized cumulative return is 56.78, and CF exceeds it by 2.89 points (seed-t 95% CI $[1.86, 3.91]$; crossed-bootstrap 95% CI $[2.48, 3.36]$). These 95% intervals apply to the additional comparison; the original two primary contrasts retain their 97.50% intervals. Averaging changes target variance and gradients as well as within-anchor target differences; the contrast measures these changes jointly. The paired CF-minus-mean differences in normalized return are 3.35, 2.76, and 2.56 points for training seeds 11, 12, and 13, respectively. The control adds three new models and 1,500 complete episodes (450,000 steps), yielding 6,000 four-arm episodes on the same reset panel. The additional comparison reuses the original saved 20,000 paired seed/root bootstrap indices. Its primary endpoint is CF-minus-mean normalized return; this 95% family is separate from the original two-comparison 97.50% family. Mean-minus-Base is positive in mean but has a seed interval crossing zero; mean-minus-rollout is negative with both intervals below zero. Final-step success for CF minus mean is 3.73 pp, with seed-t 95% CI $[-7.22, 14.69]$ and crossed-bootstrap CI $[-1.53, 9.80]$; both intervals include zero.

Table 23: Fast64 UMaze paired differences in normalized return. Intervals are 97.50%, as prespecified for the two primary comparisons.

Contrast	Difference	Seed-t 97.50% CI	Crossed-bootstrap 97.50% CI
CF − Base	+5.20	[+1.54, +8.86]	[+4.05, +6.34]
CF − rollout	+1.17	[+0.45, +1.89]	[+0.79, +1.54]

Table 24: Fast64 additional mean-label contrasts in normalized return. Both intervals are 95%; only CF minus mean is the added primary comparison.

Contrast	Difference	Seed-t 95% CI	Crossed-bootstrap 95% CI
CF − mean	2.89	$[1.86, 3.91]$	$[2.48, 3.36]$
Mean − Base	2.31	$[-0.16, 4.79]$	$[1.43, 3.39]$
Mean − rollout	-1.72	$[-3.03, -0.41]$	$[-2.29, -1.24]$

C.2 UMaze configuration comparison and earlier PointMaze results

Fast64 is one member of a broader set of UMaze configurations. The following comparison records each setup's corresponding Base, ordinary-rollout, and CF conditions. The displayed configurations share the frozen S1 actor but differ in TD support, auxiliary targets, collection budget and evaluation roots; they are reported separately. Fast64 was selected for the primary presentation during development after observing comparative results. "UMaze (Sep. 23)" denotes the expanded-TD configuration in Table 3; Rollout25 denotes the quarter-rollout control and S80/F320 the success/failure mixture. Historical names are retained in the detailed source tables for traceability. Table 25 identifies the data and target changes.

Table 25: Definitions of the displayed UMaze configurations. TD counts are transitions; ordinary increments are environment actions. All use the frozen S1 actor, 20,000 critic updates per model and 500 evaluation roots. Budgets describe each setup's data support, including reused data, rather than new collection in every run.

Configuration	Base TD / ordinary increment	CF targets and pool	Seeds
Fast64	1,000,000 / 153,600	Absolute returns; 181 retained failed-parent anchors	11–13
UMaze (Sep. 23)	1,153,600 / 1,261,200	Same 181-anchor absolute-return pool	61–63
Rollout25	1,000,000 / 315,300 (R25), 1,261,200 (R)	Same absolute-return CF control; R25 uses 1,051 of 4,204 ordinary episodes	61–63
S80/F320	120,000 / 1,484,400	State-stratified pool from 80 successful and 320 failed parents; absolute and relative returns	61–65

S80/F320 uses 64 successful and 192 failed TD rows per batch. Its branch budget is 1,484,400 actions, matching the ordinary increment. Anchors are stratified by goal distance and speed before outcome filtering; the retained pool uses the same return-range threshold of one. Relative targets subtract the candidate mean at each anchor and center each Q head over its five candidates, with gradients through that Q mean. The other CF arms regress absolute returns. Rollout25 reuses a fixed identity-hashed quarter of the full ordinary episodes; its CF arm is a shared control, not a newly trained quarter-budget CF model.

In Table 27, B is Base, C is CF, R is ordinary-rollout, and R25 uses 25% of the ordinary-rollout increment. For S80/F320, A and L denote the absolute and relative CF targets. The September 23 rows retain the original three-comparison correction despite showing only two comparisons. Rollout25's two primary comparisons concern R25, whereas its CF contrasts are descriptive 95% intervals. S80/F320 retains all four primary contrasts and their 98.75% intervals. Bootstrap intervals cross seeds and paired roots. The three-seed September 23 configuration has near-zero CF-minus-Base gain and negative CF-minus-rollout estimates; its corrected seed-t intervals cross zero. S80/F320's corrected seed-t intervals also cross zero. Rollout25's negative CF-minus-R contrast is descriptive.

Table 26: UMaze configuration-specific three-arm comparisons. Results retain their own actor, data, training seeds, and evaluation panel.

Configuration	Arm	Mean return	Seeds	Roots	Episodes / arm
Fast64	Base	54.47	3	500	1,500
Fast64	Ordinary rollout	58.50	3	500	1,500
Fast64	CF	59.67	3	500	1,500
UMaze (Sep. 23)	Base	59.64	3	500	1,500
UMaze (Sep. 23)	Ordinary rollout	61.09	3	500	1,500
UMaze (Sep. 23)	CF	59.82	3	500	1,500
Rollout25	Base	55.38	3	500	1,500
Rollout25	Ordinary rollout (25%)	59.91	3	500	1,500
Rollout25	Ordinary rollout (100%)	61.08	3	500	1,500
Rollout25	CF	60.28	3	500	1,500
S80/F320-v2	Base	60.56	5	500	2,500
S80/F320-v2	Ordinary rollout	60.63	5	500	2,500
S80/F320-v2	CF absolute	60.29	5	500	2,500
S80/F320-v2	CF relative	60.53	5	500	2,500

Distinct evaluation panels and intervals; not pooled.

Table 27: Separate UMaze contrasts in normalized-return points. P denotes a prespecified comparison and D a descriptive comparison; confidence levels retain each source's original comparison family. Both interval types are shown.

Configuration	Contrast	Difference	CI% / role	Seed-t CI	Bootstrap CI
Fast64	C–B	+5.20	97.50 / P	[+1.54, +8.86]	[+4.05, +6.34]
Fast64	C–R	+1.17	97.50 / P	[+0.45, +1.89]	[+0.79, +1.54]
Sep. 23	C–B	+0.18	98.33 / P	[-0.53, +0.89]	[-0.22, +0.57]
Sep. 23	C–R	-1.27	98.33 / P	[-2.85, +0.31]	[-1.78, -0.87]
Rollout25	R25–R	-1.17	97.50 / P	[-3.35, +1.02]	[-1.75, -0.43]
Rollout25	R25–B	+4.53	97.50 / P	[-1.79, +10.86]	[+2.57, +6.19]
Rollout25	C–B	+4.90	95 / D	[+1.45, +8.35]	[+3.46, +6.10]
Rollout25	C–R25	+0.37	95 / D	[-0.63, +1.36]	[-0.10, +0.82]
Rollout25	C–R	-0.80	95 / D	[-1.35, -0.25]	[-1.07, -0.50]
S80/F320	A–B	-0.27	98.75 / P	[-0.69, +0.15]	[-0.54, +0.03]
S80/F320	A–R	-0.35	98.75 / P	[-0.86, +0.16]	[-0.67, -0.04]
S80/F320	L–B	-0.03	98.75 / P	[-0.17, +0.12]	[-0.17, +0.12]
S80/F320	L–R	-0.10	98.75 / P	[-0.63, +0.43]	[-0.41, +0.20]

C.3 EARLIER UMAZE FOUR-ARM STUDY

This earlier fixed-pool study compared Base IQL (B), filtered CF (C), unfiltered branch supervision (M), and uniform candidate selection (U). M uses unfiltered supervision from the same branch pool. Filtering at $G \geq 1$ changes both pool support and sampling weights. On this distinct panel, C scores 60.41 versus 55.44 for B, while C$-$M is +0.06 with intervals crossing zero. It is retained as a separate configuration.

Data, actor and critics. The environment is the locked Minari `D4RL/pointmaze/umaze-v2` evaluation specification, with fixed goal, sparse continuing rewards, horizon 300 and success radius 0.45. Success does not terminate the episode. The Base dataset contains one million transitions. Actor, initial Q/V/target-Q and normalization come from the same v2 baseline bundle, held fixed as the initialization for B/C/M. The actor is a time-augmented squashed Gaussian MLP; state dimension is seven, action dimension two, and its two hidden layers have width 256 and ReLU activations. Log standard deviation is state dependent and bounded to $[-5, 2]$. Q and V use two width-256 ReLU hidden layers and linear outputs trained on discounted native returns.

The historical branch pool has ten anchors per root, 330 roots and 3,300 anchors. Each candidate has eight discounted-return continuations with the frozen K1 actor. For anchor u, $G_u = \max_k z_{uk} - \min_k z_{uk}$ uses the float64 mean returns. The threshold $G_u \geq 1$ retains 181 anchors from 144 roots. C samples a retained root uniformly, then an anchor uniformly within that root, with probability $1/(144 n_u)$ for each anchor in a root containing n_u retained anchors. M samples one of 330 roots and one of its ten anchors, giving probability $1/3300$ per anchor. Common anchors have identical labels. Filtering and changed weighting are one joint intervention based on finite-sample return differences.

Table 28: UMaze v2 four-arm comparison. Score is normalized cumulative return, not success percentage. B/C/M have three training seeds; U is measured once per root. SD is across training seeds.

Arm	Auxiliary pool	Score	Seed SD
B	None	55.44	1.23
C	Filtered branch pool	60.41	0.29
M	Unfiltered branch pool	60.35	0.34
U	None (uniform selector)	44.31	—

Training. Seeds 61/62/63 each initialize B/C/M from identical baseline weights and fresh Adam state. There are 20,000 updates per model, Base batch 256, learning rate 10^{-4}, $\gamma = 0.99$, expectile 0.7, Adam $(0.9, 0.999)$ with $\epsilon = 10^{-8}$ and no weight decay. Each update optimizes V, then Q, then updates target Q with rate 0.005. The base Q loss sums the two head MSEs; the auxiliary loss averages the two head Huber losses with threshold one and weight 0.1. C/M each draw 16 anchors and all five candidates, giving 80 targets. Branch transitions do not enter the TD pool. Base row and remaining-horizon draws (uniform over 1–300) are identical across B/C/M within a seed; auxiliary RNG streams are separate. Only the final 20,000-update online Q is evaluated, without checkpoint selection, early stopping or retuning.

Evaluation. The panel contains 500 reset roots with seeds 9100000–9100499, checked for overlap with the historical root manifests. Each root evaluates U once and all nine trained critics. Paired resets, candidate RNG and tie RNG are shared; after state divergence the numerical candidates need not coincide. U selects a uniform candidate at each step, without a critic. Each episode has exactly 300 steps: 5,000 episodes and 1,500,000 evaluation steps in total.

The primary score and dataset reference constants are defined in Equation 5. Raw native return, discounted return, ever-success, final-success, first arrival and last-30-step occupancy are secondary. Non-arrivals are right-censored at 300 and included at 300 in the restricted mean first-arrival time.

Uncertainty and interpretation. We report individual training seeds, mean and sample SD. The primary comparisons are C−B and C−M, with M−B and B/C/M−U secondary. The 10,000-draw crossed seed×paired-root bootstrap uses seed 20260922; seed-t with two degrees of freedom is auxiliary. U is one shared reference with 500 roots. Its root-only normalized-score 95% interval is $[+43.55, +45.08]$. Differences against U retain its shared-root dependence. The positive-effect criterion requires all three seed estimates and both interval lower bounds to be positive. This exploratory study uses a fixed actor, map, and branch pool.

Table 29: PointMaze normalized-return contrasts. Both interval types are reported.

Contrast	Difference	Crossed bootstrap 95% CI	Seed-t 95% CI
C-B	4.97	[+3.70, +6.05]	[+1.90, +8.05]
C-M	0.06	[-0.49, +0.52]	[-1.23, +1.35]
M-B	4.91	[+3.23, +6.12]	[+1.04, +8.79]
B-U	11.12	[+10.04, +12.47]	[+8.06, +14.18]
C-U	16.10	[+15.45, +16.74]	[+15.38, +16.81]
M-U	16.03	[+15.39, +16.67]	[+15.19, +16.88]

Table 30: PointMaze individual training seeds and shared U reference.

Arm	Seed	Normalized score
B	61	56.86
B	62	54.70
B	63	54.75
C	61	60.43
C	62	60.11
C	63	60.68
M	61	59.97
M	62	60.64
M	63	60.43
U	—	44.31

Table 31: PointMaze auxiliary outcomes (mean and training-seed SD where applicable). Success and occupancy are fractions; return and arrival use their native units.

Arm	Metric	Mean	Seed SD
B	native_return	127.25	2.53
B	discounted_return	30.81	0.62
B	ever_success	1.00	0.00
B	final_success	0.50	0.02
B	first_arrival_step	60.48	0.16
B	last30_occupancy	0.50	0.01
C	native_return	137.45	0.59
C	discounted_return	33.23	0.07
C	ever_success	1.00	0.00
C	final_success	0.54	0.01
C	first_arrival_step	62.05	0.38
C	last30_occupancy	0.51	0.01
M	native_return	137.33	0.70
M	discounted_return	33.24	0.17
M	ever_success	1.00	0.00
M	final_success	0.55	0.01
M	first_arrival_step	61.66	0.32
M	last30_occupancy	0.52	0.01
U	native_return	104.42	—
U	discounted_return	25.93	—
U	ever_success	1.00	—
U	final_success	0.33	—
U	first_arrival_step	67.74	—
U	last30_occupancy	0.39	—

Acquisition cost. Collecting the branch pool required 21.78 million environment steps. The four-arm comparison reused that pool or the Base data, trained nine critics for 180,000 updates, and evaluated 1.50 million steps. Collection is counted here in environment steps; LIBERO uses action-chunk acquisition budgets.

D DIAGNOSTICS AND VISUAL EVIDENCE

D.1 EARLIER 240-STATE LOCAL G/U DIAGNOSTIC

This earlier diagnostic uses a different 240-state roster from the main 240-root AUC evaluation in Appendix B.4; the rosters do not overlap. It uses the same final Base and CF critics as the LIBERO matrix, with three training seeds for each algorithm. A fixed roster contains six roots per task across 40 tasks. For each root, the frozen actor generates a K1 parent trajectory. Uniform reservoir sampling selects one nonterminal H5 decision with at least five actions remaining, without selecting on candidate outcomes. The resulting 240 states each receive five H10 proposals. Each candidate's first five actions are executed, followed by the frozen K1 actor to success or the suite horizon, with six continuation repeats. Prefix replay restores the same pre-intervention state for each branch. These are local, reference-policy outcomes.

For anchor u, candidate k, and repeat m, let $Y_{ukm} \in \{0,1\}$, $p_{uk} = \frac{1}{6}\sum_{m=1}^{6} Y_{ukm}$, and $q_{uk} = \min_j Q_j(s_u, a_{uk})$. Define

$$G_u(Q) = \frac{100}{\binom{5}{2}} \sum_{i<j} \operatorname{sign}(q_{ui} - q_{uj})(p_{ui} - p_{uj}),$$
$$U_u(Q) = 100 p_{u,k^*}, \qquad k^* = \arg\max_k q_{uk}.$$

Score ties contribute zero to G_u; U_u breaks ties by the first original candidate index. Equal observed success fractions contribute zero regardless of score order. Unlike the return range used to filter UMaze anchors, $G_u(Q)$ here is a signed preference diagnostic, in percentage points; U_u is a success percentage. Differences in either metric are in percentage points.

We average with equal weights over suites, tasks within a suite, roots within a task, and anchors within a root, then over the three trained seeds. ΔG and ΔU subtract Base from CF using the same states and labels. The seed-t interval uses paired seed differences with two degrees of freedom. The 20,000-draw bootstrap (seed 20260924) resamples tasks and then whole roots within each fixed suite, carrying all anchors, methods, seeds, candidates and repeats together. Its trained seeds stay fixed. The two algorithm-specific ΔG comparisons use 97.50% intervals; ΔU is descriptive.

Table 32: Local candidate-scoring results for CF minus Base. Both types of 97.50% intervals apply to the pairwise preference diagnostic. ΔU is descriptive.

Critic	ΔG (pp)	Seed-t CI	Bootstrap CI	ΔU (pp)
IQL	+0.23	[-0.89, +1.35]	[-0.79, +1.28]	+0.02
CQL	+0.19	[-0.01, +0.39]	[-0.64, +1.04]	+0.32

All reported intervals for ΔG include zero. CQL uses $\alpha = 0.1$. Of the 240 states, 153 have no observed candidate-outcome contrast, and 78.42% of candidate pairs have equal observed success fractions. Together with six continuations per candidate and three training seeds, this limits sensitivity to small effects.

D.2 CONTROLLED-MDP MEASUREMENTS

Dynamics and scoring for the separation measurement. The state is $x_h = (p_h, d_h)$ with $x_0 = (0,0)$, horizon $H = 10$ and $K = 8$. At each decision, draw independent candidates $a_{hk} \sim \mathcal{N}(0, I_2)$ and score noise $\eta_{hk} \sim \mathcal{N}(0,1)$. For $\psi = (1,0)$ or $(-1,0)$, score each candidate by

$$s_{hk} = p_h + 0.5\psi^\top a_{hk} + \eta_{hk}.$$

After selecting k_h, draw $w_h \sim \mathcal{N}(0, 0.3^2 I_2)$ and update

$$p_{h+1} = p_h + 0.05 + a_{hk_h,1} + w_{h,1}, \qquad d_{h+1} = d_h + a_{hk_h,2} + w_{h,2}.$$

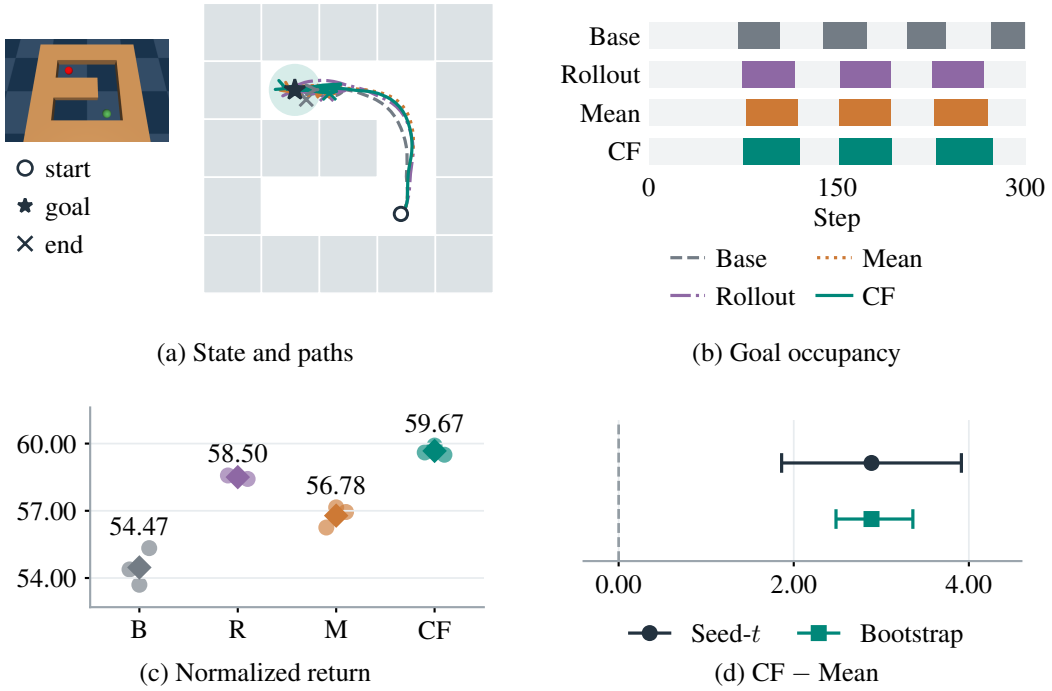
Success is $Y = \mathbf{1}\{p_H \geq \theta\}$, with $\theta = 0.05H + 0.8\sqrt{H}\,0.3$; no cliff constraint is active. Argmax selects the largest noisy score, while uniform samples one of the eight indices independently of the scores. The dynamics and all parameters are fixed between the aligned and anti-aligned scorer conditions.

Prediction metric and random streams. For a uniform-reference trajectory, compute $\bar{s} = H^{-1}\sum_h K^{-1}\sum_k s_{hk}$. The reported AUC ranks successful versus failed trajectories using $\bar{s}$, not a raw state–action Q score or a selected-action score. The implementation uses the Mann–Whitney rank statistic; the continuous score noise makes exact ties negligible. Each condition has $N = 20{,}000$ episodes per seed and 24 seeds $s_i = 7 + 1000i$, $i = 0, \ldots, 23$. For each scorer, the closed-loop difference subtracts the uniform mean at seed $s_i + 1$ from the argmax mean at seed s_i. Prediction AUC uses uniform rollouts at seed s_i. These are seed-level differences of sample means, not trajectory-paired treatment effects. The saved aggregate interval is the seed mean plus or minus 1.96 times the sample standard error. The observed AUCs are 0.91 and 0.90; the selection effects are +55.74 and -39.65 pp.

Supplementary score-only probe. Uniform logs with directions $(1, 0)$ and $(-1, 0)$ use seeds s_i and $s_i + 7919$. Features are the trajectory averages of each candidate set's mean, maximum and standard deviation. The $2N$ examples are shuffled with seed s_i and split equally for fitting and evaluation. An intercept-augmented least-squares classifier predicts the scorer direction; its held-out AUC is averaged across the same 24 seeds. This probe uses only the score statistics defined above.

D.3 Visual evidence and deterministic example selection

Figures 2 and 5 show recorded observations and trajectories. We render the maze layout, initial pose, and goal with the simulator's CPU renderer. Robot frames come from the camera buffers, flipped vertically to correct the simulator's bottom-up order.



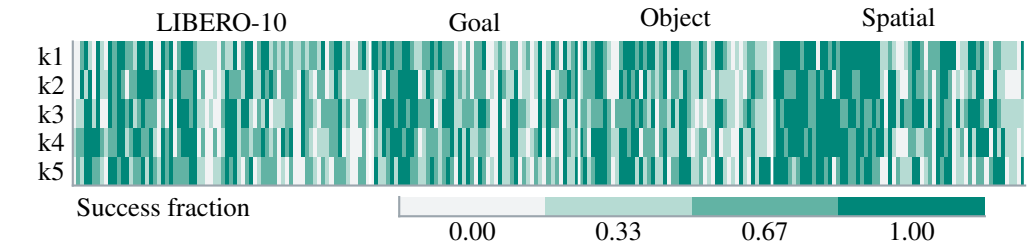
(a) State and paths

(b) Goal occupancy

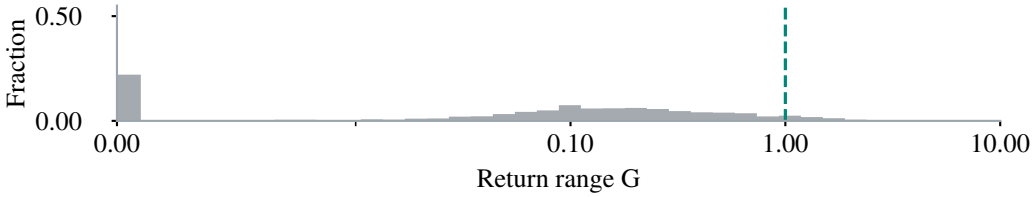
(c) Normalized return

(d) CF − Mean

Figure 5: Fast64 UMaze with an anchor-mean control. (a) Recorded initial state and paired trajectories, selected by the four-arm median-return rule. Circles, stars, and crosses mark the start, goal, and endpoints; the shaded disk is the goal region. Gray dashed, purple dash-dotted, orange dotted, and teal solid paths denote Base, Rollout, Mean, and CF, respectively. (b) Goal-occupancy bands: colored segments indicate steps inside the goal region; light-gray segments indicate steps outside it. (c) Normalized cumulative return over 500 shared roots per model; dots show all three training seeds, while diamonds and numerical labels show means. B, R, and M denote Base, ordinary-rollout, and action-conditioned anchor mean. (d) CF-minus-anchor-mean return difference (2.89 points), with circular seed-t and square crossed seed/root bootstrap 95% intervals.



(a) LIBERO candidate success fractions



(b) UMaze candidate-return range

Figure 6: The supervision pools behind the examples. (a) All CF237 candidate success fractions, sorted by suite and then anchor ID, without sorting candidates by outcome. All LIBERO anchors were outcome-selected. (b) Complete candidate mean-return range distribution for the historical UMaze pool. The dashed threshold is $G = 1$, retaining 181 of 3,300 anchors. The horizontal scale is linear near zero and logarithmic above 0.01; all anchors, including exact zeros, contribute. Fractions denote mass per bin, not probability density. Filtering changes support and sampling weights jointly.

LIBERO example. We rank anchors by absolute distance to the median candidate-label range in CF237, using integer success-count ranges to preserve exact ties at the common $M = 3$, breaking ties by anchor ID and requiring complete images and both observed outcomes. The image-complete subset contains 93 anchors; the earlier 144 contribute to the full heatmap but are not eligible for this camera example. The selected anchor is LIBERO-10 task 30 (global task index), initial configuration 21, frame 105, with the instruction "put both the alphabet soup and the tomato sauce in the basket." Its five success counts are $(0, 0, 0, 1, 0)$ out of three; its anchor mean is $1/15$. The endpoints are candidate 4/repeat 1 (success, 338 branch steps) and candidate 1/repeat 1 (failure, 415 steps), using one-based indices. For each outcome we take the first available candidate/repeat in index order.

Fast64 anchor-mean example. Figure 5 shows the control from Appendix C.1. We fix seed 12 and compute each arm's median native return over all 500 shared initial states. We select the state minimizing the sum of absolute deviations from the four arm-wise medians, breaking ties by root ID. The selected zero-based root ordinal is 41. In Base, ordinary-rollout, anchor-mean, and CF order, the native returns are $(126, 123, 125, 132)$ and the arm-wise medians are $(127, 126.50, 125, 132)$. Each path includes the initial position and all 300 recorded steps; the time bands use the corresponding goal-occupancy flags, checked against the radius 0.45 and step rewards. This example illustrates deployment behavior; the aggregate results use all evaluation roots.

Maze collection example. The branch illustration selects the UMaze retained anchor nearest the median G, breaking ties by root and anchor ID: root 39 at time 30, $G = 1.35$. For each candidate, its displayed continuation minimizes absolute distance to the mean of that candidate's eight recorded returns, with repeat-index ties. The five solid arrows encode recorded action vectors at a common display scale; the fine curves are the resulting recorded state trajectories, not action chunks.

E SUPPLEMENTARY LOGGING ANALYSIS AND HISTORICAL SCORING

E.1 OBSERVED-FORK SIBLING-RANKING MEASUREMENTS

These scorers and their actor belong to a separate historical experiment. This historical exam uses a separate $\pi_{0.5}$ actor with one deterministic continuation per candidate. The reset exam evaluates the native progress score (FPF) and two separately trained expectile-Q scorers, denoted IQL-offline and IQL-RR. Each scorer receives the same stored observations and candidate actions. A pivotal anchor has both a successful and a failed realized continuation. At each such anchor, sibling AUC is the fraction of successful–failed candidate pairs correctly ordered by the score, with half credit for score ties. Suite AUC is the unweighted mean over pivotal anchors. Random and observed-fork oracle success average, over all anchors, the mean and maximum of the five realized candidate outcomes, respectively.

Table 33: Sibling-ranking AUC on the fixed, single-continuation reset exam. Anchor counts give the complete roster and the pivotal subset used for AUC.

Suite	All anchors	Pivotal	FPF	IQL-offline	IQL-RR
`spatial`	868	335	0.49	0.50	0.50
`object`	926	290	0.52	0.51	0.48
`goal`	726	271	0.50	0.50	0.48
`libero_10`	1058	432	0.52	0.52	0.49

E.2 RANDOMIZED INDEX LOGGING AND ESTIMATION COST

Let $\mathcal{H}_h$ be the recorded pre-choice history, including the current candidate set and scores. Logging and target rules use the same initial law, candidate-generation kernels, and environment/outcome kernels; only their index probabilities differ. Logging randomization is independent of future environment randomness conditional on $\mathcal{H}_h$. Assume trajectory absolute continuity, implemented by $\sigma_b(k \mid \mathcal{H}_h) > 0$ whenever $\sigma'(k \mid \mathcal{H}_h) > 0$ on target-reachable histories. Define

$$R_h = \frac{\sigma'(k_h \mid \mathcal{H}_h)}{\sigma_b(k_h \mid \mathcal{H}_h)}, \qquad W = \prod_{h=1}^{H} R_h, \qquad \widehat{J}_{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} W_i Y_i.$$

Proposition 4 (Sequential importance weighting)**.** *For independent logged episodes and $0 \le Y \le 1$, $J(\sigma') = \mathbb{E}_b[WY]$ and $\mathbb{E}_b[W] = 1$. Ordinary IS is unbiased and strongly consistent. Self-normalized IS, $\sum_i W_i Y_i / \sum_i W_i$, is strongly consistent whenever the denominator is nonzero; any fixed convention for zero-denominator finite samples has the same limit. If $\mathbb{E}_b[W^2] < \infty$, ordinary IS has MSE at most $\mathbb{E}_b[W^2]/n$.*

Proof. Factor a trajectory law into the initial distribution and, at each decision, the candidate kernel, index probability, and environment kernel, followed by the outcome kernel. Shared factors cancel in the Radon–Nikodym derivative, leaving $dP_{\sigma'}/dP_b = W$. Integrating Y and 1 proves both identities. Bounded Y implies integrability of WY since $\mathbb{E}_b W = 1$, so the strong law proves consistency of its sample mean. The sample mean of W converges to 1, which also proves consistency of the ratio and eventual positivity of its denominator. Finally, independence gives $\mathbb{E}[(\widehat{J}_{\mathrm{IS}} - J)^2] = \mathrm{Var}_b(WY)/n \le \mathbb{E}_b[W^2]/n$. These are the standard change-of-measure identities used in counterfactual learning and sequential off-policy evaluation (Bottou et al., 2013; Jiang & Li, 2016). □

Proposition 5 (A bound in terms of differing decisions)**.** *Suppose the rules agree outside a fixed set of D decision times. At each of those times, assume $\sigma_b(k \mid \mathcal{H}_h) \ge p > 0$ for every target-supported index, uniformly over histories. Then*

$$\mathbb{E}_b[W^2] \le p^{-D}, \qquad \mathbb{E}[(\widehat{J}_{\mathrm{IS}} - J)^2] \le \frac{p^{-D}}{n}.$$

For a softmax logger with temperature $\tau > 0$ and a uniform bound R on the within-set score range, one may take $p = [1 + (K-1)e^{R/\tau}]^{-1}$.

Proof. At a differing time, condition on $\mathcal{H}_h$ to obtain

$$\mathbb{E}_b[R_h^2 \mid \mathcal{H}_h] = \sum_k \frac{\sigma'(k \mid \mathcal{H}_h)^2}{\sigma_b(k \mid \mathcal{H}_h)} \le p^{-1} \sum_k \sigma'(k \mid \mathcal{H}_h)^2 \le p^{-1}.$$

At an agreeing time the factor is 1. Applying the tower property backwards to $\mathbb{E}_b[\prod_h R_h^2]$ gives one factor p^{-1} per differing time. The MSE bound follows from Proposition 4. For softmax, the smallest probability is bounded below by $e^{s_{\min}/\tau}/[e^{s_{\min}/\tau} + (K-1)e^{(s_{\min}+R)/\tau}]$. Thus the bound depends on the number of differing factors when the stated p remains fixed. The Gaussian-score experiments below report empirical ESS and a fitted decay factor for their realized weight distributions. □

Proposition 6 (A rare-path lower bound)**.** *Fix $1 \le D \le H$ and $0 < p \le 1/2$. There is a family of finite-horizon, finite-action models $\{P_\theta : 0 \le \theta \le 1\}$ for which the target follows a specified branch at D decisions, the logger chooses it with independent probability p at each decision, and the target value is $J_\theta = \theta$. For $n \ge 1$ independent logged episodes,*

$$\inf_{\widehat{J}} \sup_{\theta \in [0,1]} \mathbb{E}_\theta[(\widehat{J} - \theta)^2] \ge \frac{1}{512} \min\{1, (np^D)^{-1}\}.$$

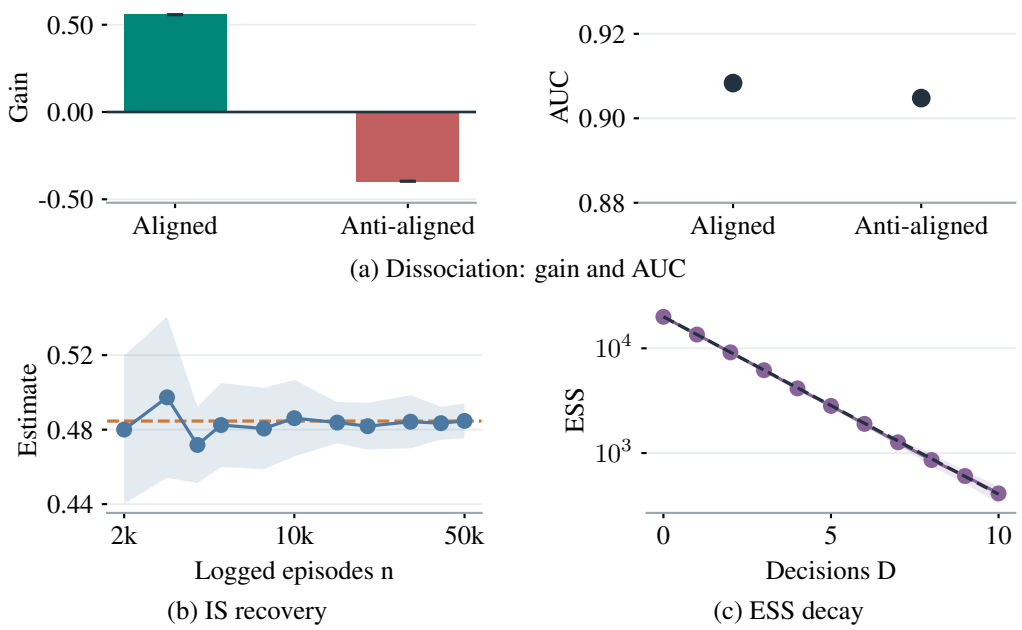
Within this family an estimator achieves MSE at most $\min\{1/4, (np^D)^{-1}\}$.

Proof. Use a state flag recording whether all specified branches so far have been taken. It is updated deterministically by the executed action. All transitions, observations, and action probabilities are independent of θ. At the terminal time draw $Y \sim \mathrm{Bernoulli}(\theta)$ if every specified branch was taken (event E), and $Y \sim \mathrm{Bernoulli}(1/2)$ otherwise. The target always takes those branches, hence $J_\theta = \theta$. Under the logger $q = \Pr_b(E) = p^D$. Extra steps use identical rules and θ-independent transitions. The complete logged path, including the flag, therefore contains information about θ only through Y on E.

Set $\epsilon = \min\{1/4, (128nq)^{-1/2}\}$ and $\theta_\pm = 1/2 \pm \epsilon$. The KL divergence between one-episode laws is $q\,\mathrm{kl}(\theta_- \| \theta_+)$. The Bernoulli bound $\mathrm{kl}(a\|b) \le (a-b)^2/[b(1-b)]$ gives $\mathrm{kl}(\theta_- \| \theta_+) \le (64/3)\epsilon^2$, so the n-episode KL divergence is at most $1/6$. Pinsker's inequality bounds total variation by $\sqrt{1/12} < 1/2$. Any estimator defines a test by thresholding at $1/2$; a wrong decision incurs squared error at least ϵ^2. The average risk over the two hypotheses is therefore at least $\epsilon^2(1 - \mathrm{TV})/2 \ge \epsilon^2/4$. Finally, $\epsilon^2 = \min\{1/16, (128nq)^{-1}\} \ge (1/128)\min\{1, (nq)^{-1}\}$, proving the lower bound.

For the upper bound, $W = \mathbf{1}_E/q$ and $\mathbb{E}_b[W^2] = 1/q$, so ordinary IS has MSE at most $1/(nq)$. The constant estimator $1/2$ has worst-case MSE $1/4$. Choosing between these estimators using known n, q establishes the stated upper bound. This family illustrates the cost of rare informative paths; trajectory importance-sampling variance is studied by Liu et al. (2020). □

Additional logging measurements. Figure 7(a) supports the motivating construction (Appendix D.2). Panels (b–c) concern the separate problem of evaluating selectors from randomized logs. Panel (b)'s protocol and statistics follow below. Panel (c) varies the number D of decisions at which the target selector differs from the logger, at fixed horizon $H = 10$ and softmax temperature 0.3. With 20,000 episodes per condition and 12 seeds, a log-linear fit of mean effective sample size against D gives $r^2 = 1.00$ and $\Lambda = 1.48$. This is an estimation-cost measurement under randomized logging, separate from the critic-training comparisons.



(a) Dissociation: gain and AUC

(b) IS recovery

(c) ESS decay

Figure 7: Controlled-MDP validation with panel-specific configurations. (a) Two constructed scorers have similar trajectory-score AUCs under uniform selection (right, dots) but opposite success changes under argmax selection relative to uniform selection (left, bars). AUC uses candidate-mean scores averaged over a reference trajectory. Means are across 24 seeds; error bars on the success changes show normal-approximation 95% intervals. (b) Self-normalized importance sampling with the higher terminal threshold $\theta = 7$ (direct success 0.48), at 11 logged-episode counts; points show the mean and shading shows ± 1 sample standard deviation across 24 seeds. The dashed line is the independent direct-rollout reference; connecting lines are visual guides. (c) Effective sample size exhibits approximately exponential decay with the number of diverging decisions D. Panel (c) uses 12 seeds; its band is the 10th–90th seed percentile and its dashed curve uses the fitted decay factor $\Lambda = 1.48$. Panels (a) and (c) retain the original configuration. Panel (b) statistics and protocol are in Appendix E.3.

E.3 Higher-threshold importance-sampling recovery

Figure 7(b) uses $H = 10$, $K = 8$, softmax logging temperature 0.2, and the aligned score direction $(1, 0)$. Only the terminal success threshold is raised to $\theta = 7$; the additional cliff constraint is disabled. This threshold was fixed before the run, near the analytically predicted target mean final progress of 6.87, to avoid near-ceiling success. Transition dynamics and candidate scoring are held fixed. Each of the 11 sample sizes uses all 24 seeds $7 + 1000i$, $i = 0, \ldots, 23$; sample-size conditions sharing this roster are not treated as independent replications. An independent 200,000-episode argmax rollout with seed 900001 gives the Monte Carlo reference 0.48. Logging and target rollouts are newly generated for this configuration.

The sequential weight is $W = \prod_{h=1}^{10} \frac{\mathbf{1}\{k_h = \arg\max_k s_{hk}\}}{\sigma_b(k_h|h)}$. We retain both ordinary IS, $n^{-1} \sum_i W_i Y_i$, and self-normalized IS, $\sum_i W_i Y_i / \sum_i W_i$. Shading in panel (b) is the sample standard deviation across seeds. At $n = 2000$, individual self-normalized estimates range from 0.40 to 0.55, despite their mean being close to the reference. All seeds and sample sizes are retained without smoothing or post-hoc selection.

Table 34: Higher-threshold panel (b): estimates and errors in probability units. RMSE is across 24 seeds against the independent direct-rollout reference; ESS is the seed mean.

n	SN mean	SN SD	SN RMSE	Ordinary RMSE	ESS
2,000	0.48	0.04	0.04	0.05	154.00
3,000	0.50	0.04	0.04	0.05	229.20
4,000	0.47	0.02	0.02	0.03	294.70
5,000	0.48	0.02	0.02	0.04	372.50
7,500	0.48	0.02	0.02	0.02	567.50
10,000	0.49	0.02	0.02	0.03	753.00
15,000	0.48	0.01	0.01	0.02	1139.40
20,000	0.48	0.01	0.01	0.02	1486.20
30,000	0.48	0.01	0.01	0.02	2232.40
40,000	0.48	0.01	0.01	0.01	2944.70
50,000	0.48	0.01	0.01	0.01	3664.80