DISTRIBUTIONAL EPISTEMIC ROBUSTIFICATION FOR SEQUENTIAL PORTFOLIO OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Portfolio reinforcement-learning systems must operate with policies frozen before future market regimes are observed, because return dynamics shift between historical training and deployment. Overcoming this requires separating ordinary market-outcome risk from uncertainty about which portfolio action remains valuable, rather than treating all predictive dispersion as one risk signal. Distributional value learning is well suited to representing stochastic portfolio returns, yet its within-model return law leaves epistemic ambiguity unresolved once historical data support multiple plausible value functions, especially in offline frozen-policy extraction. Within that subcase, distributional critic ensembles address model uncertainty through predictive disagreement, but aggregate dispersion can mix within-critic outcome variation with between-critic disagreement about expected action values and thereby penalize the wrong uncertainty. We therefore propose Distributional Epistemic Robustification (DER), which separates return variation from disagreement across learned value functions and yields exact structural guarantees for the extractor together with a conditional shift-aware bound. Experiments on a strictly time-ordered, cost-aware A-share benchmark show that DER attains the highest aggregate and worst-calendar Active ARR among the external baselines in the primary comparison. A matched uncertainty-target control shows that penalizing cross-critic expected-value disagreement preserves substantially more return than penalizing within-critic or total dispersion while remaining competitive with ensemble-mean extraction.

1 INTRODUCTION

Portfolio reinforcement learning is ultimately deployed after training, when an agent must convert market observations and current holdings into capital allocations under transaction costs and liquidity constraints (Moody et al., 1998; Moody & Saffell, 2001; Jiang et al., 2017; Wang et al., 2021b). A policy selected from historical data must continue trading through later bull, bear, and sideways regimes without observing future rewards and cannot refit to the test market. The specific problem studied here is how to extract a fixed portfolio policy whose decisions remain dependable when the market regime changes after training.

Solving this problem requires separating ordinary market-outcome risk from uncertainty about whether an estimated action ranking will transfer. For a risk-neutral portfolio mandate, the core capability is to preserve genuine return variation while robustifying only ambiguity that can change the chosen allocation.

Scalar value methods collapse future outcomes to conditional means (van Hasselt et al., 2016; Kumar et al., 2020). Distributional reinforcement learning enriches that representation with the full return law (Bellemare et al., 2017; Dabney et al., 2018b;a), yet a law conditional on one fitted critic cannot reveal whether alternative fits disagree about the action ordering. Deep ensembles expose variation across fitted value functions (Osband et al., 2016; Lakshminarayanan et al., 2017), and offline pessimism converts uncertainty into conservative action values (An et al., 2021; Ghasemipour et al., 2022). These layers leave a remaining methodological gap. Without separating within-model outcome variation from between-model ambiguity, a frozen portfolio rule cannot identify which uncertainty should alter its action. This narrows the problem to distributional critic ensembles, which contain both sources in one architecture.

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.

Distributional critic ensembles represent stochastic returns and multiple fitted models jointly (Eriksson et al., 2022). They can support uncertainty-aware action selection through predictive dispersion, but a score built from combined dispersion is drawn from both the return law and model variation, so it can penalize genuine market-outcome risk when only the expected ranking is fragile. Penalizing within-critic variance alone has the same mismatch for a risk-neutral mandate.

We therefore propose Distributional Epistemic Robustification (DER), a two-stage method that learns categorical portfolio-return laws and then performs epistemically robust policy extraction. Its core insight is to reduce each critic to its risk-neutral expectation before constructing a local ambiguity set from disagreement across those expectations, so ordinary outcome variation cannot enter the robustification term. This construction yields exact robust-duality, selective-switching, and action-stability results together with a conditional shift-aware error bound.

We evaluate DER in a cost-aware A-share benchmark with strict time-based separation and a policy frozen before all future test regimes. Among the external baselines in the primary comparison, DER has the highest five-year mean and worst-calendar Active ARR, the highest mean Sharpe, Sortino, and Calmar, and the second-highest low-volatility information ratio.

Contributions.

1. We introduce DER, a two-stage portfolio method that combines categorical return-law learning with a mean-ambiguity policy extractor that is risk-neutral within critics and ambiguity-averse only across their expected action values.
2. We derive robust duality, selective Pareto switching, finite ambiguity paths, deterministic action stability, and a frozen-policy error decomposition that keeps residual market shift explicit.
3. We provide a strictly time-ordered, cost-aware A-share evaluation with yearly disclosure, six external comparators, and a matched five-repeat control that isolates the uncertainty target used at extraction.

2 Related Work

Reinforcement learning for portfolio management. Portfolio agents range from direct reinforcement objectives to deep allocation networks, market encoders, and hierarchical execution systems (Moody et al., 1998; Moody & Saffell, 2001; Jiang et al., 2017; Wang et al., 2021b;a; Ye et al., 2020). FinRL-Meta and TradeMaster emphasize realistic market environments and reproducible comparison (Liu et al., 2022; Sun et al., 2023a), while expert mixtures and meta-policies improve adaptation across changing market conditions (Sun et al., 2023b; Niu et al., 2022; Qin et al., 2024; Zong et al., 2024). Recent financial systems use transformed-data conservative targets and meta-controllers over heterogeneous risk agents (Yuan et al., 2026; Chen et al., 2026). Backtest selection can nevertheless inflate apparent performance when calendars are repeatedly reused (White, 2000; Bailey et al., 2017; Harvey et al., 2016). Our evaluation therefore targets one cost-aware portfolio policy frozen before five future calendars, with every annual outcome disclosed.

Offline value learning under deployment shift. Modern offline methods constrain unsupported actions, reduce extrapolated values, and preserve proximity to behavior support (Fujimoto et al., 2019; Kumar et al., 2020; Levine et al., 2020; Wu et al., 2022). High-confidence and baseline-bootstrapped methods instead emphasize conservative policy improvement (Thomas et al., 2015; Ghavamzadeh et al., 2016; Laroche et al., 2019; Chandak et al., 2020). DER complements these training-time approaches by addressing the practically important post-training decision of how a fixed critic should rank allocation actions in a later market.

Distributional value and financial outcome risk. Categorical, quantile, implicit-quantile, and nonparametric critics represent the return distribution rather than only its expectation (Bellemare et al., 2017; Dabney et al., 2018b;a; Morimura et al., 2010). Conservative distributional offline learning modifies training to control overestimation (Ma et al., 2021); distributional ensembles model aleatoric and epistemic risk jointly (Eriksson et al., 2022); and DRL-ORA adapts epistemic risk online (Wu et al., 2026). Building on distributional critics and ensemble estimation, DER contributes

a hierarchical extractor that keeps the learned return law risk-neutral internally while robustifying only disagreement about its expected value.

Ensemble pessimism and ambiguity sets. Bootstrapped critics and deep ensembles provide practical estimates of predictive disagreement (Osband et al., 2016; Lakshminarayanan et al., 2017). Offline ensembles can support uncertainty-aware pessimism through diversified, independently trained Q-functions (An et al., 2021; Ghasemipour et al., 2022), while CQL learns conservative values without requiring an ensemble (Kumar et al., 2020). Robust MDPs and distributionally robust optimization act against explicit uncertainty sets (Iyengar, 2005; Nilim & Ghaoui, 2005; Ben-Tal et al., 2009; Namkoong & Duchi, 2017), and recent epistemic-robust offline RL constructs structured sets over action-value vectors inside Bellman targets (Chenreddy & Delage, 2026). DER supplies a post-training, backbone-compatible portfolio operator whose adversary acts only on critic expectations. Unlike training-time ensemble pessimism and ambiguity sets embedded in Bellman targets, it leaves the learned return laws and Bellman backup unchanged and modifies only frozen-policy extraction. Its distinctive object is therefore not an uncertainty set by itself, but the risk-neutral-within/robust-across composition and the selective allocation path induced by that composition.

3 Problem Setup

We formulate portfolio management as an offline finite-action MDP with state s, allocation action $a \in \mathcal{A}$, net block reward r, discount γ, and discounted portfolio return $G_t = \sum_{k\geq 0} \gamma^k r_{t+k}$. The state contains price-volume features computed only from current and past information, current holdings, and portfolio value. Each action selects one long-only portfolio construction rule for the next twenty-day block rather than directly predicting an unconstrained weight vector. Because the simulator is price-taking, the training panel supplies counterfactual transitions for every candidate allocation from each reachable holdings state, while transaction costs and liquidity constraints enter the realized reward. A transition is written as $(s_t, a_t, r_t, s_{t+1}, \delta_t)$, where r_t is the net portfolio return over one decision block and δ_t marks termination. The action changes the portfolio held during the next block, so its value combines the immediate net return with the discounted value of the resulting holdings state. This construction aligns the learned ranking with deployment because every candidate is scored from the same pre-decision information, and the selected rule is executed only at the next admissible price.

Definition 1 (Distributional critic)**.** *A distributional critic $Z_j(s,a)$ is a learned random variable approximating the conditional law of G_t for ensemble member j. Its risk-neutral action value is $q_j(s,a) = \mathbb{E}[Z_j(s,a)]$.*

Definition 2 (Epistemic disagreement)**.** *For M critics, define*

$$\mu(s,a) = \frac{1}{M}\sum_{j=1}^{M} q_j(s,a), \qquad \sigma_{\mathrm{epi}}(s,a) = \left[\frac{1}{M}\sum_{j=1}^{M}\big(q_j(s,a) - \mu(s,a)\big)^2\right]^{1/2}. \tag{1}$$

The first quantity is ensemble consensus and the second is disagreement about expected return.

Introduce a uniformly sampled member index J and a critic-induced mixture variable $\widetilde{G}$ satisfying $\widetilde{G} \mid (J = j, s, a) \sim Z_j(s,a)$. The law of total variance gives

$$\mathrm{Var}(\widetilde{G} \mid s,a) = \mathbb{E}_J\Big[\mathrm{Var}(\widetilde{G} \mid J,s,a)\Big] + \mathrm{Var}_J\Big(\mathbb{E}[\widetilde{G} \mid J,s,a]\Big). \tag{2}$$

In Equation 2, the second term equals $\sigma^2_{\mathrm{epi}}(s,a)$. The equality follows because J is uniform and $\mathbb{E}[\widetilde{G} \mid J = j, s, a] = q_j(s,a)$. DER penalizes only this between-model term. It remains risk-neutral to within-model return dispersion. Thus two actions with identical critic expectations receive identical DER scores even when their learned within-critic return distributions have different variances.

4 Method

DER comprises categorical return-law learning, expectation-level uncertainty summarization, robust action extraction, and frozen deployment. Its key design choice is to preserve outcome variation

inside each critic while applying robustness only to disagreement across critic expectations. The extractor selects the action with the largest worst-case local mean, after which its policy and ambiguity coefficient remain fixed throughout deployment. The following subsections present the extractor and its guarantees, the categorical learning objective, and the frozen deployment procedure.

4.1 Selective Robust Policy Extraction

The analysis follows the deployed computation and separates three objects. The categorical laws describe outcome variation, critic expectations define plausible mean values, and a temporal residual represents disagreement not explained by the frozen ensemble. The guarantees apply to the policy extractor while the Bellman backup retains its standard categorical form.

For fixed (s, a), let $\mathbf{q} = (q_1, \ldots, q_M)^\top$, let $\mathbf{1}$ be the all-ones vector, and define $\mathbf{d} = (\mathbf{q} - \mu\mathbf{1})/\sqrt{M}$. The critic-induced local mean-ambiguity set is

$$\mathcal{U}_\beta(s, a) = \{\mu(s, a) + \mathbf{u}^\top \mathbf{d}(s, a) : \mathbf{u} \in \mathbb{R}^M,\ \|\mathbf{u}\|_2 \le \beta\}. \tag{3}$$

It contains scalar expected values generated by norm-bounded perturbations along observed disagreement coordinates. We use it as a data-dependent robust envelope over critic-implied mean values. The hierarchy therefore defines the nested deployment functional

$$\mathfrak{R}_\beta(Z_{1:M}(s, a)) = \inf_{v \in \mathcal{U}_\beta(s,a)} v. \tag{4}$$

Equation 4 first reduces each critic to its risk-neutral expectation; only then does the outer adversary act across plausible means. Consequently, within-critic aleatoric dispersion cannot enter the robustification term. The normalization by $\sqrt{M}$ makes $\|\mathbf{d}\|_2$ equal to the population standard deviation in Equation 1, so β is a dimensionless ambiguity price and $\beta\sigma_{\text{epi}}$ has the same units as expected return. At $\beta = 0$, the operator reduces exactly to ensemble-mean extraction. Increasing β changes an action only when its reduction in disagreement compensates for its loss in consensus value, as formalized next.

Theorem 1 (Critic-induced robust duality)**.** *For every state-action pair and $\beta \ge 0$,*

$$\inf_{v \in \mathcal{U}_\beta(s,a)} v = \mu(s, a) - \beta\sigma_{\text{epi}}(s, a). \tag{5}$$

Consequently, DER maximizes the worst local mean in Equation 3.

Proof sketch. Equation 1 gives $\|\mathbf{d}\|_2 = \sigma_{\text{epi}}$. The result is the support function of a Euclidean ball, attained by the perturbation antiparallel to $\mathbf{d}$. Appendix E gives the full proof.

Define

$$\widehat{\pi}_\beta(s) \in \arg\max_{a \in \mathcal{A}} L_\beta(s, a), \qquad L_\beta(s, a) = \mu(s, a) - \beta\sigma_{\text{epi}}(s, a). \tag{6}$$

Let $a_\mu(s) \in \arg\max_a \mu(s, a)$, using the same deterministic tie rule as DER. When the state is fixed, write $\mu_a = \mu(s, a)$ and $\sigma_a = \sigma_{\text{epi}}(s, a)$.

Proposition 1 (Selective robustification)**.** *For $\beta > 0$, if DER selects $b \ne a_\mu(s)$, then*

$$0 \le \mu(s, a_\mu) - \mu(s, b) \le \beta\big[\sigma_{\text{epi}}(s, a_\mu) - \sigma_{\text{epi}}(s, b)\big]. \tag{7}$$

Thus DER can leave the mean-greedy action only for an action with no larger epistemic disagreement, and only when its uncertainty reduction offsets its consensus-value loss.

This exact switching condition makes the intervention state dependent. When the mean winner is already no more uncertain than every competitor, DER and mean-greedy extraction coincide.

Theorem 2 (Pareto selectivity and finite ambiguity path)**.** *Fix a state and deterministic tie rule, and write $a_\beta = \widehat{\pi}_\beta(s)$. Then the following statements hold.*

1. *For $\beta > 0$, a_β is mean-uncertainty Pareto admissible, meaning no action b satisfies both $\mu_b \ge \mu_{a_\beta}$ and $\sigma_b \le \sigma_{a_\beta}$, with at least one strict inequality.*
2. *If $0 \le \beta_1 < \beta_2$, then $\sigma_{a_{\beta_2}} \le \sigma_{a_{\beta_1}}$ and $\mu_{a_{\beta_2}} \le \mu_{a_{\beta_1}}$.*

3. *The selected action is piecewise constant in β. It can change only at a nonnegative pairwise threshold*

$$\beta_{a,b}(s) = \frac{\mu(s,a) - \mu(s,b)}{\sigma_{\mathrm{epi}}(s,a) - \sigma_{\mathrm{epi}}(s,b)} \tag{8}$$

for actions with unequal disagreement.

Proof sketch. A dominated action cannot maximize the linear scalarization $\mu - \beta\sigma$. Adding the optimality inequalities at β_1 and β_2 proves monotone movement toward lower disagreement and weakly lower consensus value. Finally, the upper envelope of finitely many affine scores changes only at pairwise intersections. The theorem turns β into an explicit price of ambiguity. DER traces supported points of the mean-epistemic Pareto frontier rather than applying indiscriminate pessimism.

Decision-path interpretation. For a fixed state, every action defines a line with intercept μ_a and slope $-\sigma_a$ as a function of β. DER selects the upper envelope of these lines. An action strictly dominated in both consensus value and disagreement can never appear on this envelope, while each actual switch has the threshold in Equation 8. This geometry explains why robustification is selective rather than a uniform downward adjustment. Actions with equal disagreement retain their mean ordering, and states with a sufficiently separated mean winner need not change at all. It also yields an auditable deployment path because the score, the selected action, and every possible switching point are determined by the same two action-level statistics.

Theorem 3 (Critic-perturbation action certificate)**.** *Let $\widetilde{q}_j(s,a)$ be perturbed critic expectations satisfying $\max_j |\widetilde{q}_j(s,a) - q_j(s,a)| \le \varepsilon$ for every action. Let $\widetilde{\mu}$, $\widetilde{\sigma}_{\mathrm{epi}}$, and $\widetilde{L}_\beta$ be recomputed by Equations 1 and 6. Then*

$$|\widetilde{L}_\beta(s,a) - L_\beta(s,a)| \le (1+\beta)\varepsilon. \tag{9}$$

Let a_β be the unique DER action and $g_\beta = L_\beta(s,a_\beta) - \max_{b \neq a_\beta} L_\beta(s,b)$. If $g_\beta > 2(1+\beta)\varepsilon$, the selected action is unchanged by every admissible critic perturbation.

Proof sketch. Centering is an orthogonal projection, so both the ensemble mean and standard deviation change by at most ε. Equation 9 follows by the triangle inequality; a score gap larger than twice this deviation preserves the maximizer. The result converts the deployed robust-score margin into a deterministic stability certificate without treating ensemble members as independent posterior draws. It also separates two notions that are otherwise easy to conflate. The quantity σ_{epi} ranks actions under the fitted ensemble, whereas ε describes a hypothetical perturbation of that fitted ensemble. The certificate therefore concerns decision invariance and does not require a probabilistic interpretation of either quantity.

We next characterize performance under deployment shift. Let $Q^\star_{\mathrm{T}}$ be the optimal action value in a future deployment MDP, where the subscript T denotes that deployment environment rather than a time index.

Assumption 1 (Independently bounded deployment residual)**.** *There exist a nonnegative shift residual $\Delta_{\mathrm{T}}(s,a)$ and a finite constant $\bar{\Delta}$, specified independently of the fitted critics, such that, simultaneously for all actions at each deployment state under consideration,*

$$|Q^\star_{\mathrm{T}}(s,a) - \mu(s,a)| \le \beta\sigma_{\mathrm{epi}}(s,a) + \Delta_{\mathrm{T}}(s,a). \tag{10}$$

The deployment class satisfies $\sup_{s,a} \Delta_{\mathrm{T}}(s,a) \le \bar{\Delta}$.

Theorem 4 (Shift-aware action regret)**.** *Let $a^\star_{\mathrm{T}}(s) \in \arg\max_a Q^\star_{\mathrm{T}}(s,a)$. If Assumption 1 holds at s, then*

$$\begin{aligned} &Q^\star_{\mathrm{T}}(s,a^\star_{\mathrm{T}}) - Q^\star_{\mathrm{T}}(s,\widehat{\pi}_\beta(s)) \\ &\quad \le 2\beta\sigma_{\mathrm{epi}}(s,a^\star_{\mathrm{T}}) + \Delta_{\mathrm{T}}(s,a^\star_{\mathrm{T}}) + \Delta_{\mathrm{T}}(s,\widehat{\pi}_\beta(s)). \end{aligned} \tag{11}$$

Proof sketch. Apply the upper side of Equation 10 to the optimal action, the lower side to the selected action, and insert the ordering of their robust scores. The selected action's epistemic term cancels, leaving uncertainty at the truly optimal action and two explicit shift residuals.

The local statement extends to the full frozen policy. Let $V^\star_{\mathrm{T}}$ and $V^{\widehat{\pi}_\beta}_{\mathrm{T}}$ denote discounted deployment values. Define $\bar{\sigma}^\star = \sup_s \sigma_{\mathrm{epi}}(s, a^\star_{\mathrm{T}}(s))$.

Theorem 5 (Frozen-policy performance under deployment shift)**.** *Suppose deployment rewards are bounded, $\bar{\sigma}^\star$ is finite, Assumption 1 holds over the deployment state set, and $0 \leq \gamma < 1$. Then*

$$\left\| V_{\mathrm{T}}^\star - V_{\mathrm{T}}^{\widehat{\pi}_\beta} \right\|_\infty \leq \frac{2\beta\bar{\sigma}^\star + 2\bar{\Delta}}{1-\gamma}. \tag{12}$$

The Bellman resolvent propagates the action regret of Theorem 4 to sequential deployment. When $\Delta_{\mathrm{T}} = 0$, Equations 11 and 12 isolate the epistemic component in Equation 10. The resulting bound keeps ensemble disagreement and deployment shift explicit. All terms in Equation 10 are action-value quantities under the deployment MDP. The function Δ_{T} absorbs temporal error shared by all critics, while $\beta\sigma_{\mathrm{epi}}$ records error visible through their disagreement. Under the independently bounded residual, the theorem explicitly separates uncertainty visible to the ensemble from temporal error shared by its critics.

4.2 Learning Portfolio Return Laws

Each member is a categorical critic with fixed atoms $\{z_i\}_{i=1}^N$ and probabilities $p_{\theta_j,i}(s,a)$. Its expectation is

$$q_j(s,a) = \sum_{i=1}^N z_i p_{\theta_j,i}(s,a). \tag{13}$$

Equation 13 is the scalar summary used by DER. Appendix E derives the exact C51 projection identity in Equation 17, connecting the learned law with this summary. The within-critic law is retained during learning, while its expectation supplies the coherent risk-neutral summary used for robust extraction. Categorical Bellman targets preserve the multi-outcome structure during representation learning, while Equation 17 shows how that structure maps to the extracted mean on the bounded support. The architecture therefore separates the role of the return law during learning from the role of cross-member disagreement during action selection. Members use different random initializations and minibatch orderings but the same offline transition set. Each target distribution is obtained from the target network's greedy next action and projected onto the common categorical support (Bellemare et al., 2017). This choice preserves information about the return law while leaving epistemic disagreement identifiable through the expectations of independently fitted members.

4.3 Frozen Deployment Procedure

After training, Equation 1 is evaluated for all eleven actions. A coefficient β is selected only on the validation calendar and then frozen. Training weights and target networks are also frozen before the first test date. The method performs no test-time update, regime detection, rolling refit, and no look-ahead selection. This deployment stage is the epistemic component of DER; the preceding categorical ensemble supplies its distributional value representation. Their separation aligns the implementation with Theorem 1.

4.4 Training and Deployment Objective

For critic member j, categorical temporal-difference learning minimizes

$$\ell_j(\theta_j) = \mathbb{E}_{(s,a,r,s',\delta)\sim\mathcal{D}} \left[-\sum_{i=1}^N m_{j,i}(s,a,r,s',\delta) \log p_{\theta_j,i}(s,a) \right], \tag{14}$$

where $\delta \in \{0,1\}$ marks termination and $\mathbf{m}_j = (m_{j,1},\ldots,m_{j,N})$ is the projected C51 target, held fixed while updating θ_j. Equation 16 in Appendix C gives its exact construction. The deployed decision then solves Equation 6. Equations 14 and 6 express the complete method. The former learns plausible return laws; the latter performs ambiguity-averse selection over their expected values.

Algorithm 1 in Appendix C summarizes the complete train-select-freeze pipeline.

5 Experiments

Our experiments address three questions that mirror the empirical argument of the paper. The first asks whether a portfolio policy fixed before deployment remains competitive across unseen regimes,

the second examines temporal and risk-adjusted reliability, and the third identifies which uncertainty target should be robustified under matched critic banks. The following setup is shared by all three analyses.

5.1 SETUP

Data and temporal separation. The daily A-share panel contains 790 securities, 3886 trading days, and 29 past-only price-volume features from 2010-01-04 through 2025-12-31. The universe comprises the union of historical quarterly CSI 300 constituents. Using only information available through 2019, point-in-time liquidity filtering forms two disjoint pools of 256 securities. Training uses 2010–2019, validation uses calendar 2020, and the untouched test comprises calendars 2021–2025. Each repeat is trained on the historical split; checkpoints and hyperparameters are selected on 2020, after which all weights, normalization statistics, pools, and decision rules are frozen. There is no rolling refit and no test-time adaptation.

State, action, and transition interface. The state combines past-only market features with pre-decision holdings and portfolio value. The eleven long-only actions are daily low volatility, a twenty-day hold version, and nine ridge-ranked portfolios obtained by crossing prediction horizons $\{5, 20, 60\}$ with cardinalities $\{10, 20, 50\}$. Offline trajectories use a twenty-day decision horizon, discount $\gamma = 0.98$, two trajectory initializations, four particles, and fixed offsets $\{0, 5, 10, 15\}$. Because the simulator is price-taking, the same reachable state can be evaluated under every discrete portfolio option, giving all methods an identical counterfactual action interface.

Execution and reward. The initial portfolio value is 10^7, weights are nonnegative and sum to at most one, and an action chosen from close information at day t executes at the adjusted OHLC average on day $t + 1$. The simulator imposes a 10% participation cap, 3 basis points commission and 5 basis points slippage on each side, plus the historical sell-side stamp tax (10 basis points before 2023-08-28 and 5 thereafter). The primary metric is annualized active return relative to the same-pool daily low-volatility benchmark, denoted Active ARR; CSI300-relative Active ARR is a secondary external-market diagnostic. The controlled same-pool benchmark removes much of the broad long-only market component while retaining allocation and timing differences. Appendix A documents delistings, price alignment, and the complete simulator contract.

Comparators and aggregation. We compare DER with Dueling DQN (Wang et al., 2016), CQL (Kumar et al., 2020), and Double DQN (van Hasselt et al., 2016). We additionally include interface-matched adaptations of three recent published designs: ROI-DICE-Finite (Kim et al., 2024), Decision Mamba-Finite (Lv et al., 2024), and EarnMore-Option (Zhang et al., 2024). The suffixes identify interface-matched adaptations at the finite-action and option levels within the common experimental protocol. Neural methods share the trajectory data, training budget, hidden width where applicable, and validation checkpoint rule. Method-specific architectures and losses follow their definitions. Let $X^{(b)}_{r,y,p}$ denote Active ARR for repeat r, calendar y, pool p, and benchmark b. With $R = 5$ common repeats (indices 0–4 for every displayed method) and $P = 2$ pools, the annual repeat-level estimand and five-year repeat-level estimand are

$$\bar{X}^{(b)}_{r,y} = \frac{1}{P}\sum_{p=1}^{P} X^{(b)}_{r,y,p}, \qquad \bar{X}^{(b)}_{r} = \frac{1}{5}\sum_{y=2021}^{2025} \bar{X}^{(b)}_{r,y}. \tag{15}$$

Following Equation 15, year columns report the mean and sample standard deviation of $\bar{X}^{(b)}_{r,y}$ over r, while the final columns apply the same summaries to $\bar{X}^{(b)}_{r}$ under the two benchmarks. This order gives equal weight to the two pools and to each deployment calendar within a repeat. The repeat is the unit used for reported variability; years and pools are fixed deployment strata disclosed separately rather than treated as additional independent replicates. Across all tables, the best mean in each comparable column is marked in **bold**, and the second-best is underlined. Experiments used NVIDIA A100 GPUs with PyTorch 2.2.0; Appendix C gives the optimizer, target-update, support, and repeat settings. The validation sweep selects $\beta = 2$ from the prespecified candidate set using the 2020 metric; Appendix C records the archived selection configuration.

Table 1: Frozen-policy active return across deployment calendars. Year columns use the same-pool low-volatility benchmark; the final columns summarize the two stated benchmarks. Best in bold, second best underlined.

Method	2021	2022	2023	2024	2025	Controlled mean ± SD	CSI300 mean ± SD
DER	**14.20 ± 5.42**	2.91 ± 4.03	**0.15 ± 4.18**	**-8.51 ± 4.13**	**26.11 ± 5.41**	**6.97 ± 1.13**	**17.50 ± 1.13**
Dueling DQN	13.00 ± 6.54	9.20 ± 6.24	-1.27 ± 1.42	-15.43 ± 6.24	21.96 ± 1.93	5.49 ± 1.46	16.02 ± 1.46
Double DQN	6.06 ± 6.24	8.83 ± 3.63	-2.40 ± 1.84	-16.77 ± 3.90	23.86 ± 3.77	3.92 ± 1.91	14.44 ± 1.91
CQL	4.41 ± 3.71	**9.40 ± 4.91**	-1.67 ± 1.56	-14.00 ± 2.90	20.86 ± 4.94	3.80 ± 2.80	14.32 ± 2.80
ROI-DICE-Finite	10.92 ± 1.93	7.42 ± 3.92	-1.83 ± 2.60	-11.58 ± 2.93	22.04 ± 6.24	5.39 ± 1.34	15.92 ± 1.34
Decision Mamba-Finite	11.23 ± 2.85	8.88 ± 5.28	-2.44 ± 2.96	-9.22 ± 2.88	23.51 ± 6.04	6.39 ± 1.42	16.92 ± 1.42
EarnMore-Option	13.95 ± 5.60	3.08 ± 3.97	-2.80 ± 4.39	-18.39 ± 3.25	24.97 ± 3.82	4.16 ± 2.50	14.69 ± 2.50

Table 2: Risk-adjusted deployment diagnostics. Lower MDD is better. Best in bold, second best underlined.

Method	Sharpe	Sortino	MDD (%)	Calmar	Low-vol. IR
DER	**0.890 ± .066**	**1.386 ± .120**	14.59 ± .57	**1.648 ± .285**	0.472 ± .149
Dueling DQN	0.846 ± .040	1.308 ± .076	15.14 ± 1.12	1.565 ± .160	0.445 ± .085
Double DQN	0.776 ± .069	1.182 ± .117	15.10 ± .47	1.310 ± .236	0.303 ± .115
CQL	0.791 ± .107	1.197 ± .177	14.35 ± .54	1.410 ± .311	0.295 ± .206
ROI-DICE-Finite	0.874 ± .062	1.310 ± .104	**14.13 ± .30**	1.522 ± .270	**0.488 ± .042**
Decision Mamba-Finite	0.877 ± .079	1.344 ± .119	15.04 ± .54	1.579 ± .257	0.460 ± .133
EarnMore-Option	0.826 ± .087	1.258 ± .144	14.17 ± .40	1.580 ± .178	0.379 ± .192

5.2 Frozen Portfolio Performance

To answer the first research question, we compare policies fixed before deployment across all five unseen calendars. Table 1 reports every test year. DER has the largest five-year controlled mean among the six displayed external baselines, exceeding the strongest comparator, Decision Mamba-Finite, by 0.58 percentage points. Its CSI300-relative Active ARR averages 17.50 percent. DER leads in 2021, 2023, 2024, and 2025, while CQL leads in 2022. The five repeats quantify training variability under a common protocol, with every held-out calendar and the corresponding repeat dispersion reported explicitly.

5.3 Financial Reliability Across Regimes

To answer the second research question, we summarize temporal, downside, and risk-adjusted reliability. The following cross-period summaries are deterministic functions of the annual results in Table 1. Across the five test calendars, DER has positive mean Active ARR in four years and the highest worst-calendar Active ARR (-8.51%). It leads in four of five calendars, with annual margins of 0.25, 1.42, 0.71, and 1.14 percentage points in 2021, 2023, 2024, and 2025, respectively. Its five-year mean is 0.58 points higher than Decision Mamba-Finite, complementing its lead in four of five deployment calendars.

Risk-adjusted diagnostics. For each repeat, Table 2 averages frozen-Test risk metrics over the same five years and two pools, then reports mean and sample SD over five repeats. DER ranks first in Sharpe, Sortino, and Calmar; ROI-DICE-Finite has the lowest maximum drawdown and the highest low-volatility information ratio (IR). These diagnostics complement Active ARR by characterizing return quality, downside exposure, and benchmark-relative efficiency. Repeat-level reporting is especially important because Sharpe-ratio estimates are sensitive to sampling variation (Lo, 2002). The two benchmark columns in Table 1 serve different purposes. The same-pool low-volatility benchmark controls for the investable opportunity set faced by each policy, whereas CSI300 provides a common external market reference. DER ranks first under both, so the main ordering does not depend on choosing one benchmark convention. Likewise, its leading Sharpe, Sortino, and Calmar show that the active-return ranking is accompanied by competitive risk-adjusted summaries rather than being produced solely by a single high-return calendar.

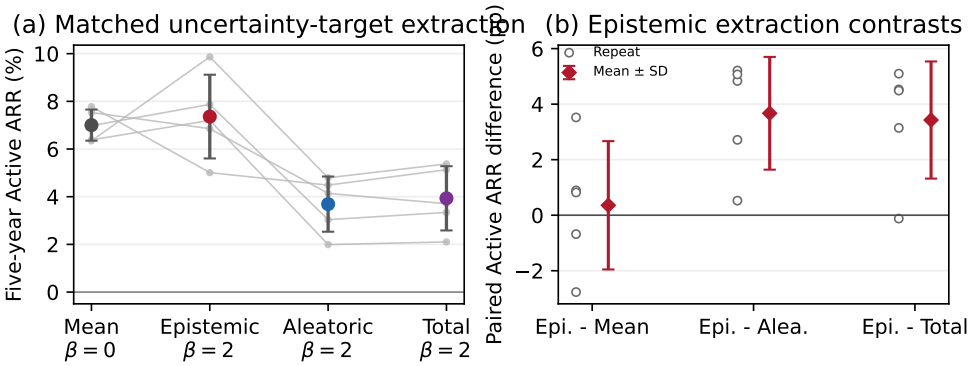

Figure 1: Matched uncertainty-target ablation in an independently retrained five-repeat control cohort. In panel (a), thin lines connect extractors sharing the same C51 critics. In panel (b), open circles are individual repeat contrasts, while red diamonds and bars show mean ± sample SD.

5.4 Which Uncertainty Should Be Robustified?

To answer the third research question, we hold an independently retrained C51 critic bank fixed within each repeat and change only the uncertainty component used by the extractor. Figure 1 compares ensemble-mean extraction with fixed-$\beta = 2$ penalties based on cross-critic expected-value disagreement (epistemic), average within-critic outcome variance (aleatoric), or their sum (total).

At the same ambiguity coefficient, epistemic extraction exceeds aleatoric and total-variance extraction by 3.67 and 3.43 percentage points, with paired advantages in five of five and four of five repeats, respectively. Relative to ensemble-mean extraction, it gains 0.36 points with three of five wins. This pattern supports the uncertainty-target choice under the stated risk-neutral mandate. Appendix B reports the remaining score-level audits.

6 Conclusion

We introduced Distributional Epistemic Robustification (DER) for frozen portfolio decisions under changing market conditions. Against the external baselines in the primary comparison, DER achieved the highest aggregate and worst-calendar Active ARR, together with leading risk-adjusted metrics. In the matched uncertainty-target control, epistemic extraction exceeds aleatoric and total-variance extraction by 3.67 and 3.43 percentage points, respectively, and exceeds ensemble-mean extraction by 0.36 points with three of five repeat-level wins. Together, these results support separating outcome variation from model disagreement when constructing an ambiguity-aware extractor for the stated risk-neutral mandate.

AI USE STATEMENT

Generative AI tools assisted with research development, technical implementation, analysis, and manuscript preparation. The authors reviewed and verified all AI-assisted work and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

Appendix A specifies the data construction and evaluation contract, Appendix C records the implementation and archived selection configuration, and Appendix E provides the complete proofs. The submission code contains the executable pipelines and result provenance; all reported test tables exclude validation outcomes.

REFERENCES

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified Q-ensemble. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 20(4):39–69, 2017.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, 2017.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

Yash Chandak, Scott M. Jordan, Georgios Theocharous, Martha White, and Philip S. Thomas. Towards safe policy improvement for non-stationary mdps. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jiayi Chen, Jing Li, and Guiling Wang. MARS: A meta-adaptive reinforcement learning framework for risk-aware multi-agent portfolio management. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 20092–20099, 2026. doi: 10.1609/aaai.v40i24.39095.

Abhilash Reddy Chenreddy and Erick Delage. Epistemic robust offline reinforcement learning. *arXiv preprint arXiv:2604.07072*, 2026.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning*, 2018a.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence*, 2018b.

Hannes Eriksson, Debabrota Basu, Mina Alibeigi, and Christos Dimitrakakis. Sentinel: Taming uncertainty with ensemble based distributional reinforcement learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 631–640, 2022.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062, 2019.

Kamyar Ghasemipour, Shixiang Shane Gu, and Ofir Nachum. Why so pessimistic? estimating uncertainties for offline rl through ensembles, and why their independence matters. In *Advances in Neural Information Processing Systems*, 2022.

Mohammad Ghavamzadeh, Marek Petrik, and Yinlam Chow. Safe policy improvement by minimizing robust baseline regret. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*, 2017.

Woosung Kim, Hayeong Lee, Jongmin Lee, and Byung-Jun Lee. ROIDICE: Offline return on investment maximization for efficient decision making. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pp. 3652–3661, 2019.

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

Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Dan Wang, Zhaoran Wang, and Jian Guo. Finrl-meta: Market environments and benchmarks for data-driven financial reinforcement learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 35, 2022.

Andrew W. Lo. The statistics of sharpe ratios. *Financial Analysts Journal*, 58(4):36–52, 2002.

Qi Lv, Xiang Deng, Gongwei Chen, Michael Yu Wang, and Liqiang Nie. Decision mamba: A multi-grained state space model with self-evolution regularization for offline RL. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001.

John Moody, Lizhong Wu, Yuansong Liao, and Matthew Saffell. Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, 17(5–6):441–470, 1998. doi: 10.1002/(SICI)1099-131X(1998090)17:5/6⟨441::AID-FOR707⟩3.0.CO;2-\#.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *International Conference on Machine Learning*, 2010.

Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Hui Niu, Siyuan Li, and Jian Li. Metatrader: An reinforcement learning approach integrating diverse policies for portfolio optimization. In *ACM International Conference on Information and Knowledge Management*, pp. 1573–1583, 2022. doi: 10.1145/3511808.3557363.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in Neural Information Processing Systems*, 2016.

Molei Qin, Shuo Sun, Wentao Zhang, Haochong Xia, Xinrun Wang, and Bo An. Earnhft: Efficient hierarchical reinforcement learning for high frequency trading. In *AAAI Conference on Artificial Intelligence*, 2024.

Shuo Sun, Molei Qin, Wentao Zhang, Haochong Xia, Chuqiao Zong, Jie Ying, Yonggang Xie, Lingxuan Zhao, Xinrun Wang, and Bo An. Trademaster: A holistic quantitative trading platform empowered by reinforcement learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023a.

Shuo Sun, Xinrun Wang, Wanqi Xue, Xiaoxuan Lou, and Bo An. Mastering stock markets with efficient mixture of diversified trading experts. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2109–2119, 2023b.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence policy improvement. In *International Conference on Machine Learning*, pp. 2380–2388, 2015.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI Conference on Artificial Intelligence*, 2016.

Rundong Wang, Hongxin Wei, Bo An, Zhouyan Feng, and Jun Yao. Commission fee is not enough: A hierarchical reinforced framework for portfolio management. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 626–633, 2021a.

Zhicheng Wang, Biwei Huang, Shikui Tu, Kun Zhang, and Lei Xu. Deeptrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 643–650, 2021b.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning*, 2016.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

Jialong Wu, Haixu Wu, Zihan Qiu, Jianmin Wang, and Mingsheng Long. Supported policy optimization for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yupeng Wu, Wenyun Li, Wenjie Huang, and Chin Pang Ho. DRL-ORA: Distributional reinforcement learning with online epistemic risk adaptation. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 7484–7503, 2026.

Yunan Ye, Hengzhi Pei, Boxin Wang, Pin-Yu Chen, Yada Zhu, Ju Xiao, and Bo Li. Reinforcement-learning based portfolio management with augmented asset movement prediction states. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 1112–1119, 2020.

Haochen Yuan, Minting Pan, Yunbo Wang, Siyu Gao, and Xiaokang Yang. Metatrader: Learning to generalize rl trading policies beyond offline data. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 28023–28031, 2026. doi: 10.1609/aaai.v40i33.40027.

Wentao Zhang, Yilei Zhao, Shuo Sun, Jie Ying, Yonggang Xie, Zitao Song, Xinrun Wang, and Bo An. Reinforcement learning with maskable stock representation for portfolio management in customizable stock pools. In *Proceedings of the ACM Web Conference*, pp. 187–198, 2024. doi: 10.1145/3589334.3645615.

Chuqiao Zong, Chaojie Wang, Molei Qin, Lei Feng, Xinrun Wang, and Bo An. Macrohft: Memory augmented context-aware reinforcement learning on high frequency trading. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4712–4721, 2024.

A Benchmark Datasets and Evaluation Protocol

A.1 Data Source and Access

Daily prices and historical CSI 300 constituent snapshots were obtained through the BaoStock client and converted by the acquisition scripts in the submission code. The frozen panel hash below, the instrument manifest, acquisition code, and aggregate results support reconstruction and integrity

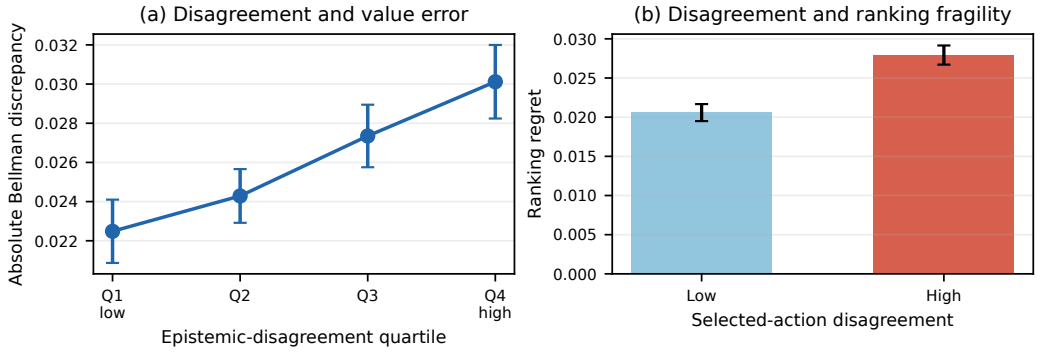

Figure 2: Bellman discrepancy and ranking regret across critic-disagreement strata in the frozen-policy audit. Points and bars report the mean and sample standard deviation across five repeats.

checks. Raw market data are not redistributed because the client's BSD software license does not establish redistribution rights for the underlying records. The reported panel hash identifies the exact experimental data object, while the acquisition scripts and instrument manifest define the reconstruction path under the provider's current access terms.

A.2 Frozen Evaluation Contract

The panel hash is `7fd832dee28890bd46b99d228b95c68748042d7c4c479f64b094704c5b0dbe37`. It spans 2010-01-04 through 2025-12-31. Pool membership is fixed point-in-time at the end of 2019. The initial portfolio value is 10^7. Portfolio weights are long-only, nonnegative, and sum to at most one. Trading uses a ten percent participation cap, three basis points commission on both sides, five basis points slippage on both sides, and sell-side stamp tax of ten basis points before 2023-08-28 and five basis points thereafter. Delisting recovery and the cash rate are zero.

A decision at close t executes at the adjusted OHLC average on $t+1$. Trajectory construction uses two initializations, four particles, offsets $0, 5, 10, 15$, a twenty-day horizon, and $\gamma = 0.98$. The eleven actions are low volatility, low volatility held for twenty days, and ridge portfolios for prediction horizons $5, 20, 60$ crossed with top-$10, 20, 50$ cardinalities.

B Decision-Mechanism Audit

The audit evaluates the frozen 2021–2025 transitions without affecting model or hyperparameter selection. For each state-action pair, absolute Bellman discrepancy is $|\mu(s,a) - y(s,a)|$, where $y(s,a) = r + \gamma(1-\delta)\max_b \mu(s^+, b)$. State-action pairs are grouped within each repeat by quartiles of $\sigma_{\text{epi}}(s,a)$. Ranking regret is $\max_b y(s,b) - y(s,a_\mu)$ for the ensemble-mean action $a_\mu = \arg\max_a \mu(s,a)$; its low and high groups use the lower and upper quartiles of $\sigma_{\text{epi}}(s,a_\mu)$. Metrics are averaged within each repeat before reporting the across-repeat mean and sample standard deviation. From the lowest to the highest disagreement quartile, absolute Bellman discrepancy and ranking regret increase by factors of 1.34 ± 0.06 and 1.36 ± 0.08, respectively. Table 3 measures how DER changes the deployed decision relative to ensemble-mean extraction from the same trained critics. A switch is a deployment state in which the two rules choose different actions. Consensus-value sacrifice is the ensemble-mean score of the mean-greedy action minus that of the DER action. Epistemic-disagreement reduction is the corresponding decrease in cross-critic standard deviation. A Pareto violation would select an action with lower consensus value and no lower disagreement.

C Implementation Details

Algorithm 1 records the complete train-select-freeze procedure used in every repeat.

Table 3: Decision-mechanism audit with critic-score changes conditioned on an action switch.

Metric	Aggregate
Action-switch rate (%)	49.38 ± 9.95
Consensus-value sacrifice ($\times 10^{-3}$)	4.67 ± 1.17
Epistemic-disagreement reduction ($\times 10^{-3}$)	4.46 ± 1.04
Pareto-violation rate (%)	0.00 ± 0.00

Algorithm 1 Distributional Epistemic Robustification

Require: Offline transitions $\mathcal{D}$, validation set $\mathcal{V}$, actions $\mathcal{A}$, ensemble size M

for member $j = 1, \dots, M$ **do**
  Train categorical critic Z_j on $\mathcal{D}$ by Equation 14; select its checkpoint on $\mathcal{V}$
end for
Select β using only $\mathcal{V}$; freeze $\{Z_j\}_{j=1}^{M}$ and β
for each deployment state s **do**
  for each $a \in \mathcal{A}$ **do**
    $q_j(s,a) \leftarrow \mathbb{E}[Z_j(s,a)]$ for $j = 1, \dots, M$
    Compute $\mu(s,a)$ and $\sigma_{\mathrm{epi}}(s,a)$ using Equation 1
    $L_\beta(s,a) \leftarrow \mu(s,a) - \beta\sigma_{\mathrm{epi}}(s,a)$
  end for
  Execute $\arg\max_{a\in\mathcal{A}} L_\beta(s,a)$ with a fixed tie rule
end for

C.1 Shared Training Configuration

DER and the conventional neural value baselines use two hidden layers of width 128 with ReLU activations. Training runs for forty epochs over the fixed offline transition table, reshuffled each epoch into batches of 128 states. Adam uses learning rate 3×10^{-4}, coefficients 0.9 and 0.999, numerical constant 10^{-8}, and zero weight decay. Gradients are clipped at ten, the discount is $\gamma = 0.98$, and the target network is copied every two epochs. All neural checkpoints are selected by the two-pool mean Active ARR on the 2020 validation calendar; no replay buffer is used.

C.2 Our Method's Hyperparameters

DER uses three independently initialized C51 critics, each with fifty-one atoms on $[-1, 1]$. The ambiguity coefficient is selected from $\{0, 0.1, 0.25, 0.5, 1, 2\}$ using the 2020 validation metric. The deployed configuration uses the original action-state representation and selected coefficient $\beta = 2$. The submission provenance maps this choice to the archived validation-selection record. Neither the representation nor the coefficient changes during testing. Let $\bar{\theta}_j$ denote member j's target-network parameters. The C51 target used in Equation 14 is

$$a_j^+(s') \in \arg\max_{b\in\mathcal{A}} \sum_{i=1}^{N} z_i p_{\bar{\theta}_j,i}(s', b),$$
$$Y_j = r + \gamma(1-\delta) Z_{\bar{\theta}_j}(s', a_j^+(s')), \qquad m_{j,i} = \left[\Pi_{\mathcal{Z}} \operatorname{Law}(Y_j)\right]_i, \tag{16}$$

using the same fixed tie rule as deployment.

C.3 Baseline Implementations and Hyperparameters

All comparators are implemented in the common evaluation codebase from the cited algorithm descriptions. For the three recent portfolio designs, public architectural specifications define the interface-matched adaptations. These adaptations preserve each method's defining objective and architecture while standardizing data access, action choices, transaction costs, checkpoint selection, and evaluation conditions. Canonical result records store the runtime configuration and SHA-256 hashes of the executed local source files, which provide the exact implementation version used in the paper. Dueling DQN and Double DQN use the shared scalar-critic configuration. CQL uses

coefficient one. ROI-DICE-Finite uses a finite-action divergence objective with coefficient one and downside quantile 0.5. Decision Mamba-Finite uses token width five, hidden width 128, and self-evolution weight 0.10 multiplied by training progress. EarnMore-Option uses mask fraction 0.20 and reconstruction weight 0.10.

C.4 Repeat Protocol

Every method in the primary comparison uses the common completed repeat indices 0–4. Each neural repeat contains three independently initialized ensemble members. All reported test results are generated only after validation-based checkpoint selection and policy freezing.

D Computational Resources

Experiments used NVIDIA A100-SXM4 GPUs with Python 3.12.13, PyTorch 2.2.0, CUDA 12.1, NumPy 1.26.4, and pandas 3.0.1.

E Theorem Proofs

Categorical mean coherence. Let $\Pi_{\mathcal{Z}}$ be the C51 projection onto equally spaced atoms with endpoints $v_{\min}$ and $v_{\max}$. For any integrable target return Y, let Y^{Π} have distribution $\Pi_{\mathcal{Z}}\,\mathrm{Law}(Y)$. Then

$$\mathbb{E}[Y^{\Pi}] = \mathbb{E}[\mathrm{clip}(Y, v_{\min}, v_{\max})]. \tag{17}$$

Equation 17 gives the exact projected expectation on the common categorical support.

Derivation. For a scalar $y \in [z_i, z_{i+1}]$, categorical projection assigns weights $(z_{i+1} - y)/(z_{i+1} - z_i)$ and $(y - z_i)/(z_{i+1} - z_i)$ to the adjacent atoms. Their weighted mean is exactly y. Values below $v_{\min}$ and above $v_{\max}$ are assigned to the corresponding endpoint. Taking expectation over Y therefore gives Equation 17.

Proof of Theorem 1. By definition,

$$\|\mathbf{d}\|_2^2 = \frac{1}{M}\sum_{j=1}^{M}(q_j - \mu)^2 = \sigma_{\mathrm{epi}}^2.$$

For any feasible $\mathbf{u}$, Cauchy-Schwarz gives $\mathbf{u}^\top \mathbf{d} \geq -\|\mathbf{u}\|_2\|\mathbf{d}\|_2 \geq -\beta\sigma_{\mathrm{epi}}$. If $\sigma_{\mathrm{epi}} > 0$, equality is attained by $\mathbf{u} = -\beta\mathbf{d}/\|\mathbf{d}\|_2$. If it is zero, every perturbation term vanishes. Adding μ proves Equation 5. □

Proof of Proposition 1. Let $a_\mu = a_\mu(s)$ and suppose DER selects $b \neq a_\mu$. Mean-greedy optimality gives $\mu_{a_\mu} - \mu_b \geq 0$. DER optimality gives $\mu_b - \beta\sigma_b \geq \mu_{a_\mu} - \beta\sigma_{a_\mu}$. Rearranging yields Equation 7. Its nonnegative left side also implies $\sigma_b \leq \sigma_{a_\mu}$. □

Proof of Theorem 2. Fix s and suppress it from notation. For the first claim, suppose $\beta > 0$ and an action b weakly improves the mean and weakly reduces disagreement relative to a_β, with at least one strict inequality. Then $\mu_b - \beta\sigma_b > \mu_{a_\beta} - \beta\sigma_{a_\beta}$, contradicting the definition of a_β.

For the second claim, write $i = a_{\beta_1}$ and $j = a_{\beta_2}$. Optimality at the two coefficients gives

$$\mu_i - \beta_1\sigma_i \geq \mu_j - \beta_1\sigma_j, \qquad \mu_j - \beta_2\sigma_j \geq \mu_i - \beta_2\sigma_i.$$

Therefore

$$\beta_1(\sigma_i - \sigma_j) \leq \mu_i - \mu_j \leq \beta_2(\sigma_i - \sigma_j).$$

Because $\beta_2 > \beta_1$, these inequalities require $\sigma_i \geq \sigma_j$, and then imply $\mu_i \geq \mu_j$.

Finally, every action score is affine in β. The maximizer of finitely many affine functions is constant between their pairwise intersections. Equating the scores of actions a and b with unequal disagreement gives Equation 8; only nonnegative intersections lie in the admissible domain. The fixed tie rule makes the selected action well defined at an intersection. □

Proof of Theorem 3. Fix an action and collect its member expectations in vectors $\mathbf{q}$ and $\widetilde{\mathbf{q}}$. Let $P = I_M - \mathbf{1}\mathbf{1}^\top/M$ be the centering projection, where I_M is the M-dimensional identity. The perturbation assumption gives $\|\widetilde{\mathbf{q}} - \mathbf{q}\|_2 \le \sqrt{M}\varepsilon$, hence

$$|\widetilde{\mu} - \mu| \le \varepsilon, \qquad |\widetilde{\sigma}_{\mathrm{epi}} - \sigma_{\mathrm{epi}}| = \frac{1}{\sqrt{M}}\big|\|P\widetilde{\mathbf{q}}\|_2 - \|P\mathbf{q}\|_2\big| \le \varepsilon,$$

because P is nonexpansive. Therefore $|\widetilde{L}_\beta - L_\beta| \le (1+\beta)\varepsilon$. The winning score can fall by at most this amount and any competing score can rise by at most the same amount. A gap above $2(1+\beta)\varepsilon$ therefore preserves the unique maximizer. □

Interpretation of deployment shift. In Equation 10, the ensemble term represents disagreement visible to DER, while Δ_{T} accounts for temporal change, shared ensemble bias, and error outside that observed disagreement. Keeping these components separate prevents the deployment analysis from attributing every source of shift to the ensemble.

Proof of Theorem 4. Fix s, write $\widehat{a} = \widehat{\pi}_\beta(s)$, and abbreviate $a^\star = a^\star_{\mathrm{T}}(s)$, $\Delta_a = \Delta_{\mathrm{T}}(s,a)$. The two sides of Equation 10 give

$$Q^\star_{\mathrm{T}}(s,a^\star) \le \mu_{a^\star} + \beta\sigma_{a^\star} + \Delta_{a^\star}$$

and

$$Q^\star_{\mathrm{T}}(s,\widehat{a}) \ge \mu_{\widehat{a}} - \beta\sigma_{\widehat{a}} - \Delta_{\widehat{a}}.$$

Equation 6 gives $\mu_{\widehat{a}} - \beta\sigma_{\widehat{a}} \ge \mu_{a^\star} - \beta\sigma_{a^\star}$. Subtracting the lower bound from the upper bound and inserting this ordering proves Equation 11. □

Proof of Theorem 5. Let $\pi = \widehat{\pi}_\beta$ and define the optimal-action residual $e(s) = V^\star_{\mathrm{T}}(s) - Q^\star_{\mathrm{T}}(s,\pi(s))$. Theorem 4 gives $0 \le e(s) \le B := 2\beta\bar{\sigma}^\star + 2\bar{\Delta}$. Let $\mathcal{T}^\pi_{\mathrm{T}}$ and P^π_{T} denote the deployment Bellman operator and transition kernel under π. Using $V^\pi_{\mathrm{T}} = \mathcal{T}^\pi_{\mathrm{T}} V^\pi_{\mathrm{T}}$ and $Q^\star_{\mathrm{T}}(s,\pi(s)) = (\mathcal{T}^\pi_{\mathrm{T}} V^\star_{\mathrm{T}})(s)$,

$$V^\star_{\mathrm{T}} - V^\pi_{\mathrm{T}} = e + \gamma P^\pi_{\mathrm{T}}(V^\star_{\mathrm{T}} - V^\pi_{\mathrm{T}}).$$

Taking the sup norm and using that P^π_{T} is nonexpansive gives $\|V^\star_{\mathrm{T}} - V^\pi_{\mathrm{T}}\|_\infty \le B + \gamma\|V^\star_{\mathrm{T}} - V^\pi_{\mathrm{T}}\|_\infty$. Rearrangement proves Equation 12. □

F LIMITATIONS

The evaluation focuses on a strictly time-ordered A-share setting with two fixed pools and a finite portfolio action interface. Extending the same frozen-policy protocol to additional asset classes, market microstructures, and action spaces would further characterize its scope.

The present study uses expectation-level disagreement from independently initialized categorical critics and a Euclidean local mean-ambiguity geometry. Alternative ensemble constructions, ambiguity geometries, and validation criteria are natural extensions.