# DECISION RETENTION IS NOT TRAJECTORY RETENTION: PRESERVING CROSS-HERO CAPABILITIES DURING OFFLINE-TO-ONLINE GAME REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Offline imitation can produce behavior that transfers beyond the demonstrated task, while narrow online reinforcement learning may erase those untrained capabilities. We study this problem in the licensed Honor of Kings Gamecore using a corpus of 1.78 million decisions from one controlled hero. Across six training seeds and 1,422 Gamecore matches, twenty source-only PPO updates reduce cross-hero scheduled win rate from 38.9% to 29.2% (four of six seeds negative) without a reliable source-reward gain. Source-policy distillation restores cross-hero reward and win rate to near their pre-adaptation levels. We introduce capability-support distillation (CSD), which stores reward-free states and legal-action constraints from sentinel heroes and distills the frozen pre-adaptation policy during online learning. On 39,094 sentinel states, CSD raises teacher agreement from 75.78% under source distillation to 79.35%. In a one-time held-out evaluation released by a hashed decision file fixed before any held-out CSD game, CSD improves over unconstrained PPO by 14.59 reward points [3.21, 26.47] and 20.83 percentage points [8.33, 33.33]; a second independent held-out run shows the same direction with smaller gaps, and pooled over both runs the advantage is 10.63 reward points [2.68, 19.82] and 13.2 percentage points [5.6, 21.5]. Pooled, CSD matches but does not exceed the pre-adaptation policy, and its reward differences from behavioral cloning, hard replay, and source distillation include zero. On sentinel heroes, where the five methods win one of 270 games, CSD agrees with the teacher most closely yet loses fewer games and reaches the step cap more often than the teacher. Decision retention on stored states is therefore measurable but is not by itself evidence of trajectory retention.

## 1 INTRODUCTION

Game agents are often trained in two stages. First, imitation or offline reinforcement learning turns a static corpus into a competent policy. Second, online reinforcement learning specializes that policy through interaction. The update distribution in the second stage can be much narrower than the capabilities acquired during pretraining. In a multi-hero game, a policy adapted online with one hero can preserve, improve, or erase behavior that was useful when controlling other heroes.

We measure this change in Honor of Kings Arena (Wei et al., 2022), using the HoK-Off corpus (Qu et al., 2023) and terminal matches in the licensed Gamecore. We distinguish source-task adaptation from cross-task retention, use the training seed as the unit of inference, and report terminal outcomes beside shaped return. Our first experiment shows that one-hero behavioral cloning produces measurable cross-hero behavior, but narrow PPO (Schulman et al., 2017) reduces cross-hero task completion without a reliable source-task gain. Hard source replay reduces the damage; Figure 1 previews the main results.

We then ask what a small memory must contain. Our proposed capability-support distillation stores observations and legal-action constraints from non-source sentinel heroes, but no reward, outcome, or expert-action label. A frozen pre-adaptation policy supplies soft targets during source-hero PPO.



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.

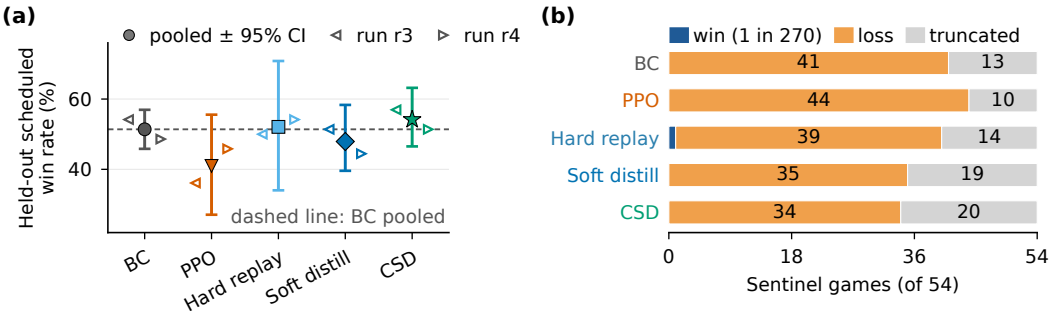


Figure 1: The two halves of the finding. (a) Held-out scheduled win rate on four controlled heroes never used for training or selection: source-only PPO falls below the pre-adaptation BC policy, and capability-support distillation (CSD) returns to BC's level without exceeding it (pooled over two independent evaluation runs; bars are 95% seed-bootstrap intervals over six training seeds; Table 3). (b) On the sentinel heroes whose states CSD stores, CSD agrees with the BC teacher most closely (Table 2), yet its games end differently from the teacher's: fewer losses and more truncations at the step cap (one hard-replay win in 270 games).

No CSD state, trajectory, or result from the held-out heroes is observed during method selection; the frozen CSD checkpoints are evaluated on them once after the release gate is passed.

Our contributions are:

- a source–sentinel–held-out protocol that separates source adaptation, historical-support behavior, and cross-task outcomes in a commercial-game research engine;
- a six-seed, 1,422-match measurement showing that narrow source-hero PPO reduces cross-hero completion without a reliable source-task gain, together with matched hard- and soft-replay controls; and
- capability-support distillation, which uses reward-free sentinel states and legal-action structure, and a one-time held-out evaluation, released by a hashed pre-committed decision, showing improvement over unconstrained PPO while exposing the gap between fixed-state agreement and transferable trajectory retention.

## 2 Related Work

### 2.1 Offline-to-Online Reinforcement Learning

WSRL (Zhou et al., 2025) uses a small set of warm-up rollouts from the pretrained policy to recalibrate a value function before online fine-tuning. Earlier, Lee et al. (2021) observed that state-action distribution shift during fine-tuning can cause bootstrap error severe enough to destroy the offline-trained policy, and countered it with a balanced replay of online and near-on-policy offline samples; RLPD (Ball et al., 2023) showed that off-policy methods with a few design changes can use offline data directly during online learning. Recent work frames offline-to-online learning through stability and plasticity (Anonymous, 2026). These methods measure success on the task that supplies online reward. Our setting preserves behavior on tasks that provide neither online trajectories nor online rewards, so final source return alone is insufficient.

### 2.2 Continual Reinforcement Learning

Policy distillation was introduced as a way to compress a policy or consolidate several task policies into one network (Rusu et al., 2016). In continual RL, CLEAR mixes on-policy updates with off-policy replay and behavioral cloning, reducing forgetting without requiring task identities (Rolnick et al., 2019). Policy Consolidation instead distills behavior across a cascade of policy timescales and evaluates continuous distribution changes and competitive self-play (Kaplanis et al., 2019). DisCoRL (Traoré et al., 2019) combines state representation learning and distillation across sequential navigation tasks. RECALL (Zhang et al., 2023) applies policy distillation on replayed old-task

states and shows that naive replay can fail even with perfect memory. Parameter-space alternatives such as EWC (Kirkpatrick et al., 2017) slow learning on weights important for earlier tasks instead of constraining outputs. Teacher distillation also accelerates RL, as in kickstarting (Schmitt et al., 2018), and small variations among distillation objectives change what is optimized (Czarnecki et al., 2019). Closest to our question, Wołczyk et al. (2024) identify forgetting of pre-trained capabilities during RL fine-tuning: the model deteriorates on parts of the downstream state space it does not visit early in fine-tuning, and standard knowledge-retention techniques mitigate this in NetHack and Montezuma's Revenge. For fine-tuned language and robot policies, forgetting tracks the KL to the base policy on the new task (Shenfeld et al., 2025), and Luo et al. (2026) argue that the KL should instead be measured on prior-task behavior; CSD measures it on stored prior-task states, and we find that lowering it there need not preserve deployed outcomes. We study the cross-task analogue, in which the forgotten behavior belongs to controlled heroes that never supply online reward, and we ask whether retention measured on stored states carries over to deployment. Our contribution is therefore not KL regularization or a frozen teacher. It is the construction and leakage-resistant evaluation of a reward-free capability support for emergent transfer in a complex game.

### 2.3 Learning Agents for Honor of Kings

Honor of Kings has served as a testbed for large-scale RL: Ye et al. (2020) train agents for a large hero pool with curriculum self-play and policy distillation among other techniques, and Honor of Kings Arena (Wei et al., 2022) was introduced to study generalization across heroes and opponents in the 1v1 game. We use the same engine at a much smaller scale to isolate what one-hero online adaptation does to behavior on other heroes.

## 3 Problem Formulation

Let $\mathcal{M}_0$ denote the source task and $\{\mathcal{M}_i\}$ the tasks produced by changing controlled and opponent heroes while keeping the structured interface fixed. Behavioral cloning produces $\pi_0$ from source-only offline data. Online adaptation produces $\pi_\theta$ using trajectories and rewards exclusively from $\mathcal{M}_0$. We ask whether adaptation changes source performance and whether it preserves behavior that $\pi_0$ exhibited on non-source tasks.

We measure *decision retention* on a frozen state support $\mathcal{D}_{\mathrm{cap}}$ using legal-head teacher agreement and $D_{\mathrm{KL}}(\pi_0 \| \pi_\theta)$. We measure *trajectory retention* by deploying $\pi_\theta$ and recording shaped return, scheduled win rate, and terminal rate. These quantities need not agree because each deployed action changes the next-state distribution. This is the sequential distribution-shift problem underlying compounding error in imitation learning (Ross & Bagnell, 2010; Ross et al., 2011), here applied to retention after online updates rather than to initial imitation. Sentinel tasks construct $\mathcal{D}_{\mathrm{cap}}$ and select the method; disjoint held-out tasks test whether retained behavior transfers beyond it.

The observation, action and legality interface, the reward vector and its verified coefficients, the match-outcome rule, and the offline corpus and its split are specified in Appendix A.

## 4 Methods

### 4.1 Behavioral Cloning and PPO

The compact recurrent actor–critic uses a 725-to-128 tanh encoder, a 128-unit LSTM cell, six linear policy heads, and a scalar value head. Behavioral cloning runs for 5,000 AdamW updates on episode-safe sequences of length 32, of which the first 16 steps are burn-in and the remaining 16 are supervised. Online adaptation starts from each seed's behavioral-cloning checkpoint and resets the optimizer. All methods receive 20 PPO updates (unconstrained: PPO20), at most 3,000 source-hero decisions per update, and the same eight-opponent cycle; every update in every run ended at a terminal state, so one update is one full source episode. PPO uses two optimization epochs per update, discount $\gamma = 0.995$, GAE $\lambda = 0.95$, clip ratio 0.2, value coefficient 0.5, entropy coefficient 0.01, Adam with learning rate $10^{-4}$, gradient-norm clipping at 0.5, and per-update advantage normalization; log-probabilities and entropies are computed only over active heads and legal actions.

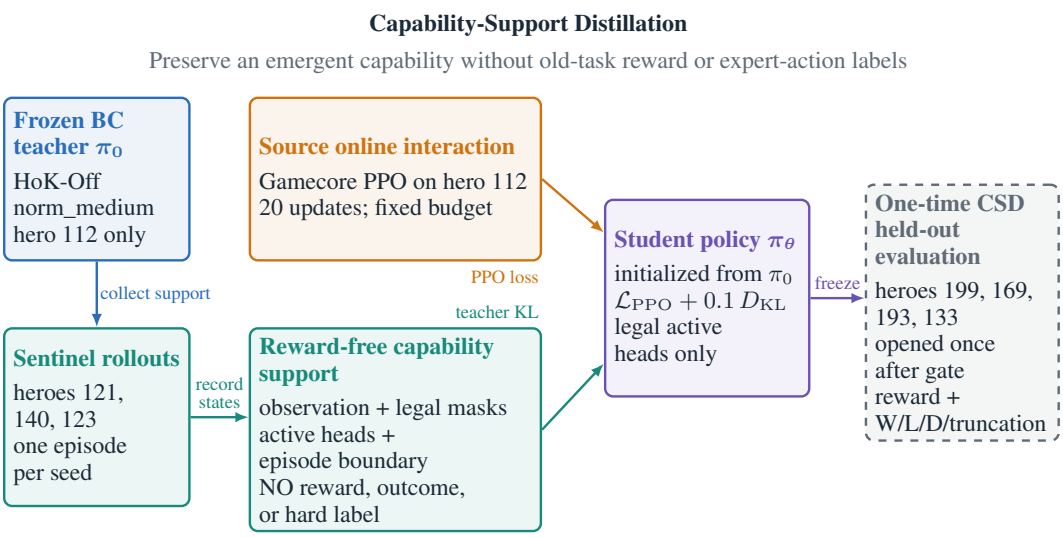


Figure 2: Capability-support distillation. Sentinel trajectories contribute reward-free state and legality support; the frozen behavioral-cloning policy supplies soft targets while PPO receives reward only from source hero 112. CSD trajectories and results on held-out heroes remain sealed until the selection gate is frozen and passed.

### 4.2 Source Replay Baselines

Hard source replay applies the recorded categorical labels from 64 source episodes. Soft source distillation replaces each hard label with the frozen behavioral-cloning distribution. Both use batches of eight sequences of length 32, 16 burn-in steps followed by 16 loss-bearing steps, one replay batch per optimization epoch, and coefficient 0.1. The replay sampler is seeded identically across methods, so hard replay, soft distillation, and capability-support distillation draw the same sequence positions within a training seed; the comparison between them is paired at both the batch and the seed level.

### 4.3 Capability-Support Distillation

Figure 2 summarizes the training and evaluation boundary.

For each training seed, its frozen behavioral-cloning policy plays one scheduled game in each of three sentinel matchups under the same 3,000-decision cap as evaluation; in four of the six seeds one of the three collection games reached the cap without a terminal state, and its states are kept. The memory stores observations, legal-action masks, active-head masks, episode boundaries, hero metadata, and the teacher button needed to select the button-conditioned target legality mask. It stores no reward, outcome, or expert-action label. The retained button is the teacher's greedy (argmax) button at collection time; it is used solely to reconstruct the button-conditioned active-head and target-legality context and is never optimized as a hard target. Teacher and student consume the recurrent burn-in prefix. For state $s$, the auxiliary loss is

$$\mathcal{L}_{\mathrm{cap}}(s) = \sum_{h \in \mathcal{H}(s)} D_{\mathrm{KL}}(\pi_{0,h}(\cdot \mid s) \,\|\, \pi_{\theta,h}(\cdot \mid s)), \tag{1}$$

where illegal choices are masked, $\pi_0$ is frozen, and $\pi_\theta$ is updated by $\mathcal{L}_{\mathrm{PPO}} + 0.1\mathcal{L}_{\mathrm{cap}}$.

## 5 Experimental Protocol

Training seeds are 11, 23, 47, 71, 89, and 101. Hero 112 is the source; heroes 121, 140, and 123 are sentinels; heroes 199, 169, 193, and 133 are held out. Sentinel states may be recorded without rewards or action labels. No held-out CSD states, trajectories, outcomes, or diagnostics are available during method selection. The source suite controls hero 112 against opponents

Table 1: Frozen six-seed Gamecore result, 144 scheduled matches per method and suite. Source: hero 112 against eight opponents; cross-hero: the eight cross-hero pairs of Section 5. Reward is mean shaped return with a 95% percentile interval that bootstraps the six training-seed means; W/L/D/T counts wins, losses, draws and truncations; ↑: higher is better.

| | Source suite (hero 112) | | | Cross-hero suite | | |
|---|---|---|---|---|---|---|
| Method | Reward↑ [95% CI] | W/L/D/T | Win %↑ | Reward↑ [95% CI] | W/L/D/T | Win %↑ |
| *Offline pre-adaptation policy* | | | | | | |
| BC | 27.22 [24.42, 29.05] | 142/0/0/2 | 98.6 | −17.46 [−21.03, −14.15] | 56/56/0/32 | 38.9 |
| *Source-only online adaptation (20 PPO updates)* | | | | | | |
| PPO20 | 27.82 [18.15, 36.19] | 140/1/0/3 | 97.2 | −26.40 [−38.55, −15.97] | 42/60/0/42 | 29.2 |
| + hard replay | 29.06 [23.76, 34.70] | 141/2/0/1 | 97.9 | −20.81 [−28.68, −13.54] | 54/50/0/40 | 37.5 |
| + soft distill | 29.17 [23.88, 34.60] | 141/1/0/2 | 97.9 | −18.64 [−26.07, −11.75] | 55/49/0/40 | 38.2 |

$\{112, 121, 123, 133, 140, 169, 193, 199\}$; every opponent hero is played by the engine's built-in Common AI. The cross-hero schedule is $(112, 121)$, $(121, 112)$, $(123, 133)$, $(133, 123)$, $(140, 193)$, $(169, 199)$, $(193, 140)$, and $(199, 169)$, where each pair is (controlled, opponent); $(112, 121)$ is the only cross-hero pair whose controlled hero is the source hero, and it also appears in the source suite. The three controlled-hero sentinel pairs are $(121, 112)$, $(140, 193)$, and $(123, 133)$; the remaining four non-source controlled heroes form the held-out suite. The split is disjoint at the level of the controlled hero, which is the entity whose identity block and action semantics the policy must handle; it is not disjoint at the level of the opponent, since held-out heroes 193 and 133 appear as opponents in the sentinel memory, and two held-out matchups are role swaps of sentinel matchups. BC, PPO, and hard-replay outcomes on these matchups predate the split and were known, so this is a one-time evaluation of the CSD checkpoints rather than a fully blind task benchmark. We call it *registered* because the selection summary and the release rule were fixed in an immutable, SHA-256-identified decision file before any held-out CSD game was run; this is an internal commitment, not a public preregistration. The release decision was written at 05:32:15 local time on 2026-09-13 and the six held-out evaluations started between 05:32:24 and 05:47:09, as recorded in the immutable decision file and the per-condition logs. Three consecutive Gamecore trajectories are run per matchup. Because Gamecore does not expose an addressable environment seed, repeats reduce trajectory noise but are not treated as independent inferential units.

For every method and suite, we first average scheduled matches within each training seed. We then report the mean across six seeds and a percentile interval from bootstrap resamples of those six seed-level values (100,000 resamples for Table 1 and the effects in §6.1, 20,000 for the capability tables). Method effects are paired within seed. We also report the number of positive seed-level effects and an exact two-sided sign test over non-zero differences; with six seeds the smallest attainable two-sided $p$ is $0.03125$, and a percentile bootstrap on six values is anti-conservative, so intervals are read as descriptions of repeatability rather than as tests. Scheduled win rate uses every assigned game in its denominator, so truncations remain visible as failures to complete rather than disappearing from a terminal-only rate.

## 6 Results

### 6.1 Narrow PPO Reduces Cross-Hero Win Rate

PPO20 minus BC changes source reward by $+0.61$ with a 95% seed-bootstrap interval of $[-6.86, 7.55]$. Its cross-hero scheduled win rate decreases by 9.72 percentage points $[-18.06, -1.39]$ (negative in four seeds, zero in one, positive in one; sign-test $p = 0.375$) and its terminal rate by 6.94 points $[-14.58, -0.69]$, while the cross-hero reward change, $-8.94$ $[-20.72, 0.39]$, includes zero. Hard replay minus PPO20 improves cross-hero reward by $+5.59$ $[1.93, 10.09]$, with all six seed-level effects positive and an exact two-sided sign-test $p = 0.03125$. The hard-replay policy still does not clearly exceed the original BC policy. Soft source distillation improves cross-hero reward over PPO20 by $+7.76$ $[3.41, 13.01]$ (6/6 seeds) and scheduled win rate by 9.03 percentage points $[0.69, 16.67]$ (3/6 seeds, two tied). Compared with hard replay it

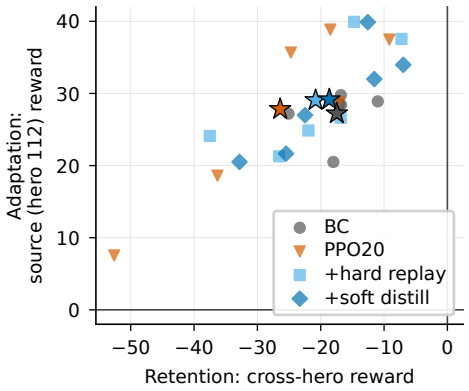


Figure 3: Adaptation–retention coordinates: source-suite against cross-hero-suite mean shaped reward. Markers are training seeds; stars are six-seed means (Table 1).

Table 2: Decision retention on the same 39,094 sentinel support states: legal-head teacher argmax agreement and $D_{\mathrm{KL}}(\pi_0 \| \pi_\theta)$, with 95% percentile bootstraps of six seed means. ↑/↓: higher/lower is better; bold: best in column.

| Method | Agreement %↑ [95% CI] | KL↓ [95% CI] |
|---|---|---|
| PPO20 | 69.40 [65.87, 74.38] | .248 [.175, .311] |
| *Source-only replay* | | |
| Hard replay | 74.55 [72.64, 76.41] | .152 [.125, .179] |
| Soft distill | 75.78 [73.74, 77.66] | .133 [.109, .166] |
| CSD (ours) | **79.35** [77.23, 81.98] | **.099** [.077, .121] |

adds $+2.17$ reward $[0.47, 3.95]$, while their win-rate difference is only $+0.69$ percentage points $[-4.86, 6.94]$. The teacher distribution is therefore a stronger source-only retention control than one sampled demonstration label, but it still revisits no non-source state.

### 6.2 Capability-Support Distillation

Across the 39,094 stored sentinel states, unconstrained PPO retains 69.40% teacher argmax agreement, hard source replay 74.55%, soft source distillation 75.78%, and capability-support distillation 79.35%. Capability-support distillation exceeds soft source distillation by 3.57 percentage points [1.61, 6.04], with positive paired effects in all six seeds. Its teacher-to-student KL is lower by 0.0337 $[-0.0589, -0.0101]$. Thus direct sentinel support changes the intended policy-retention diagnostic more than source-only replay.

The sentinel game outcome does not follow agreement. Over 54 sentinel games per method, hard replay wins one and every other method none, so a terminal sentinel game is almost always a loss. Here capability-support distillation improves shaped reward over PPO by 9.62 [4.03, 15.16] (positive in 5/6 seeds, sign-test $p = 0.22$) and terminates 63.0% of scheduled games (34 losses, 20 truncations), against 81.5% for PPO (44, 10) and 75.9% for the BC teacher (41, 13); the CSD−PPO terminal-rate difference of $-18.5$ percentage points $[-22.2, -14.8]$ is the one sentinel effect that is consistent across all six seeds ($p = 0.031$). Soft distillation shifts the same way (64.8%). The method closest to the teacher's sentinel actions thus departs further than PPO from the teacher's outcomes, losing less and stalling more without winning (Section 6.3). On the source suite, CSD is indistinguishable from the other online methods: reward 28.93, 142/0/0/2, $+1.11$ reward $[-5.73, 6.94]$ and $+1.39$ win-rate points [0.00, 2.78] relative to PPO. This divergence is evidence that policy agreement, shaped reward, and task completion measure different properties.

On the held-out suite, CSD improves reward over PPO20 by $+14.59$ $[3.21, 26.47]$ (positive in 4/6 seeds, sign-test $p = 0.69$) and scheduled win rate by 20.83 percentage points $[8.33, 33.33]$ (5/6 seeds, $p = 0.063$). It also improves reward over hard replay by $+7.47$ $[0.40, 14.85]$ (4/6 seeds, $p = 0.69$); the corresponding win-rate difference is $+6.94$ points $[-4.17, 19.44]$. CSD has the highest absolute held-out mean reward and scheduled win rate, but its paired intervals against BC and soft source distillation include zero: $+3.71$ $[-8.25, 15.45]$ and $+6.75$ $[-2.48, 16.55]$ reward, respectively. The held-out terminal rate of CSD (72.2%) is above PPO's (55.6%), so the sentinel terminal-rate shift did not recur on the held-out heroes, where terminal games include wins; one run on four heroes cannot say whether it is specific to the sentinel matchups. Because twelve games per seed is a noisy cell, we repeated the held-out evaluation once on fresh trajectories after the results were frozen (run r4, Table 3). The replicate has smaller gaps, and CSD's advantage over hard replay does not recur; pooled, the win-rate ordering is PPO $<$ soft distillation $<$ BC $<$ hard replay $\approx$ CSD. Pooled (24 games per seed and method), CSD improves on PPO by $+10.63$ reward [2.68,

Table 3: Held-out controlled heroes (four heroes, 72 scheduled matches per method and run). *Run r3* is the registered one-time evaluation: its CSD trajectories and outcomes were opened only after the release gate, while the baseline rows are the held-out subset of the cross-hero games in Table 1 and were already known. *Run r4* repeats every method on fresh trajectories after the results were frozen (Appendix B). *Pooled* averages both runs within seed (144 matches per method); intervals are 95% percentile bootstraps of the six seed means. No draws occurred; terminal % counts games ending before the step cap. Bold: highest point estimate in the column; most pooled intervals overlap (paired contrasts: Table 4, Appendix B).

| | Run r3 (registered) | | | | Run r4 | Pooled r3 + r4 | |
|---|---|---|---|---|---|---|---|
| Method | Reward↑ | W/L/T | Win %↑ | Term. % | Win %↑ | Win %↑ [95% CI] | Reward↑ [95% CI] |
| BC | −3.30 | 39/15/18 | 54.2 | 75.0 | 48.6 | 51.4 [45.8, 56.9] | −1.96 [−4.61, +1.09] |
| PPO20 | −14.19 | 26/14/32 | 36.1 | 55.6 | 45.8 | 41.0 [27.1, 55.6] | −12.47 [−31.60, +2.44] |
| + hard replay | −7.07 | 36/11/25 | 50.0 | 65.3 | **54.2** | 52.1 [34.0, 70.8] | −4.45 [−15.46, +5.93] |
| + soft distill | −6.35 | 37/14/21 | 51.4 | 70.8 | 44.4 | 47.9 [39.6, 58.3] | −7.55 [−18.10, +2.80] |
| CSD (ours) | **0.40** | 41/11/20 | **56.9** | 72.2 | 51.4 | **54.2** [46.5, 63.2] | **−1.84** [−12.95, +7.98] |

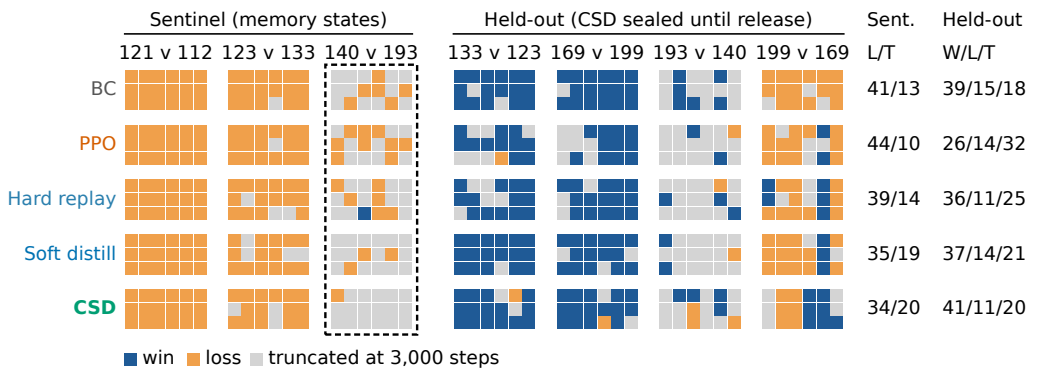


Figure 4: Every sentinel and held-out game of run r3, one square per Gamecore match (no game selected or dropped). Rows are methods; columns are (controlled, opponent) hero pairs; within a block, columns are the six training seeds (11 to 101) and rows the three repeats. Right: sentinel losses/truncations and held-out wins/losses/truncations per method (Table 3). The dashed box marks the matchup that carries most of the sentinel shift.

19.82] and 13.2 win-rate points [5.6, 21.5] (5/6 seeds, $p = 0.063$), is indistinguishable from BC (+0.12 reward, +2.8 points), and is not reliably above hard replay (+2.61 reward [−0.46, 5.71]); the pooled PPO−BC win-rate loss is −10.4 points [−27.1, 2.1]. We therefore state the held-out claim as mitigation relative to unconstrained PPO, to the level of the pre-adaptation policy (pooled win rate 54.2% for CSD, 51.4% for BC, 41.0% for PPO). The reward improvement over hard replay holds in the first run only, and the evidence does not establish superiority to the frozen BC policy or to source-only distillation: against soft distillation the pooled win-rate interval, $+6.2$ $[0.7, 13.9]$, excludes zero but only three of six seeds are positive, and the reward interval includes zero.

### 6.3 Case Study: Every Game, Matchup by Matchup

Figure 4 shows individual matches. On two of the three sentinel matchups the methods are indistinguishable: every method loses all 18 games of 121 v 112 and 14–17 of 123 v 133. The sentinel outcome shift sits in 140 v 193, where BC loses 6 games and PPO 9, while CSD loses 1 and runs the other 17 to the step cap; 8 of CSD's 10 fewer sentinel losses than PPO come from this matchup. The policy closest to the teacher on stored states thus avoids defeat on one sentinel matchup by stalling there. On held-out heroes, PPO's deficit is mostly wins turned into truncations (32 truncations against BC's 18, with 14 and 15 losses), spread over three matchups. CSD reaches BC's total

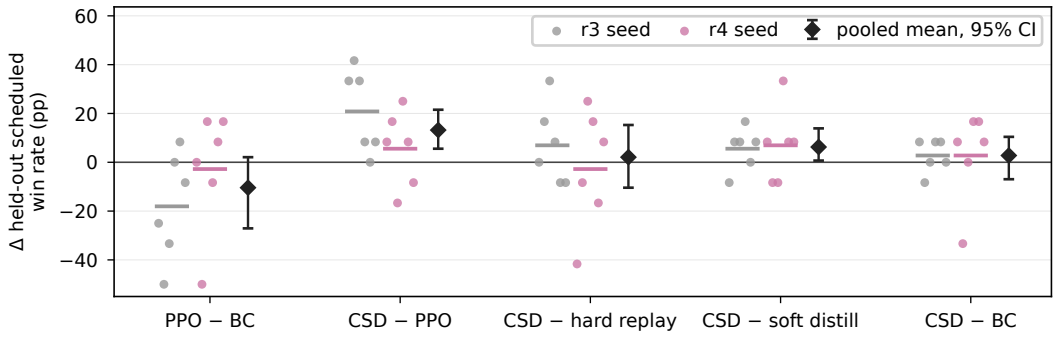


Figure 5: Seed-level paired effects on held-out scheduled win rate in the first (r3) and second (r4) held-out runs; dots are training seeds (12 games per seed and method in each run), bars are six-seed run means, and diamonds are pooled means with their 95% seed-bootstrap interval (Table 4). Effects move in steps of $1/12$, so seeds overlap.

with a different profile, winning 7 of 18 games of 199 v 169, where BC wins none, and 1–2 fewer than BC on each of the other three; aggregate parity with BC is not game-by-game parity.

### 6.4 Per-Run and Per-Seed Evidence Behind the Held-Out Claims

Table 4 (Appendix B) gives every held-out contrast per run and pooled; Figure 5 and Table 5 (Appendix B) show the seed-level effects behind them. Pooling averages all 24 games of a seed and method before forming paired differences, so the runs weigh equally and a pooled interval describes six seed-level values, not 144 independent games.

The second run reverses part of the first run's ordering (Table 3): PPO is last in r3 but not in r4, and hard replay moves from second-lowest to highest. In both runs CSD's scheduled win rate is above that of PPO, soft distillation, and BC; its reward advantage over hard replay ($+7.47$ in r3) becomes $-2.25$ $[-7.65, 4.39]$ in r4. Seed-level win-rate effects repeat poorly: their sign agrees across r3 and r4 in 2/6 seeds for PPO$-$BC and CSD$-$soft distillation, 3/6 for CSD$-$PPO and CSD$-$hard replay, and 5/6 for CSD$-$BC, counting one seed that is zero in both runs. Per seed (Table 5), the cross-hero PPO$-$BC loss of Table 1 is negative in four seeds, zero in one, and positive in one ($+4.2$ points). The pooled CSD$-$PPO win-rate gain is positive in five seeds and zero in seed 101, but uneven: seeds 11 and 23 contribute $+20.8$ and $+29.2$ points and the other four lie between 0.0 and $+16.7$. On reward, seed 101 gains $+11.71$ without any win-rate gain, and seed 89 is the only negative seed ($-0.15$). The CSD$-$PPO improvement is therefore consistent in sign but not in size, and a twelve-game cell is too noisy to rank the stronger methods within a seed.

### 6.5 Where the Cross-Hero Competence Resides

Two pre-registered analyses in Appendix B bear on why source-only updates can change behavior on other heroes. Deployment-time perturbations of the frozen BC policy (Table 10, Figure 6) show that the controlled-hero identity input is inert (win-rate changes of at most $+2.8$ points, every interval including zero) and that unseen heroes depend on the same feature blocks as the source hero, with no block showing an unseen-hero-specific dependency. Cross-hero competence therefore rides on features that source-hero PPO updates also use, so it can change without any non-source state being visited; whether PPO reshapes these blocks is not tested, because only BC is perturbed. The reward decomposition of the replicate run (Table 8) spreads PPO's held-out reward loss relative to BC across health, money, death, energy, and experience rather than the tower term.

## 7 Discussion

The results expose two forms of retention. *Decision retention* asks whether the adapted policy matches the teacher on a fixed state support. *Trajectory retention* asks whether deploying the adapted policy induces state sequences that still complete the task. Capability-support distillation directly

improves the first quantity, but the second can fail through compounding action differences and a shift away from the teacher's state occupancy. An agent can collect less negative reward by surviving to the cap instead of winning, as CSD does on one sentinel matchup (Figure 4).

For evaluation, teacher KL and action agreement diagnose optimization behavior but here did not predict terminal outcomes, and shaped return rose on the sentinel suite without any added win. A retention claim therefore requires agreement on historical support, outcomes under deployment, and transfer to disjoint tasks. The held-out result reinforces this ordering: CSD has the highest pooled held-out win rate and improves on PPO with pooled intervals that exclude zero, yet it only matches the pre-adaptation policy, and source soft distillation, which collects no non-source state, is within its reward interval. The most immediate method question is therefore how support states should be selected to constrain the future occupancy induced by the adapted policy, rather than how closely a student can match a teacher on a static memory alone.

## 8 LIMITATIONS

**Statistical resolution.** Each run gives twelve held-out games per seed and method, and Gamecore random seeds are inaccessible, so repeats cannot be matched across methods. Seed-level effects repeat in sign between the two held-out runs in only two or three of six seeds for most contrasts (Section 6.4), six seeds bound the exact sign test at $p = 0.03125$, and a percentile bootstrap over six values is anti-conservative. The held-out claims therefore rest on the pooled effects, and the ordering among BC, hard replay, soft distillation, and CSD is not resolved. We report many paired contrasts without multiplicity correction; the individual intervals should be read accordingly.

**Scope of the design.** All methods use 20 PPO updates, one auxiliary coefficient (0.1), and one source hero; we do not measure how damage or mitigation evolves with longer adaptation or other coefficients. The sentinel memory comes from one game per sentinel matchup per seed, and decision retention is measured only on sentinel states, because held-out states were sealed during selection. Because the sentinel opponents are almost never beaten, the sentinel suite measures trajectory retention only through losses and truncations, not through wins. We do not ablate the ingredients of CSD (sentinel states versus legality masks versus teacher targets), nor compare it with parameter-space regularizers such as EWC.

**Protocol and artifact.** The Common AI opponent pool does not establish competition-level strength. Replay, distillation, and adapter hyperparameters were screened on training seed 101 before the six-seed confirmation; seed 101 is included in the reported means. Excluding it leaves the pooled held-out CSD$-$PPO effect at $+10.42$ reward $[0.87, 21.74]$ and $+15.8$ win-rate points $[8.3, 23.3]$ over five seeds (verification code in the supplementary material). Several long evaluation conditions lost their simulator child process and were restarted from game zero on isolated servers; no partial condition entered any aggregate. Held-out heroes are disjoint from sentinel heroes only as controlled entities, not as opponents. The compact recurrent model is not the official HoK-Off reference architecture. Evidence from one 1v1 game cannot establish general game-agent retention.

## 9 CONCLUSION

In one 1v1 game, source-only PPO reduced the cross-hero win rate of a policy cloned from a one-hero offline corpus without a reliable source-task gain. Hard replay and source-policy distillation reduce this damage, making source distillation a necessary control for capability-preservation claims. Capability-support distillation further improves teacher agreement on reward-free sentinel states and, in the registered one-time held-out evaluation, improves reward and scheduled win rate over unconstrained PPO (positive in both runs; pooled intervals exclude zero), returning held-out performance to the level of the pre-adaptation policy. It does not establish a reliable advantage over behavioral cloning or source distillation, and its higher sentinel agreement coexists with a sentinel outcome profile (fewer losses, more truncations, no wins) that is further from the teacher's than PPO's. Offline-to-online game-agent studies should therefore report fixed-state decision retention, deployed trajectory outcomes, and disjoint-task transfer separately. Future work should select support states for their coverage of policy-induced occupancy and test whether the distinction generalizes beyond 1v1 Honor of Kings.

## REPRODUCIBILITY STATEMENT

The implementation builds on a fixed revision of the public Honor of Kings Arena environment (Wei et al., 2022) and keeps an experiment manifest with the training and evaluation code of every run. Contract tests assert the 725-dimensional observation, 172-entry legal-action structure, active-head masking, hero identity, and episode-safe recurrent sampling. Each checkpoint, capability-memory file, selection summary, and release decision is identified by SHA-256. Gamecore evaluation writes one machine-readable record per scheduled match and a condition summary; aggregation refuses incomplete seed–method suites. The held-out release decision is immutable and stores the hash of the exact selection summary that triggered it.

Code, per-match records, manifests, and the pre-registered analysis plans are provided in the supplementary material. The HoK-Off corpus, interface-derived code, manifests, aggregate statistics, and plots can be redistributed. The licensed Gamecore binary cannot; reproducing terminal outcomes therefore requires authorized engine access. This boundary is why machine-readable per-match records, contract tests, and the distinction between public data and licensed execution are part of the reported artifact.

## AI USE STATEMENT

We used a generative AI assistant (Claude, Anthropic), under the authors' direction, to give feedback on experimental design, to implement analysis and plotting code, to help interpret results, to draft and edit the text, to create and edit figures and tables, to identify and summarise related literature, and to format references. We did *not* use generative AI tools to generate synthetic data reported as measurements, or for translation. We have reviewed all AI-assisted work: every reported number was checked by the authors against the logged experimental outputs, and every reference was checked against its published record. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

## REFERENCES

Anonymous. The three regimes of offline-to-online reinforcement learning. *Manuscript under review*, 2026. URL https://openreview.net/forum?id=IeZf5RvF2i. Cited only as a contemporary framing; not treated as established peer-reviewed evidence.

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2302.02948.

Wojciech Marian Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant M. Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019. URL https://arxiv.org/abs/1902.02186.

Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Policy consolidation for continual reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3242–3251. PMLR, 2019. URL https://proceedings.mlr.press/v97/kaplanis19a.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, 2021. URL https://arxiv.org/abs/2107.00591.

Mao-Lin Luo, Zhe-Xu Wang, Zi-Hao Zhou, Bo Ye, Jian Zhao, Min-Ling Zhang, and Tong Wei. RL forgets! Towards continual policy optimization. *arXiv preprint arXiv:2607.04364*, 2026. URL https://arxiv.org/abs/2607.04364.

Yun Qu, Boyuan Wang, Jianzhun Shao, Yuhang Jiang, Chen Chen, Zhenbin Ye, Lin Liu, Junfeng Yang, Lin Lai, Hongyang Qin, Minwen Deng, Juchao Zhuo, Deheng Ye, Qiang Fu, Wei Yang, Guang Yang, Lanxiao Huang, and Xiangyang Ji. Hok-off: Real game dataset from honor of kings and its offline reinforcement learning benchmarks. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2408.10556.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 348–358, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html.

Stephane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pp. 661–668. PMLR, 2010. URL https://proceedings.mlr.press/v9/ross10a.html.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011. URL https://arxiv.org/abs/1011.0686.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06295.

Simon Schmitt, Jonathan J. Hudson, Augustin Zidek, Simon Osindero, Carl Doersch, Wojciech M. Czarnecki, Joel Z. Leibo, Heinrich Kuttler, Andrew Zisserman, Karen Simonyan, and S. M. Ali Eslami. Kickstarting deep reinforcement learning. *arXiv preprint arXiv:1803.03835*, 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*, 2025. URL https://arxiv.org/abs/2509.04259.

René Traoré, Hugo Caselles-Dupré, Timothée Lesort, Te Sun, Guanghang Cai, Natalia Díaz-Rodríguez, and David Filliat. Discorl: Continual reinforcement learning via policy distillation. *arXiv preprint arXiv:1907.05855*, 2019.

Hua Wei, Jingxiao Chen, Xiyang Ji, Hongyang Qin, Minwen Deng, Siqin Li, Liang Wang, Weinan Zhang, Yong Yu, Lin Liu, Lanxiao Huang, Deheng Ye, Qiang Fu, and Wei Yang. Honor of kings arena: An environment for generalization in competitive reinforcement learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022. URL https://arxiv.org/abs/2209.08483.

Maciej Wołczyk, Bartłomiej Cupiał, Mateusz Ostaszewski, Michał Bortkiewicz, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Fine-tuning reinforcement learning models is secretly a forgetting mitigation problem. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.02868.

Deheng Ye, Guibin Chen, Wen Zhang, Sheng Chen, Bo Yuan, Bo Liu, Jia Chen, Zhao Liu, Fuhao Qiu, Hongsheng Yu, Yinyuting Yin, Bei Shi, Liang Wang, Tengfei Shi, Qiang Fu, Wei Yang, Lanxiao Huang, and Wei Liu. Towards playing full MOBA games with deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2011.12692.

Tiantian Zhang, Kevin Zehua Shen, Zichuan Lin, Bo Yuan, Xueqian Wang, Xiu Li, and Deheng Ye. Replay-enhanced continual reinforcement learning. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2311.11557.

Zhiyuan Zhou, Andy Peng, Qiyang Li, Sergey Levine, and Aviral Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HN0CYZbAPw.

# A Environment and Data Contract

The 725-dimensional policy input consists of the official 705 game features plus a 20-way controlled-hero identity. Because the identity is constant during single-hero cloning, it is inert at deployment: zeroing or falsifying it changes the frozen policy's win rate by under 3 pp on both seen and unseen heroes, and block knockouts show that unseen-hero competence rests on the same own-hero, enemy-hero, and enemy-soldier features that carry source-hero play (Appendix B). It includes friendly and enemy hero blocks, a public controlled-hero block, soldier and structure slots for both camps, and global features. The action has six categorical heads of sizes $(12, 16, 16, 16, 16, 8)$: button, move X/Y, skill-offset X/Y, and target. Gamecore supplies 172 legal-action entries, including a $12 \times 8$ target mask conditioned on the button. Illegal logits are masked before sampling, and inactive heads are omitted from every policy or imitation loss.

The interface returns a ten-entry per-step reward vector whose final entry is the shaped sum that PPO optimizes. Entry 6 is identically zero and is dropped by the environment wrapper; the remaining eight entries are, in order, death, energy ratio, experience, health, kill, last hit, money, and tower health. We verified the resolved coefficients empirically: a least-squares fit of the shaped sum on the eight per-game entry sums over 150 evaluation games is exact (zero residual) with weights $-1.0$, $0.75$, $0.006$, $2.0$, $-0.6$, $0.5$, $0.006$, and $5.0$, identical to the configuration file, so per-entry attributions of every reward difference are well defined (Appendix B). Evaluation derives win, loss, or draw from the terminal state of each camp's crystal (the base structure, organ type 24, 7,000 hit points at the start): the learner wins if only the opponent's crystal is destroyed, loses if only its own is, and draws otherwise. Matches exceeding 3,000 agent steps are truncations and count as non-wins in scheduled win rate.

The HoK-Off `norm_medium` dataset contains 1,776,809 decisions in 940 complete episodes, all controlling hero 112. The official sampling recipe generates `norm_medium` with matched level-1 pretrained policies in Gamecore; we therefore call it offline policy data rather than human demonstration data. Every seed uses the same episode-disjoint hash split: 256 training episodes (478,927 rows) and 64 validation episodes (129,573 rows).

# B Replicate, Reward Decomposition, and Mechanism Ablations

This appendix reports a pre-registered evaluation matrix run after the main results were frozen (hypotheses recorded before any result; the plan is in the supplementary material): a second independent held-out evaluation of every method on fresh trajectories (r4), the per-entry decomposition of the interface's reward vector, and deployment-time ablations of the frozen BC policy that test where its cross-hero competence resides. Every number below is computed from the per-match records by the released aggregation code.

**Replicate and pooled held-out results.** Tables 3 and 4 give both runs and their pool. Three facts survive pooling: unconstrained PPO has the lowest held-out win rate in the first run and in the pool; CSD improves on PPO with intervals that exclude zero on reward and win rate in the pool; and CSD equals the pre-adaptation policy rather than exceeding it. Two claims that the first run alone would have supported do not survive: the PPO$-$BC win-rate loss (interval includes zero when pooled) and CSD's advantage over hard replay (sign reverses in r4). Seed-level paired win-rate effects agree in sign between the runs on only two to three of six seeds for every contrast except CSD$-$BC (five of six, one of them zero in both runs; Section 6.4), so a twelve-game cell is dominated by trajectory noise and per-seed sign counts should not be over-read at this cell size.

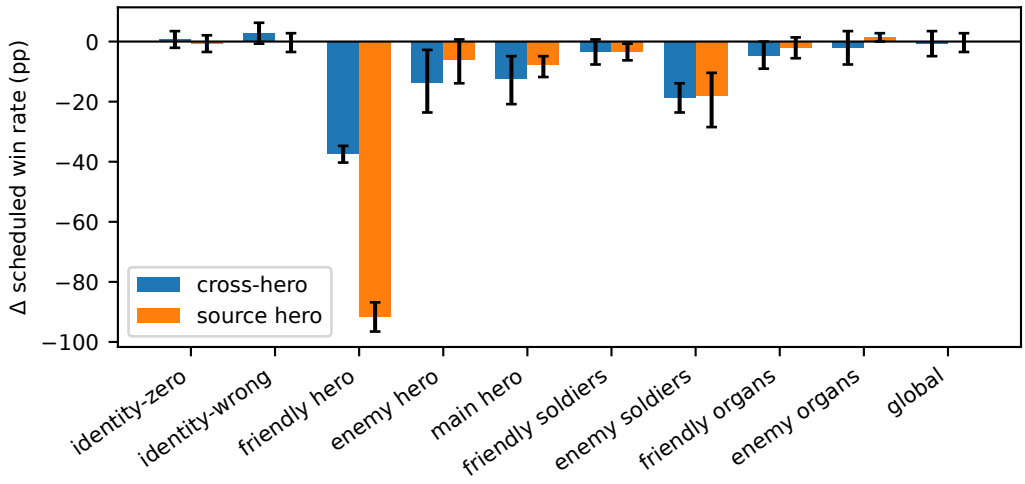


Figure 6: Change in scheduled win rate of the frozen BC policy under each deployment-time perturbation, on unseen heroes (cross-hero) and on the source hero, paired against the unperturbed control within seed; bars are six-seed means with percentile-bootstrap intervals.

**Reward decomposition.** Table 8 attributes each reward difference to the eight weighted entries. Our pre-registered prediction that PPO's held-out loss would be carried by the tower and death terms was wrong: the loss is spread across health ($-3.09$), money ($-3.33$), death ($-1.69$), energy ($-1.15$), and experience ($-1.09$), with the tower term near zero ($-0.26$); the only entry whose interval excludes zero is the kill term, which moves in PPO's favor ($+1.02$ because its coefficient is negative). CSD's recovery over PPO is likewise spread (health $+1.32$, money $+1.91$, death $+1.01$, tower $+0.86$), with death and tower intervals excluding zero. The shaped reward therefore does not isolate one behavior that online PPO breaks; it degrades laning economy and survival together.

**Mechanism ablations.** Table 10 and Figure 6 report the identity and block-knockout perturbations against the contemporaneous unperturbed control (37.5% scheduled wins on unseen heroes, 98.6% on the source hero). Both pre-registered hypotheses survive. H1 (identity): zeroing or falsifying the hero one-hot changes the win rate by at most $+2.8$ pp on either suite and every interval includes zero, so the frozen policy carries no hero-conditional behaviour through the identity input; the one-hot was constant during behaviour cloning and is inert at deployment (the identity-wrong interval on unseen heroes, $[-0.7, +6.3]$, is the only one whose upper bound exceeds the pre-registered 3 pp margin). H2 (knockouts): the blocks that matter for unseen heroes are the blocks that matter for the source hero. Removing the friendly-hero block destroys play everywhere (0.0% and 6.9% wins); removing the enemy-soldier block costs about 18 pp on both suites ($-18.8$ and $-18.1$); removing the enemy-hero block or the 14-entry main-hero block costs 12–14 pp on unseen heroes and 6–8 pp on the source hero; the friendly-soldier, organ, and global blocks are within noise on both. The cross-minus-fixed contrast excludes zero only for the friendly-hero block, and in the direction opposite to a cross-specific dependency: it reflects the floor at 0% on a suite whose control already wins only 37.5%, against a 92 pp drop on the source hero. The two candidates for a cross-specific dependency, enemy hero ($-7.6$ pp, $[-19.4, +3.5]$) and main hero ($-4.9$ pp, $[-9.7, +0.0]$), have intervals that include or touch zero. We therefore read the transfer as distributed: cross-hero competence sits in the same own-hero, enemy-hero, and enemy-soldier features that carry source-hero play, not in a block that is consulted only when the controlled hero changes. One unplanned observation is that dropping the friendly-soldier block raises source-hero reward by $+7.7$ $[+4.8, +10.7]$ while lowering its win rate by 3.5 pp; we report it without interpretation. These ablations perturb the frozen BC policy only; whether online PPO reshapes the same blocks is left to future work.

Table 4: Paired held-out effects per run and pooled (six seeds; 95% percentile bootstrap of the seed-level differences; last column: positive seeds and exact two-sided sign test on the win-rate effect of that run). r3 CSD rows are from the registered one-time summary; r4 and pooled rows are from the pooling code.

| Contrast | Run | $\Delta$ reward | $\Delta$ win (pp) | seeds / $p$ |
|---|---|---|---|---|
| PPO − BC | r3 | −10.88 [-29.82, +4.73] | −18.1 [-34.7, -2.8] | 1/6, 0.375 |
| | r4 | −10.15 [-28.18, +2.52] | −2.8 [-23.6, +12.5] | 3/6, 1.000 |
| | pooled | −10.52 [-28.21, +2.64] | −10.4 [-27.1, +2.1] | 2/6, 0.688 |
| CSD − PPO | r3 | +14.59 [+3.21, +26.47] | +20.8 [+8.3, +33.3] | 5/6, 0.062 |
| | r4 | +6.68 [+0.35, +15.93] | +5.6 [-5.6, +16.7] | 4/6, 0.688 |
| | pooled | +10.63 [+2.68, +19.82] | +13.2 [+5.6, +21.5] | 5/6, 0.062 |
| CSD − PPO + hard replay | r3 | +7.47 [+0.40, +14.85] | +6.9 [-4.2, +19.4] | 3/6, 1.000 |
| | r4 | −2.25 [-7.65, +4.39] | −2.8 [-20.8, +13.9] | 3/6, 1.000 |
| | pooled | +2.61 [-0.46, +5.71] | +2.1 [-10.4, +15.3] | 3/6, 1.000 |
| CSD − PPO + soft distillation | r3 | +6.75 [-2.48, +16.55] | +5.6 [-1.4, +11.1] | 4/6, 0.375 |
| | r4 | +4.67 [-1.15, +11.96] | +6.9 [-2.8, +18.1] | 4/6, 0.688 |
| | pooled | +5.71 [-1.19, +13.41] | +6.2 [+0.7, +13.9] | 3/6, 0.250 |
| CSD − BC | r3 | +3.71 [-8.25, +15.45] | +2.8 [-2.8, +6.9] | 3/6, 0.625 |
| | r4 | −3.47 [-12.83, +4.61] | +2.8 [-12.5, +13.9] | 4/6, 0.375 |
| | pooled | +0.12 [-9.89, +8.09] | +2.8 [-6.9, +10.4] | 4/6, 0.375 |
| PPO + hard replay − PPO | r3 | +7.12 [-0.35, +15.36] | +13.9 [+4.2, +23.6] | 4/6, 0.125 |
| | r4 | +8.93 [+0.09, +19.66] | +8.3 [-16.7, +30.6] | 4/6, 0.688 |
| | pooled | +8.02 [-0.02, +17.58] | +11.1 [-4.9, +27.1] | 4/6, 0.375 |

Table 5: Seed-level values behind the headline effects. PPO−BC is the scheduled win-rate change on the full cross-hero suite of Table 1 (24 games per seed); CSD−PPO is the held-out change in each run and pooled. The last row counts strictly positive seeds.

| | PPO−BC | CSD−PPO, $\Delta$ win (pp) | | | CSD−PPO, $\Delta$ reward | | |
|---|---|---|---|---|---|---|---|
| Seed | cross win (pp) | r3 | r4 | pooled | r3 | r4 | pooled |
| 11 | −12.5 | +33.3 | +8.3 | +20.8 | +34.20 | +5.23 | +19.71 |
| 23 | −20.8 | +41.7 | +16.7 | +29.2 | +29.71 | +28.73 | +29.22 |
| 47 | −25.0 | +33.3 | −16.7 | +8.3 | −2.57 | +5.28 | +1.35 |
| 71 | +0.0 | +8.3 | +25.0 | +16.7 | +1.27 | +2.62 | +1.95 |
| 89 | +4.2 | +0.0 | +8.3 | +4.2 | −1.81 | +1.51 | −0.15 |
| 101 | −4.2 | +8.3 | −8.3 | +0.0 | +26.73 | −3.31 | +11.71 |
| mean | −9.72 | +20.8 | +5.6 | +13.2 | +14.59 | +6.68 | +10.63 |
| positive | 1/6 | 5/6 | 4/6 | 5/6 | 4/6 | 5/6 | 5/6 |

Table 6: Held-out replicate (r4) on fresh trajectories: four held-out controlled heroes, three repeats, six training seeds (72 scheduled matches per method). Same protocol as Table 3.

| Method | Reward [95% CI] | Win % | Terminal % |
|---|---|---|---|
| BC | −0.61 [-5.16, +3.41] | 48.6 | 68.1 |
| PPO | −10.76 [-29.82, +3.43] | 45.8 | 63.9 |
| PPO + hard replay | −1.83 [-13.90, +9.33] | 54.2 | 68.1 |
| PPO + soft distillation | −8.76 [-21.19, +2.79] | 44.4 | 63.9 |
| CSD | −4.09 [-14.58, +4.76] | 51.4 | 65.3 |

Table 7: Paired held-out effects on the r4 replicate (six seeds; percentile bootstrap; exact two-sided sign test).

| Contrast | Δ reward | Δ win % | seeds / $p$ |
|---|---|---|---|
| PPO − BC | −10.15 [-28.18, +2.52] | −2.8 [-23.6, +12.5] | 3/6, 1.000 |
| CSD − PPO | +6.68 [+0.35, +15.93] | +5.6 [-5.6, +16.7] | 4/6, 0.688 |
| CSD − PPO + hard replay | −2.25 [-7.65, +4.39] | −2.8 [-20.8, +13.9] | 3/6, 1.000 |
| CSD − PPO + soft distillation | +4.67 [-1.15, +11.96] | +6.9 [-2.8, +18.1] | 4/6, 0.688 |
| CSD − BC | −3.47 [-12.83, +4.61] | +2.8 [-12.5, +13.9] | 4/6, 0.375 |

Table 8: Reward decomposition on the held-out r4 suite. Each cell is the paired seed-level difference of the per-game sum of one reward-vector entry multiplied by its verified coefficient (in parentheses), so the rows sum to the shaped-reward difference in the last row. Six seeds; percentile bootstrap.

| Entry (weight) | PPO − BC | CSD − PPO | CSD − BC |
|---|---|---|---|
| dead (-1) | −1.69 [-3.94, +0.18] | +1.01 [+0.18, +1.92] | −0.68 [-2.14, +0.47] |
| ep rate (0.75) | −1.15 [-2.91, +0.47] | +0.52 [-0.29, +1.33] | −0.63 [-1.70, +0.43] |
| exp (0.006) | −1.09 [-3.81, +1.25] | +0.90 [-1.06, +3.01] | −0.20 [-1.28, +0.81] |
| hp point (2) | −3.09 [-7.69, +1.07] | +1.32 [-0.67, +3.30] | −1.77 [-4.66, +1.04] |
| kill (-0.6) | +1.02 [+0.09, +2.15] | −0.64 [-1.03, -0.24] | +0.38 [-0.24, +1.17] |
| last hit (0.5) | −0.56 [-1.61, +0.78] | +0.81 [-0.50, +1.81] | +0.25 [-0.14, +0.81] |
| money (0.006) | −3.33 [-8.61, +0.54] | +1.91 [-0.26, +5.09] | −1.42 [-3.91, +0.95] |
| tower hp point (5) | −0.26 [-2.93, +1.93] | +0.86 [+0.11, +1.72] | +0.60 [-1.44, +2.73] |
| shaped reward (sum) | −10.15 [-28.18, +2.52] | +6.68 [+0.35, +15.93] | −3.47 [-12.83, +4.61] |

Table 9: Raw (unweighted) per-game entry sums, paired seed-level differences, same suite as Table 8. Raw units are the interface's own (e.g. experience points, gold, health points).

| Entry | PPO − BC | CSD − PPO | CSD − BC |
|---|---|---|---|
| dead | +1.7 [-0.2, +3.9] | −1.0 [-1.9, -0.2] | +0.7 [-0.5, +2.1] |
| ep rate | −1.5 [-3.9, +0.6] | +0.7 [-0.4, +1.8] | −0.8 [-2.3, +0.6] |
| exp | −182.5 [-634.9, +208.9] | +149.2 [-177.5, +500.9] | −33.3 [-212.6, +134.7] |
| hp point | −1.5 [-3.8, +0.5] | +0.7 [-0.3, +1.7] | −0.9 [-2.3, +0.5] |
| kill | −1.7 [-3.6, -0.2] | +1.1 [+0.4, +1.7] | −0.6 [-2.0, +0.4] |
| last hit | −1.1 [-3.2, +1.6] | +1.6 [-1.0, +3.6] | +0.5 [-0.3, +1.6] |
| money | −555.0 [-1435.7, +89.4] | +318.0 [-42.6, +848.9] | −236.9 [-651.9, +158.1] |
| tower hp point | −0.1 [-0.6, +0.4] | +0.2 [+0.0, +0.3] | +0.1 [-0.3, +0.5] |

Table 10: Deployment-time perturbations of the frozen BC policy, six seeds, 24 scheduled matches per seed and suite. Effects are paired within seed against the unperturbed BC control evaluated in the same batch. The last column tests whether a block matters more for unseen heroes than for the source hero.

| Perturbation | Cross win % | Δ cross win (pp) | Fixed win % | Δ fixed win (pp) | cross − fixed effect (pp) |
|---|---|---|---|---|---|
| none (control) | 37.5 | – | 98.6 | – | – |
| identity-zero | 38.2 | +0.7 [-2.1, +3.5] | 97.9 | −0.7 [-3.5, +2.1] | +1.4 [-2.8, +4.9] |
| identity-wrong | 40.3 | +2.8 [-0.7, +6.3] | 98.6 | −0.0 [-3.5, +2.8] | +2.8 [-2.1, +6.9] |
| knockout-friendly hero | 0.0 | −37.5 [-40.3, -34.7] | 6.9 | −91.7 [-96.5, -86.8] | +54.2 [+47.9, +60.4] |
| knockout-enemy hero | 23.6 | −13.9 [-23.6, -2.8] | 92.4 | −6.2 [-13.9, +0.7] | −7.6 [-19.4, +3.5] |
| knockout-main hero | 25.0 | −12.5 [-20.8, -4.9] | 91.0 | −7.6 [-11.8, -4.9] | −4.9 [-9.7, +0.0] |
| knockout-friendly soldiers | 34.0 | −3.5 [-7.6, +0.7] | 95.1 | −3.5 [-6.3, -0.7] | +0.0 [-5.6, +4.9] |
| knockout-enemy soldiers | 18.8 | −18.8 [-23.6, -13.9] | 80.6 | −18.1 [-28.5, -10.4] | −0.7 [-6.9, +5.6] |
| knockout-friendly organs | 32.6 | −4.9 [-9.0, +0.0] | 96.5 | −2.1 [-5.6, +1.4] | −2.8 [-9.0, +2.8] |
| knockout-enemy organs | 35.4 | −2.1 [-7.6, +3.5] | 100.0 | +1.4 [+0.0, +2.8] | −3.5 [-9.7, +2.8] |
| knockout-global | 36.8 | −0.7 [-4.9, +3.5] | 98.6 | −0.0 [-3.5, +2.8] | −0.7 [-6.2, +4.2] |