DECISION CONFUSION: LEARNING MORE CAN DECIDE WORSE UNDER WEAK SIGNAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Expressive learned policies are increasingly used in sequential decision-making because they can discover complex representations and decision rules directly from experience. However, when decision-relevant predictive signals are weak relative to noise, additional flexibility may increase finite-sample estimation error without providing commensurate gains in exploitable structure. We investigate how flexibility in policy learning affects performance under weak signals. We argue that the appropriate extent of learning is governed by predictive information: under weak signals, additional flexibility induces decision confusion, making some structure better left unlearned. Specifically, we demonstrate that, in environments characterized by high noise and bounded drift, a restricted rule can outperform an expressive learner, and that increasing the sample size may not eliminate this advantage. Using a signal-to-noise ratio (SNR)-controlled Heston-style environment, we show that learning only a selector over fixed rules matches the strongest fully learned policies at low SNR, whereas end-to-end reinforcement learning catches up only at high SNR. Building on this finding, we propose Diverse-Cue Hierarchical Reinforcement Learning, a method designed for real financial markets. Experiments on eight indices from the Shanghai and Shenzhen stock exchanges show that our method attains the best total return on six of the eight markets and the best Sharpe ratio on seven. This paper offers a new perspective on architecture design under weak signals: the granularity of learning should scale with predictive information rather than model capacity.

1 INTRODUCTION

Recent advances in machine learning (ML) have demonstrated the power of increasing learned model capacity while reducing reliance on task-specific hand engineering (Sutton, 2019). Scaling model size, data, and computation has produced remarkably predictable improvements in language modeling (Kaplan et al., 2020; Ziegler et al., 2020), while expressive deep models have achieved major advances across perception, prediction, and sequential decision-making (LeCun et al., 2015; Krizhevsky et al., 2012; Mnih et al., 2015). Reinforcement learning (RL) extends this paradigm to interactive settings, where expressive policies can acquire complex behaviors from experience (Haarnoja et al., 2018; Vinyals et al., 2019; Berner et al., 2019). These successes make learned flexibility an attractive design principle: a sufficiently expressive learner can, in principle, discover task-relevant representations and decision rules with limited manual specification.

Yet greater capacity does not by itself guarantee better decisions. The same flexibility that captures subtle signal can also fit idiosyncrasies of finite samples: highly expressive neural networks can fit randomly assigned labels (Zhang et al., 2017), and deep RL agents can substantially overfit their training environments (Cobbe et al., 2019). This problem becomes especially important when predictive signal is weak relative to noise, because additional flexibility may increase estimation variance without providing commensurate gains in exploitable structure. Thus, the ability to learn a more complex decision rule is not equivalent to the ability to make a better out-of-sample decision.

Financial markets provide a particularly demanding instance of this problem. They are data-rich yet difficult to predict: although machine-learning methods can uncover economically meaningful nonlinear structure in asset returns (Gu et al., 2020), predictable variation remains small relative to noise, and many documented predictors exhibit weak or unstable out-of-sample performance (Welch

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.

& Goyal, 2008; Goyal et al., 2024). Predictive relationships may also evolve as market participants and economic conditions change (Lo, 2004). The relevant question is therefore not simply whether enough observations are available, but whether they contain enough stable predictive information to justify the estimated degrees of freedom.

Existing financial RL research has largely addressed this difficulty by improving how learning is organized: enriching state representations, decomposing decisions across temporal scales, or introducing hierarchical policies specialized to different conditions (Wang et al., 2021b;a; Li et al., 2023; Qin et al., 2024). These approaches can improve decision quality by exposing the agent to better representations or separating subtasks. Still, they primarily ask how learning should be structured rather than whether particular mappings should be estimated from noisy data in the first place.

Fixed or pretrained components are not new to hierarchical reinforcement learning, where reusable low-level behaviors may be composed by a learned high-level policy for temporal abstraction, transfer, or reuse (Sutton et al., 1999; Andreas et al., 2017; Pertsch et al., 2021). Our question is different: can fixing part of a decision architecture be statistically preferable because the available data do not contain enough predictive signal to estimate that component reliably? More generally, how should the learning flexibility depend on the predictive information available in the environment?

We study this question through the interaction between signal-to-noise ratio (SNR), decision flexibility, and finite-sample estimation. A more expressive learned policy can reduce approximation bias by representing a broader class of decision rules. Still, it also increases the structure that must be inferred from finite and noisy data (Geman et al., 1992; Gigerenzer & Brighton, 2009). A restricted or fixed rule makes the opposite trade-off: it introduces approximation bias while reducing estimation burden. Consequently, when predictive signal is strong, the benefit of flexibility may outweigh its estimation cost; when signal is weak, reduced estimation error may instead compensate for structural restriction. We therefore ask whether SNR can characterize when additional flexibility should be learned and when it is preferable to fix it.

To test this hypothesis, we construct an SNR-controlled stochastic decision environment instantiated as a financial market. The simulator augments a Heston-style stochastic-volatility process (Heston, 1993) with an explicit latent predictive component, allowing predictive signal and market noise to be separated and their variance ratio varied while holding the remaining market dynamics fixed. Agents never observe the latent predictor or latent volatility process; all experimental features are computed only from the realized price path and trading clock. The observable decision pipeline therefore remains unchanged as SNR varies.

Within this environment, we compare fixed rules, end-to-end RL, hierarchical RL, and Rule Reinforcement Learning (RRL), a hybrid hierarchical architecture in which low-level decision rules are fixed while RL learns only a high-level selector. RRL is introduced by the hypothesis that some mappings may be better left unestimated when predictive information is scarce. The resulting comparison exhibits an SNR-dependent crossover. Under a high-noise environment (low SNR), learning only a selector over fixed rules matches the strongest HRL policies. End-to-end RL catches up only when predictive information is strong. The preferred location of learning therefore shifts with the noise in the data, rather than remaining a fixed architectural choice.

We make three main contributions. First, we characterize how SNR interacts with decision flexibility and finite-sample estimation in sequential decision making, showing that restriction can improve performance when data noise overwhelms recoverable predictive signal. Second, we introduce an SNR-controlled stochastic decision environment, providing a controlled test of the predicted learn-versus-fix crossover. Third, motivated by the theoretical analysis and simulation results, we propose Diverse-Cue Hierarchical RL, which introduces a diversity-screened library of single-cue rules to improve adaptability to real-world market environments with unknown SNR. Real-market experiments verify that this method outperforms fully learned policies. Our study offers a new perspective on the design of decision-making systems in weak-signal settings: in high-noise environments with bounded drift, selecting from a set of fixed policies may outperform a high-capacity policy learner.

2 Related Work

Restricting the learned interface is already a performance lever. Prior work has imposed structure on the learned action interface rather than treating it as unrestricted. In continuous-control

RL, continuous actions can be discretized or factorized to simplify policy optimization and control the effective action space (Tang & Agrawal, 2020; Tavakoli et al., 2018). Financial RL similarly often operates over predefined discrete trading actions, trading sizes, or position sets rather than unconstrained continuous holdings (Park et al., 2020; Leem & Kim, 2020; Qin et al., 2024). We likewise treat the granularity of the learned decision class as a design choice, asking how fine that class should be as a function of decision-relevant predictive information and whether the same restriction should extend beyond the action set to which layers of a decision architecture are estimated at all. **Hierarchy decomposes the learning problem, but typically assumes that each resulting component should be learned.** Classical hierarchical methods use abstraction to restrict the information available to different levels of a policy. Feudal architectures hide low-level execution details from managers (Dayan & Hinton, 1993); MAXQ decompositions enable state abstraction by removing variables irrelevant to a subtask (Dietterich, 2000); and options provide temporal abstraction by reducing the need for repeated micro-decisions (Sutton et al., 1999). Financial hierarchical RL similarly decomposes trading across temporal scales, subtasks, or market regimes (Wang et al., 2021b;a; Li et al., 2023; Qin et al., 2024). These approaches therefore use hierarchy to simplify and structure learning, while generally retaining learned decision mappings within the resulting components. **Fixed rules can work, yet are ranked as second-best.** Restricted decision rules can perform well when the reduced estimation error outweighs the cost of limited flexibility. Fast-and-frugal heuristics can succeed through ecological fit (Gigerenzer & Goldstein, 1996; Martignon et al., 2008; Gigerenzer & Brighton, 2009), while equal-weight portfolio allocation often rivals more heavily estimated alternatives out of sample (DeMiguel et al., 2009). These results suggest that fixed rules are not merely weak baselines, but decision classes with little or no estimation burden. Their value therefore depends on whether the reduction in estimation error compensates for the structural bias introduced by restricting the decision rule.

3 PRELIMINARIES

3.1 MDP FRAMEWORK

We formulate the sequential decision problem as a finite-horizon Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. Here $\mathcal{S}$ is the state space, $\mathcal{A}$ is a finite action space, P is the transition function, R is the reward function and $\gamma \in (0, 1]$ is the discount factor. Our experiments use the undiscounted case $\gamma = 1$, with decisions taken at times $t = 0, \ldots, T-1$. The agent's behavior is given by a stochastic policy $\pi_\varphi(a \mid s)$, which specifies the probability of selecting action a in state s. The objective of reinforcement learning is to find an optimal policy that maximizes the expected discounted return: $\pi^* = \arg\max_{\pi_\varphi} \mathbb{E}_{\pi_\varphi}\left[\sum_{t=0}^{T-1} \gamma^t R_{t+1} \,\middle|\, s_0\right]$.

3.2 SINGLE-CUE DECISION RULES

Fast-and-frugal heuristics generally specify three components: a *search rule* determining which information to inspect, a *stopping rule* determining when information acquisition ends, and a *decision rule* mapping the acquired information to an action (Gigerenzer et al., 1999; Gigerenzer & Gaissmaier, 2011). Denoting these components for heuristic k by Search_k, Stop_k, and $\mathcal{D}_k$, respectively, we write $h_k = (\mathrm{Search}_k, \mathrm{Stop}_k, \mathcal{D}_k)$. In this paper, we restrict attention to *single-cue rules*, each of which evaluates only one feature (Gigerenzer & Goldstein, 1996; Martignon et al., 2008). Let $\mathcal{F} = \{f^{(1)}, \ldots, f^{(n_f)}\}$ denote the feature set, where n_f is the number of available features, and let $f^{(m_k)} \in \mathcal{F}$ be the feature selected by rule k. Let Θ_k denote its threshold set and $\mathcal{D}_k$ the associated decision mapping. Therefore, rule k reduces to $h_k = (f^{(m_k)}, \Theta_k, \mathcal{D}_k)$.

Let L_k denote the number of thresholds used by rule k, and let $\vartheta_{k,\ell}$ denote its ℓ-th threshold. The ordered threshold set is $\Theta_k = \{\vartheta_{k,1}, \ldots, \vartheta_{k,L_k}\}, \vartheta_{k,1} < \cdots < \vartheta_{k,L_k}$, which partitions the domain of $f^{(m_k)}$ into $L_k + 1$ mutually exclusive intervals. Denoting these intervals by $I_{k,j}$, we define $\mathcal{I}_k = \{I_{k,1}, \ldots, I_{k,L_k+1}\}$. Let $\mathcal{A}$ denote the action set. The mapping $\mathcal{D}_k : \mathcal{I}_k \to \mathcal{A}$ assigns one action to each interval. If $f_t^{(m_k)}$ is the observed value of rule k's feature at time t, and it falls in interval $I_{k,j}$, the outputs $a_{k,t} = \mathcal{D}_k(I_{k,j})$, if $f_t^{(m_k)} \in I_{k,j}$, where $a_{k,t} \in \mathcal{A}$ is the action prescribed by rule k. Hence, each rule implements a predefined feature–threshold–decision mapping from a single feature to an action, without learning or sequentially searching over multiple cues.

3.3 Decision Problem Instantiation

We consider a single-asset trading problem over a finite horizon with decision times $t = 0, \ldots, T-1$ and prices $p_0, \ldots, p_T$. At time t, the market is represented by OHLC prices $o_t = (p_t^o, p_t^h, p_t^l, p_t^c)$, from which a feature vector $f_t = \Psi(o_t, \ldots, o_{t-w+1})$ is constructed over a lookback window of length w. The trader chooses a long-only position $q_t \geq 0$ held from t to $t+1$. Letting $\Delta q_t = q_t - q_{t-1}$ with $q_{-1} = 0$, proportional transaction costs are $\text{cost}_t = c\, p_t^c |\Delta q_t|$. The after-cost reward and mark-to-market portfolio value are $R_{t+1} = q_t(p_{t+1}^c - p_t^c) - c\, p_t^c |\Delta q_t|, W_t = W_{t,c} + q_t p_t^c$, where $W_{t,c}$ denotes the cash balance. The objective is to maximize terminal portfolio value W_T.

To compare sequential decision architectures under financial market dynamics, we consider second-level intraday data from major Chinese equity indices as a controlled benchmark. We abstract from limit-order-book dynamics, queue position, partial fills, latency, and endogenous market impact, and interpret simulated returns as architecture-comparison measures under a common data and execution protocol rather than as directly realizable HFT profits.

4 Decision Confusion: The Crossover Between Learned and Fixed Decision Rule

Learned policies and fixed rules fail for different reasons: the former are limited by finite-sample estimation error under noise and drift, while the latter are limited by the ecological mismatch between a predefined rule and the environment. When observation noise and drift are large relative to the value separation among competing actions, a learner cannot reliably distinguish better actions from worse ones, and additional flexibility mainly increases estimation burden. We call this phenomenon *decision confusion* and use it to characterize the crossover between learned and fixed decision rules. We begin by quantifying that separation through *action gaps*, which provide the decision-relevant signal against which finite-sample noise and environmental drift will be measured. Full derivations are provided in Appendix E.

Definition 1 (Optimal Action Value Gaps)**.** Let $\mathcal{A}$ be a finite action set with $|\mathcal{A}| \geq 2$. For $t \in \{0, \ldots, T-1\}$, let $Q_t^*(s, a)$ denote the optimal action-value function, then we have:

$$Q_t^*(s,a) = \sup_{\pi} \mathbb{E}^{\pi}\Big[\sum_{j=0}^{T-1-t} \gamma^j R_{t+1+j} \;\Big|\; S_t = s, A_t = a \Big]. \tag{1}$$

Here, π ranges over admissible continuation policies from time $t+1$ onward. Then, the optimal action at state s is:

$$a_t^*(s) = \arg\max_{a \in \mathcal{A}} Q_t^*(s,a) \tag{2}$$

For $a \neq a_t^*(s)$, we define the optimal action value gap as:

$$\Delta_{t,a}(s) = Q_t^*(s, a_t^*(s)) - Q_t^*(s,a) > 0 \tag{3}$$

We assume that the learner observes $\widetilde{S}_{t+1} = S_{t+1} + e_t$, where $\mathbb{E}[e_t \mid S_t, A_t] = 0$ and $\mathbb{E}[|e_t|^2 \mid S_t, A_t] \leq \sigma_e^2$. Follow Asadi et al. (2018), each $Q_t^*(\cdot, a)$ is L_Q-Lipschitz Continuity in Model-based Reinforcement Learning , and the environmental drift is bounded as $\sup_{s,a} |Q_{t+1}^*(s,a) - Q_t^*(s,a)| \leq \delta$. Comparing estimated gaps to true gaps then yields a scalar measure of decision confusion:

Lemma 1 (Decision Confusion Index)**.** *In an environment with noise and bounded-drift non-stationarity, if action values are estimated by averaging the most recent N locally unbiased observations, then the estimated action gap decomposes as $\widehat{\Delta}_{t,N,a}(s) = \Delta_{t,a}(s) + \beta_{N,a}(s) + \zeta_{N,a}(s)$, where the sampling term contracts as $\text{Var}(\zeta_{N,a}(s) \mid s) = \gamma^2 L_Q^2 \sigma_e^2 \chi_a(s)/N$ and the pairwise staleness bias satisfies $|\beta_{N,a}(s)| \leq \delta(N-1)$. The pairwise tracking MSE is therefore $\mathcal{E}_{N,a}(s) = \text{Var}(\zeta_{N,a}(s) \mid s) + \beta_{N,a}^2(s)$. Ranking difficulty is summarized by the Decision Confusion Index*

$$\text{DCI}_{N,a}(s) = \frac{\sqrt{\text{Var}(\zeta_{N,a}(s) \mid s) + \beta_{N,a}^2(s)}}{\Delta_{t,a}(s)}. \tag{4}$$

The proof of Lemma 1 is detailed in Appendix E.3. DCI compares the scale of action-ranking error with the value separation between the two actions. Large DCI therefore means that finite-sample noise and staleness are large relative to the decision signal.

Lemma 2 (Regret of Learned Policy and Fixed Rule)**.** *For a learned policy $\hat{\pi}$, if we have $\widehat{p}_{N,a}(s) = \Pr(\hat{\pi}(s) = a \mid s)$, then the expected decision regret of the learned policy is*

$$\mathcal{R} = \mathbb{E}_S \left[\sum_{a \neq a_t^*(S)} \Delta_{t,a}(S)\, \widehat{p}_{N,a}(S) \right].$$

A fixed rule $h_k : \mathcal{S} \to \mathcal{A}$ has ecological regret

$$\mathcal{R}_k^{\text{fix}} = \sum_{a \in \mathcal{A}} u_{k,a}(1 - \alpha_{k,a}) \overline{\Delta}_{k,a}^{\text{vw}}, \tag{5}$$

where $u_{k,a} = \Pr(h_k(S) = a)$ is its action usage rate, $\alpha_{k,a} = \Pr(a_t^*(S) = a \mid h_k(S) = a)$ its action-specific validity, and $\overline{\Delta}_{k,a}^{\text{vw}}$ the mean value loss on its errors when it selects a.

The proof of Lemma 2 is detailed in Appendix E.5 and Appendix E.6. Under a Gaussian pairwise error, the probability that action a reverses its ordering with a^* is

$$\psi_{N,a}(s) = \Pr\left(\widehat{Q}_{t,N}(s,a) \geq \widehat{Q}_{t,N}(s,a^*) \mid s \right), \tag{6}$$

with drift-robust bounds $\psi_{N,a}^-(s) \leq \psi_{N,a}(s) \leq \psi_{N,a}^+(s)$ given in Appendix E.4. Learned regret depends on the unobserved selection distribution, while pairwise reversals give only per-action bounds. Lemma 3 converts those bounds into a sandwich on expected regret, so the learned-versus-fixed crossover can be located without the joint law of all action-value errors.

Lemma 3 (Regret Bound of Learned Policy)**.** *The expected decision regret of the learned policy satisfies*

$$\underline{\mathcal{R}} \leq \mathcal{R} \leq \overline{\mathcal{R}}, \tag{7}$$

where $\underline{\mathcal{R}} = \mathbb{E}_S \left[\Delta_{\min,t}(S) \max_{a \neq a_t^*(S)} \psi_{N,a}^-(S) \right]$ and $\overline{\mathcal{R}} = \mathbb{E}_S \left[\sum_{a \neq a_t^*(S)} \Delta_{t,a}(S)\, \psi_{N,a}^+(S) \right]$, with $\Delta_{\min,t}(s) = \min_{a \neq a_t^*(s)} \Delta_{t,a}(s)$.

The proof of Lemma 3 is detailed in Appendix E.5.

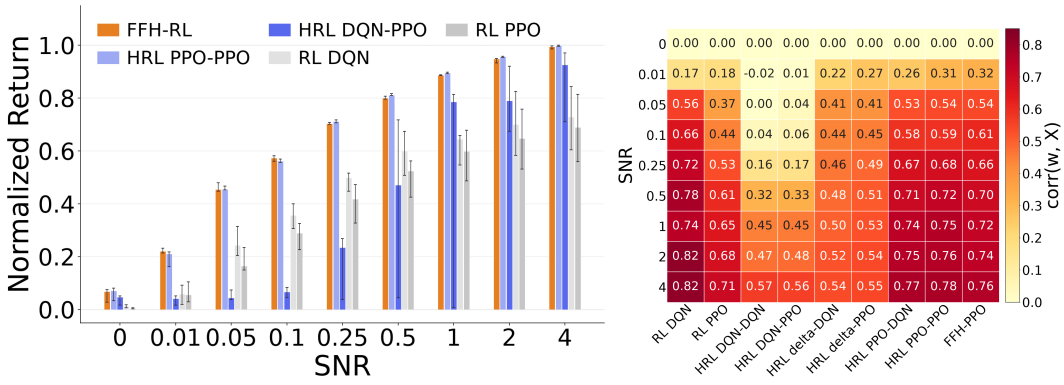
Theorem 1 (Regret-Bounded Architecture Dominance)**.** *In an environment with noise and bounded-drift non-stationarity. There exist a lower bound $\underline{\mathcal{R}}$ and an upper bound $\overline{\mathcal{R}}$ such that the regret of the learned policy R satisfies $\underline{\mathcal{R}} \leq \mathcal{R} \leq \overline{\mathcal{R}}$; a fixed-rule policy h_k has ecological regret $\mathcal{R}_k^{\text{fix}}$. Then $\overline{\mathcal{R}} < \mathcal{R}_k^{\text{fix}}$ implies that the learned structure has lower expected regret; $\underline{\mathcal{R}} > \mathcal{R}_k^{\text{fix}}$ implies that the fixed-rule policy has lower expected regret; and $\underline{\mathcal{R}} \leq \mathcal{R}_k^{\text{fix}} \leq \overline{\mathcal{R}}$ cannot be determined from pairwise information alone.*

Theorem 1 therefore partitions the comparison into three regions: a learned-dominant regime in which even a conservative bound on learned mistakes beats the fixed rule; a fixed-dominant regime in which even an optimistic bound on learned regret already exceeds the rule's ecological cost; and an indeterminate band in which pairwise information alone cannot decide the winner. This partition is the learned-versus-fixed crossover induced by decision confusion. The intermediate case is why later experiments sweep predictive signal rather than asserting a universal architectural winner. We provide an intuitive demonstration in Appendix B.

5 SNR-Controlled Simulation

5.1 Experimental Design

We provide an SNR-controlled Heston-style market (Heston, 1993) simulator in which a latent variance process and an independent stationary predictor generate second-level price paths. The scalar λ reallocates a fixed innovation variance between a persistent predictable component and idiosyncratic noise, while every agent observes only features built from the price path and the trading clock. The details of the simulator implementation is shown in Appendix A.1.



(a) (b)

Figure 1: **Left**:Total return across SNR levels and methods at $TC = 0$, bars are medians over 6 environment seeds and 6 method seeds. Complete mean±std tables for all ten methods are in Appendix D. **Right**: Decision alignment $\mathrm{corr}(w, X)$ across SNR levels and methods at $TC = 0$, w_t is the agent's normalized position weight and X is the latent predictor.

We evaluate various decision architectures over the same predefined SNR grid. Within each environment seed, the latent volatility, predictor, and innovation draws are shared across SNR conditions, so changing λ changes only the allocation between predictable and unpredictable return components. We compare the architectures spanning three families: end-to-end RL, hierarchical RL, Rule Reinforcement Learning (RRL), which fixes a library of single-cue low-level rules and learns only a high-level PPO selector, and a Single-cue rule baseline. Full method definitions are deferred to Appendix D. More implementation details of the experiments are reported in Appendix A.

5.2 Main Results

Figure 1a reports the median and the confidence interval of total return across the SNR grid for a representative subset of methods, and Figure 1b reports the decision alignment $\mathrm{corr}(w, X)$, which is defined in Appendix A.3. More experimental results are reported in Appendix D.

Fixed low-level rules remain competitive with HRL in weak-signal regimes. At $\lambda = 0$, all methods yield near-zero Sharpe ratios (Figure 3) and $\mathrm{corr}(w, X) \approx 0$. No architecture exhibits a spurious advantage when the latent predictor is absent, confirming that relative gains at positive SNR are not artifacts of the shared simulation pipeline. Over a low-SNR range, RRL matches the strongest HRL baseline (PPO–PPO): their median returns and Sharpe ratios remain close (Figure 3), while RRL executes through a fixed rule library rather than a learned low-level mapping and incurs lower turnover (Table 7). HRL that learn a DQN low-level policy have $\mathrm{corr}(w, X)$ near zero or negative in this regime. At high SNR, end-to-end DQN catches up with RRL. This crossover is qualitatively consistent with Theorem 1.

RRL gains from selecting among fixed rules rather than relying on any single rule. The high-level selector of RRL performs allocation over a library of fixed rules. The single-cue backtest exhibits a wide return distribution with large variance across environments, and its mean total return remains far below RRL at every positive SNR. This gap shows that the selector, not any one rule, accounts for RRL's gains. **RRL trades less and its turnover changes less with SNR.** The turnover of RRL is about sevenfold lower than DQN–DQN (Table 7). This comparatively stable execution style is consistent with Theorem 1. Because the low-level mapping is a restricted rule whose action usage rates $u_{k,a}$ are structural rather than re-estimated as ranking noise changes with λ, extra low-level flexibility would mainly buy estimation error. Lower trading intensity suggests greater robustness once transaction costs are introduced, which we revisit in the real-market experiments.

The decision alignment measure confirms that higher λ increases decision-relevant signal in the observation pipeline. Figure 1b provides mechanistic evidence that the SNR intervention

controls the decision-relevant information entering the fixed observation pipeline. In Figure 1b, RRL, PPO-based hierarchies, and end-to-end DQN rise from approximately zero at $\lambda = 0$ to 0.7 range as λ increases, showing that the latent predictor increasingly leaks into the shared feature map. DQN–DQN and DQN–PPO remain near zero or negative for $\lambda \leq 0.05$ and recover only once $\lambda \geq 0.5$, matching their unstable weak-signal returns. The basal action variants rise early but plateau near 0.5 even at high SNR, consistent with a truncated low-level action class that cannot fully exploit stronger ranking signal. End-to-end DQN attains the strongest alignment at high SNR, in line with the performance crossover above.

6 Diverse-Cue Hierarchical Reinforcement Learning

The simulation results for RRL reveal a systematic crossover between learned and fixed policy as decision-relevant signal increases. This suggests that the appropriate design question is not whether a decision system should be learned or fixed globally, but *where learning should occur within the decision process*. We therefore propose Diverse-Cue Hierarchical Reinforcement Learning (DC–HRL) for real-market decision making with unknown SNR: it keeps the RRL principle of fixing low-level rules while learning only a high-level selector, and constructs the low-level library by screening candidate cues for both quality and diversity.

The effective SNR available for decision making can differ across temporal resolutions. At fine timescales, individual price movements contain little predictable variation relative to short-term innovations, making reliable action ranking difficult. Over coarser horizons, persistent predictive components can accumulate while transitory noise partially averages out, potentially providing more stable information for strategy selection. This motivates an asymmetric design: micro-level decisions may benefit from stronger structural restriction, while macro-level decisions may retain sufficient signal to justify learning. DC–HRL operationalizes this idea by restricting learning at the micro level while retaining adaptation at the macro level.

Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ denote a library of fixed single-cue rules constructed from the observable feature set. At macro decision time τ, a learned high-level policy selects one rule, which remains active for the following B micro steps:

$$k_\tau \sim \pi_\varphi(\cdot \mid s_\tau^H), \qquad a_t = h_{k_\tau}(s_t), \quad t = \tau B, \ldots, \tau B + B - 1.$$

Restricting low-level learning is not sufficient if the resulting high-level action set is large and redundant. Mechanical enumeration of single-cue rules (Gigerenzer & Goldstein, 1996; Martignon et al., 2008) can produce both weak candidates and near-duplicates. Passing all candidates to the high-level learner would enlarge the selector's action space without introducing correspondingly distinct decision alternatives, effectively shifting part of the estimation burden from the low-level policy to the high-level selector.

We therefore preprocess the candidate library using the training dataset before high-level learning. Unlike HRL, which estimates a parametric low-level map, DC–HRL performs discrete quality and diversity screening over frozen single-cue rules on the same training partition. We detailed discuss their difference in Appendix C. Diversity is measured by active disagreement, which compares rules only on states where at least one takes a non-hold action. After preprocessing, the library is frozen and only the high-level policy is learned; candidate generation, thresholds, and the formal definition of active disagreement are in Appendix A.4.

7 Empirical Study

7.1 Setup

We evaluate the decision architectures on the data covering eight representative indices from the Shanghai and Shenzhen Stock Exchanges. We use index-level data because they provide broad coverage of the market and, owing to their high liquidity, offer a relatively stable setting for evaluation. Proprietary data are used because publicly available alternatives generally lack the temporal resolution required for second-level experiments. More details are provided in our released code[1].

[1] https://anonymous.4open.science/r/learnerVSfixedrule

Table 1: Held-out test performance on eight SH/SZ equity indices at TC= 1.5 bps. Each cell is the mean ± std over 6 seeds. "w/o diversity" keeps the quality-ranked pool without diversity filtering. Pink, green, and blue mark the best, second-best, and third-best, BA methods ,and "w/o diversity" is shown below the rule and is not ranked.

Market	Model	Profit	Risk-Adjusted Profit			Risk Metrics		Market	Profit	Risk-Adjusted Profit			Risk Metrics	
		TR (%)↑	ASR ↑	ACR ↑	ASoR ↑	AVOL (%)↓	MDD (%)↓		TR (%)↑	ASR ↑	ACR ↑	ASoR ↑	AVOL (%)↓	MDD (%)↓
000001.SH	DQN	8.20±4.68	0.62±0.22	0.41±0.20	0.87±0.34	6.16±1.65	9.48±1.72	000852.SH	444.53±19.22	13.99±2.28	33.99±7.00	22.72±4.14	5.84±0.88	2.43±0.48
	PPO	9.60±14.87	0.30±0.47	0.25±0.38	0.41±0.64	4.69±7.26	5.75±8.91		26.65±20.65	0.70±0.54	0.48±0.37	0.83±0.64	10.86±8.41	15.80±12.24
	DQN–PPO	-8.72±37.33	-8.93±26.08	1.11±2.85	-1.50±4.35	8.40±7.61	29.90±28.24		-33.08±41.12	-32.91±73.73	-0.46±0.69	-29.96±69.98	7.91±7.32	53.27±33.10
	PPO–PPO	19.46±47.87	0.83±1.52	0.66±1.13	0.85±1.60	11.41±4.15	25.69±13.31		937.08±77.90	20.51±2.00	17.63±8.07	12.79±1.84	3.61±0.32	5.27±3.08
	DC–HRL	**129.86±16.96**	**6.96±1.66**	**5.66±1.64**	**3.98±1.00**	**3.91±0.85**	**4.91±1.31**		**908.27±48.35**	**21.15±6.58**	**10.02±4.73**	**14.02±5.02**	**3.81±1.51**	**8.43±3.28**
	BA–DQN	1.27±7.17	0.12±0.22	0.08±0.16	0.08±0.15	10.69±6.04	19.68±12.34		-6.36±11.70	-0.65±1.18	-0.08±0.19	-0.12±0.26	7.05±5.80	23.01±19.89
	BA–PPO	59.28±30.72	2.24±1.39	1.87±1.26	1.12±0.62	8.52±4.15	12.08±8.89		268.24±10.06	7.82±1.64	5.68±1.67	4.12±1.13	5.47±1.24	7.88±2.67
	w/o diversity	-27.60±19.19	-2.13±1.80	-0.22±0.13	-1.44±1.36	5.29±0.30	43.20±12.12		679.22±39.62	24.10±1.02	6.15±0.50	11.32±0.84	2.68±0.08	10.53±0.66
000016.SH	DQN	0.43±4.19	0.02±0.31	0.04±0.15	0.07±0.37	7.49±2.01	13.04±1.88	399001.SZ	215.14±39.20	7.09±0.56	4.34±0.46	11.23±1.10	6.86±0.38	11.40±2.09
	PPO	6.28±8.06	0.23±0.27	0.18±0.22	0.29±0.37	6.90±7.69	8.40±9.40		16.65±18.24	0.41±0.45	0.28±0.30	0.52±0.57	8.25±9.03	12.24±13.41
	DQN–PPO	-42.35±44.64	-49.91±95.80	-0.08±1.65	-40.94±70.80	9.57±9.38	56.68±37.38		25.10±40.94	2.74±3.28	1.28±1.85	1.07±1.46	8.28±6.43	23.19±17.58
	PPO–PPO	-17.95±15.44	-0.54±0.23	-0.17±0.06	-0.40±0.41	11.39±8.83	33.53±25.92		744.28±183.86	19.30±4.59	19.92±3.51	14.15±2.83	3.24±0.69	3.08±0.57
	DC–HRL	**14.86±34.91**	**1.18±1.93**	**0.44±0.64**	**0.66±1.32**	**7.21±3.70**	**21.85±12.59**		**838.09±37.27**	**23.77±7.16**	**30.53±14.88**	**18.56±7.11**	**2.90±0.94**	**2.79±1.95**
	BA–DQN	23.80±37.53	0.75±1.16	0.51±0.65	0.69±1.05	10.35±2.47	24.08±13.59		168.07±34.73	4.89±1.48	3.83±2.77	3.17±1.15	5.99±1.46	9.74±4.79
	BA–PPO	-10.10±14.56	0.09±0.66	0.04±0.25	-0.13±0.30	13.56±7.21	34.77±19.59		195.10±8.96	5.22±1.05	4.17±1.72	3.00±1.05	6.08±1.02	8.12±2.10
	w/o diversity	-6.00±7.18	-1.41±1.69	-0.17±0.17	-0.14±0.28	0.90±0.81	6.46±7.24		240.70±28.67	18.28±1.22	3.46±0.22	11.64±1.49	1.90±0.08	10.07±1.04
000300.SH	DQN	17.50±5.99	0.88±0.29	0.51±0.28	1.31±0.47	9.05±1.36	18.48±6.93	399006.SZ	137.75±31.06	3.86±0.67	2.50±0.45	5.27±1.14	9.62±0.40	15.27±3.69
	PPO	21.13±10.36	0.69±0.34	0.46±0.23	0.89±0.44	11.62±5.69	17.53±8.59		48.61±23.81	0.79±0.39	0.59±0.29	1.06±0.52	19.42±9.52	26.30±12.89
	DQN–PPO	-26.98±45.07	-2.97±6.59	0.71±2.61	-3.76±6.52	7.03±6.36	42.79±36.73		15.73±24.05	1.06±1.49	0.71±1.14	0.39±0.45	13.47±7.71	31.94±20.15
	PPO–PPO	248.12±135.47	9.67±5.36	8.13±7.24	6.90±3.66	5.32±3.71	12.65±16.79		1072.35±149.81	17.79±2.25	14.04±3.33	14.16±1.25	3.95±0.38	5.24±1.45
	DC–HRL	**360.01±99.73**	**12.15±2.71**	**12.27±4.24**	**7.72±2.54**	**4.06±0.96**	**4.46±2.23**		**1375.58±18.90**	**22.43±1.29**	**31.38±7.09**	**19.21±2.56**	**3.41±0.19**	**2.56±0.67**
	BA–DQN	58.07±29.23	2.39±1.26	1.07±0.94	1.69±0.75	6.17±0.66	19.91±10.58		149.51±56.25	3.20±1.18	1.60±1.14	2.81±0.82	8.28±0.95	21.44±10.33
	BA–PPO	89.61±7.92	4.62±0.51	3.60±0.40	2.06±0.24	4.42±0.63	5.68±0.86		189.64±16.82	4.47±0.55	2.73±0.37	3.22±0.26	6.85±0.72	11.25±1.47
	w/o diversity	24.11±10.27	4.97±1.64	2.91±1.74	0.83±0.35	1.37±0.34	2.67±1.05		140.29±28.01	8.35±0.88	2.98±0.35	2.91±0.54	2.97±0.28	8.46±2.04
000688.SH	DQN	22.36±9.26	0.79±0.26	0.67±0.28	1.06±0.37	13.69±1.76	16.89±2.76	399303.SZ	714.19±69.35	27.20±1.60	57.13±6.87	33.04±3.24	3.28±0.19	1.58±0.23
	PPO	-1.19±2.90	-0.18±0.44	-0.05±0.12	-0.03±0.07	0.57±1.39	2.11±5.18		20.38±15.79	0.54±0.42	0.23±0.18	0.65±0.50	10.27±7.96	24.10±18.67
	DQN–PPO	-17.41±19.30	-8.70±11.70	-0.34±0.03	-0.78±0.89	3.98±8.75	19.26±23.02		35.47±36.08	1.70±2.06	0.40±0.39	0.83±0.66	8.86±3.93	27.76±9.93
	PPO–PPO	-0.29±34.25	0.31±1.89	0.52±1.12	-0.23±1.14	7.90±6.01	28.61±29.21		1158.58±76.52	22.06±4.68	19.60±8.90	13.27±3.12	3.39±0.76	4.46±2.15
	DC–HRL	**74.33±22.44**	**3.16±1.37**	**1.20±0.64**	**1.92±0.48**	**6.34±1.42**	**17.80±5.68**		**1043.63±41.52**	**20.44±4.99**	**17.17±8.64**	**11.97±3.21**	**3.56±0.88**	**4.93±2.27**
	BA–DQN	29.10±39.58	0.93±1.43	0.66±0.75	0.61±1.21	9.17±1.53	25.42±19.02		363.50±36.03	10.33±2.80	9.61±5.03	6.37±2.02	4.49±1.30	6.08±4.20
	BA–PPO	35.27±8.82	2.50±0.46	2.21±0.48	0.65±0.18	4.01±0.33	4.57±0.26		406.34±9.23	8.61±0.37	5.32±0.11	4.49±0.18	5.36±0.23	8.66±0.14
	w/o diversity	-12.37±4.64	-2.40±1.43	-0.28±0.10	-0.33±0.21	2.21±0.78	15.55±1.23		837.01±100.06	10.31±0.66	4.84±0.25	6.40±0.31	6.16±0.10	13.31±0.11

The data are chronologically divided into a training period from 21 January 2020 to 12 March 2024 and a test period from 13 March 2024 to 2 April 2026. All strategy construction, model training, and hyperparameter selection use only pre-test data, and the test period is held out for final evaluation. Detailed implementation settings and hyperparameters are reported in Appendix A.2.

7.2 Main Results

Table 1 reports held-out test results on the eight indices at the real-market transaction cost of 1.5 bps, over 6 seeds. The primary comparison is among end-to-end RL (DQN, PPO), hierarchical RL (DQN–PPO, PPO–PPO), and DC–HRL. The BA–DQN and BA–PPO columns use the same hierarchical interface with a basal action pool; they are included as a probe of the DC–HRL library rather than as additional competing families. As in Section 5.1, BA (basal action) denotes that the low-level policy is restricted to simple position control, rather than a library of single-cue rules.

End-to-end RL versus hierarchical RL versus DC–HRL. End-to-end RL is the weakest of the three families on most indices: PPO trails the hierarchical and DC–HRL in total return and Sharpe ratio, while DQN is competitive only in isolated cases. HRL improves when the high-level selector is PPO: PPO–PPO is the strongest hierarchical baseline and the only method that exceeds DC–HRL on total return, on 000852.SH and 399303.SZ, although DC–HRL retains the higher Sharpe ratio on 000852.SH. DQN–PPO is unstable and often unprofitable (negative mean total return on five of the eight indices), matching the weak-SNR simulation pattern in Section 5.2. DC–HRL attains the best total return on six of the eight markets and the best annualized Sharpe ratio on seven; where it leads clearly (000001.SH, 000300.SH, 399001.SZ, 399006.SZ), it also tends to post lower volatility and drawdown. Relative gains over the strongest hierarchical baseline are largest on 399006.SZ, 000300.SH, 000001.SH, and 399001.SZ. On 000016.SH, all families have small absolute returns and large variance, so the ranking is less decisive.

Role of the DC–HRL low-level library. BA–DQN and BA–PPO keep the hierarchical selector but replace the low-level mapping with simplified basal actions rather than a library of single-cue rules. They do not close the gap to DC–HRL: even the better of the two (BA–PPO) remains well below DC–HRL on profit and risk-adjusted profit in most markets, while BA–DQN is closer to the

Table 2: Macro-step ablation on eight SH/SZ indices at TC$=1.5$ bps. Cells are the mean $\pm$ sample standard deviation over 6 seeds. The $B = 60$ rows are the default DC–HRL results from Tables 1 . All variants use the diversity-screened DC–HRL library with a high-level PPO selector.

Market	Setting	Profit	Risk-Adjusted Profit			Risk Metrics		Market	Setting	Profit	Risk-Adjusted Profit			Risk Metrics	
		TR (%)↑	ASR ↑	ACR ↑	ASoR ↑	AVOL (%)↓	MDD (%)↓			TR (%)↑	ASR ↑	ACR ↑	ASoR ↑	AVOL (%)↓	MDD (%)↓
000001.SH	$B = 15$	45.87±20.82	1.66±0.56	0.98±0.61	0.94±0.26	7.02±0.89	15.07±7.13	000852.SH	$B = 15$	1054.75±15.66	29.64±13.74	31.82±19.74	19.54±10.51	3.22±2.02	3.98±3.80
	$B = 30$	74.71±1.77	2.31±0.16	2.00±0.14	1.29±0.08	7.71±0.48	8.92±0.49		$B = 30$	954.68±71.20	24.60±2.31	14.75±3.96	16.05±3.06	3.02±0.26	5.28±1.48
	$B = 45$	108.99±11.56	5.29±2.20	4.60±1.82	2.50±0.80	5.01±2.28	5.75±2.69		$B = 45$	947.86±84.10	18.22±3.59	10.21±2.70	10.88±4.01	4.17±0.87	7.59±2.07
	$B = 60$	129.86±16.96	6.96±1.66	5.66±1.64	3.98±1.00	3.91±0.85	4.91±1.31		$B = 60$	908.27±48.35	21.15±6.58	10.02±4.73	14.02±5.02	3.81±1.51	8.43±3.28
000016.SH	$B = 15$	22.56±19.14	1.51±1.31	0.74±0.64	0.76±0.89	4.22±0.92	9.52±2.29	399001.SZ	$B = 15$	910.29±31.06	21.71±0.45	25.52±0.96	14.21±0.82	3.02±0.09	2.57±0.07
	$B = 30$	31.79±15.76	2.91±1.60	1.12±0.70	1.42±0.55	3.26±0.97	8.91±3.38		$B = 30$	798.71±45.44	26.01±5.84	34.67±21.80	19.20±6.87	2.48±0.55	2.37±1.45
	$B = 45$	12.92±3.70	0.73±0.29	0.50±0.26	0.34±0.07	5.80±1.95	9.64±5.61		$B = 45$	790.25±74.29	16.92±5.00	17.17±8.54	11.46±4.45	3.94±1.43	4.67±3.24
	$B = 60$	14.86±34.91	1.18±1.93	0.44±0.64	0.66±1.32	7.21±3.70	21.85±12.59		$B = 60$	838.09±37.27	23.77±7.16	30.53±14.88	18.56±7.11	2.90±0.94	2.79±1.95
000300.SH	$B = 15$	386.30±29.84	17.57±2.51	28.41±6.25	12.14±2.10	2.86±0.39	1.80±0.37	399006.SZ	$B = 15$	1005.46±125.62	19.80±3.46	20.46±7.40	14.67±2.82	3.50±0.51	3.69±1.50
	$B = 30$	442.37±30.72	14.27±2.45	16.20±7.34	9.38±2.37	3.79±0.60	3.66±1.25		$B = 30$	1174.86±106.53	18.59±1.40	21.12±7.57	14.28±1.35	3.89±0.19	3.80±1.60
	$B = 45$	442.11±75.60	14.64±4.21	17.49±9.63	10.29±4.06	3.79±0.87	3.55±1.43		$B = 45$	1289.39±111.43	20.61±1.96	25.72±6.09	16.32±1.06	3.63±0.25	3.01±0.72
	$B = 60$	360.01±99.73	12.15±2.71	12.27±4.24	7.72±2.54	4.06±0.96	4.46±2.23		$B = 60$	1375.58±18.90	22.43±1.29	31.38±7.09	19.21±2.56	3.41±0.19	2.56±0.67
000688.SH	$B = 15$	7.20±59.16	-0.02±2.79	0.08±0.49	0.02±1.84	7.49±0.87	43.35±14.64	399303.SZ	$B = 15$	1178.89±45.96	17.28±6.59	14.11±12.16	9.71±4.25	4.56±1.50	7.90±5.10
	$B = 30$	40.62±42.78	1.31±1.46	0.54±0.60	1.07±1.03	9.13±3.00	27.13±9.92		$B = 30$	1080.94±22.21	23.89±4.32	23.58±7.99	14.32±2.78	3.00±0.62	3.28±1.39
	$B = 45$	55.36±11.12	2.67±0.59	1.19±0.48	1.40±0.34	5.57±0.50	13.37±4.18		$B = 45$	1071.97±44.27	21.88±3.67	21.47±9.64	12.83±2.44	3.25±0.60	3.98±2.43
	$B = 60$	74.33±22.44	3.16±1.37	1.20±0.64	1.92±0.48	6.34±1.42	17.80±5.68		$B = 60$	1043.63±41.52	20.44±4.99	17.17±8.64	11.97±3.21	3.56±0.88	4.93±2.27

weaker HRL variants. On 000016.SH, BA–DQN records a higher mean total return than DC–HRL, but both estimates are noisy and DC–HRL retains the higher Sharpe ratio. Thus, the advantage of DC–HRL is not explained by hierarchy alone, nor by restricting the low-level action class to a trivial pool; it tracks the fixed DC–HRL low-level library over which the high-level PPO selects.

7.3 Ablation Studies

Diversity screening. Table 1 compares a quality-ranked pool without diversity filtering to the default DC–HRL library, which further screens candidates by active disagreement. On most indices, diversity raises total return and risk-adjusted profit, often by a large margin (e.g., 000001.SH, 000300.SH, 399001.SZ, 399006.SZ), and typically lowers drawdown as well. The remaining case is 000852.SH, where the non-diverse pool attains a slightly higher Sharpe ratio but a substantially lower total return and worse Calmar and Sortino ratios. These results support the role of diversity screening in Section 6: removing near-duplicate cues shrinks a redundant high-level action set without sacrificing the fixed-rule restriction at the micro level.

Macro decision horizon. Table 2 shows the performance metrics for the macro decision horizon $B \in \{15, 30, 45, 60\}$. Cross-index mean returns remain close across horizons, and mean ASR, ACR, and ASoR likewise stay in a narrow band, with similar average volatility and drawdown. The default $B = 60$ used in Table 1 is competitive on these metrics.

8 Conclusion and Discussion

This paper proposes that when decision-relevant signals are weak relative to noise and drift, additional policy flexibility does not necessarily improve decision quality. Learned mappings are constrained by estimation error, whereas fixed rules are constrained by ecological mismatch. To test this conclusion, we construct an SNR-controlled market with fixed feature and action pipelines and show that, when signals are weak, learning only a high-level selector over fixed low-level rules matches the strongest hierarchical reinforcement learning policies. By contrast, end-to-end reinforcement learning becomes competitive only when predictive information is strong. We further propose Diverse-Cue Hierarchical Reinforcement Learning, which carries this principle to real markets. It freezes a diversity-screened library of single-cue micro-rules and learns only a macro-selector. The screening step uses the training set only to select among frozen rules; it does not estimate a parametric low-level mapping. Hierarchical controls that retain the selector but replace the library with primitive actions do not achieve the same gain, indicating that the advantage arises from the screened rules rather than from hierarchy alone.

Therefore, we argue that in weak-signal settings, the granularity of learning should scale with predictive information rather than with model capacity. Although the present evidence is limited to index-level evaluation, extending the same location-of-learning principle to other weak-signal sequential domains is a natural next step.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for manuscript formatting, language polishing, grammar checking, and implementation of part of the code. We have not used generative AI tools for the formulation of research ideas, the design of the main experiments, literature indexing, the writing of the manuscript, or any other tasks. Additionally, we used generative AI tools for manuscript formatting, language polishing, grammar checking, and implementation of part of the code. We reviewed the AI-generated code, and two of the authors independently verified its correctness. We also verified the correctness of the manuscript content after AI formatting and language polishing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work uses Level-2 (L2) order-book and transaction data from the Chinese A-share market. The data are used in accordance with that license and with applicable exchange and regulatory requirements. The license does not permit redistribution, so we do not release the raw data, nor any transformation from which the original order book could be reconstructed; we release our code in support of reproducibility.

Our study is observational and provides no means of intervening in live markets or of obtaining non-public information; we therefore do not foresee harmful applications. We nonetheless emphasize that any apparent predictive power over prices should be read subject to the backtest assumptions disclosed above, and we have taken care not to overstate its practical exploitability.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We have made substantial efforts to ensure that the results reported in this paper can be reproduced. For the theoretical results, complete proofs of all claims are provided in Appendix E. For the experiments, full implementation details are given in Appendix A.1: including hyperparameters, the construction of features, action space, and the exact definition and computation procedure for every evaluation metric. We provide an anonymized, downloadable version of our source code, including the training and evaluation scripts and the configuration files needed to reproduce every reported number. We note one limitation: the underlying A-share Level-2 dataset is obtained under a research license that does not permit redistribution, so we cannot release the raw data. To mitigate this, the released code contains the complete data-processing pipeline along with a synthetic sample in the same format, so that the pipeline can be executed end to end without access to the licensed data; licensed users can reproduce our reported results by pointing the pipeline at the original files.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Jacob Andreas, Dan Klein, and Sergey Levine. Modular multitask reinforcement learning with policy sketches. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 166–175. PMLR, 2017. URL https://proceedings.mlr.press/v70/andreas17a.html. 1

Kavosh Asadi, Dipendra Misra, and Michael L. Littman. Lipschitz continuity in model-based reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 264–273. PMLR, 2018. URL https://proceedings.mlr.press/v80/asadi18a.html. 4

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafał Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d. O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019. URL https://arxiv.org/abs/1912.06680. 1

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1282–1289. PMLR, 2019. URL https://proceedings.mlr.press/v97/cobbe19a.html. 1

Peter Dayan and Geoffrey E. Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 5, pp. 271–278, 1993. 2

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009. doi: 10.1093/rfs/hhm075. 2

Thomas G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303, 2000. doi: 10.1613/jair.639. 2

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992. doi: 10.1162/neco.1992.4.1.1. 1

Gerd Gigerenzer and Henry Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143, 2009. doi: 10.1111/j.1756-8765.2008.01006.x. 1, 2

Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual Review of Psychology*, 62:451–482, 2011. doi: 10.1146/annurev-psych-120709-145346. 3.2

Gerd Gigerenzer and Daniel G. Goldstein. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669, 1996. doi: 10.1037/0033-295X.103.4.650. 2, 3.2, 6, A.4

Gerd Gigerenzer, Peter M. Todd, and ABC Research Group. *Simple Heuristics That Make Us Smart*. Oxford University Press, New York, 1999. 3.2

Amit Goyal, Ivo Welch, and Athanasse Zafirov. A comprehensive 2022 look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 37(11):3490–3557, 2024. doi: 10.1093/rfs/hhae044. 1

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020. doi: 10.1093/rfs/hhaa009. 1

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic Algorithms and Applications. *CoRR*, abs/1812.05905, 2018. URL http://arxiv.org/abs/1812.05905. 1

Steven L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343, 1993. doi: 10.1093/rfs/6.2.327. 1, 5.1

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361. 1

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1097–1105, 2012. 1

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539. 1

JoonBum Leem and Ha Young Kim. Action-specialized expert ensemble trading system with extended discrete action space using deep reinforcement learning. *PLoS ONE*, 15(7):e0236178, 2020. doi: 10.1371/journal.pone.0236178. 2

Jiahao Li, Yong Zhang, Xingyu Yang, and Liangwei Chen. Online portfolio management via deep reinforcement learning with high-frequency data. *Information Processing & Management*, 60(3): 103247, 2023. doi: 10.1016/j.ipm.2023.103247. 1, 2

Andrew W. Lo. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, 30(5):15–29, 2004. doi: 10.3905/jpm.2004.442611. 1

Laura Martignon, Konstantinos V. Katsikopoulos, and Jan K. Woike. Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, 52(6):352–361, 2008. doi: 10.1016/j.jmp.2008.04.003. 2, 3.2, 6, A.4

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236. 1, A.4

Hyungjun Park, Min Kyu Sim, and Dong Gu Choi. An intelligent financial portfolio trading strategy using deep Q-learning. *Expert Systems with Applications*, 158:113573, 2020. doi: 10.1016/j.eswa.2020.113573. 2

Karl Pertsch, Youngwoon Lee, and Joseph J. Lim. Accelerating reinforcement learning with learned skill priors. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 672–686. PMLR, 2021. URL https://proceedings.mlr.press/v155/pertsch21a.html. 1

Molei Qin, Shuo Sun, Wentao Zhang, Hao Xia, Xinrun Wang, and Bo An. EarnHFT: Efficient hierarchical reinforcement learning for high frequency trading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14669–14676, 2024. doi: 10.1609/aaai.v38i13.29384. 1, 2

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347. A.4

Frank A. Sortino and Lee N. Price. Performance measurement in a downside risk framework. *The Journal of Investing*, 3(3):59–64, 1994. doi: 10.3905/joi.3.3.59. A.3

Richard S. Sutton. The bitter lesson. Incomplete Ideas, March 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html. 1

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211, 1999. doi: 10.1016/S0004-3702(99)00052-1. 1, 2

Yunhao Tang and Shipra Agrawal. Discretizing continuous action space for on-policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5981–5988, 2020. doi: 10.1609/aaai.v34i04.6059. 2

Arash Tavakoli, Fabio Pardo, and Petar Kormushev. Action branching architectures for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 4131–4138, 2018. doi: 10.1609/aaai.v32i1.11798. 2

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354, 2019. doi: 10.1038/s41586-019-1724-z. 1

Rundong Wang, Hongxin Wei, Bo An, Zhou Feng, and Jun Yao. Commission fee is not enough: A hierarchical reinforced framework for portfolio management. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 626–633, 2021a. doi: 10.1609/aaai.v35i1.16142. 1, 2

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 643–650, 2021b. doi: 10.1609/aaai.v35i1.16144. 1, 2

Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508, 2008. doi: 10.1093/rfs/hhm014. 1

Terry W. Young. Calmar ratio: A smoother tool. *Futures*, 20(1):40, 1991. A.3

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx. 1

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593. 1

A Implement Details

This section provides the simulator specification, hyperparameters, and implementation details for experiments. Unless stated otherwise, all experimental settings are fixed across SNR conditions.

A.1 SNR-Controlled Financial Environment

We construct synthetic financial markets in which return predictability can be varied independently of the overall innovation scale. This controlled design isolates how decision architectures respond to changes in signal strength, rather than to changes in volatility, trading frequency, or the underlying market realization.

The simulator combines stochastic volatility with a persistent latent predictor. Variance follows a Heston process,

$$\mathrm{d}\nu_t = \kappa(\theta - \nu_t)\,\mathrm{d}t + \xi\sqrt{\nu_t}\,\mathrm{d}W_t^{\nu}.$$

The initial variance is sampled from the stationary CIR distribution and simulated using full-truncation Euler. Price and variance innovations are jointly Gaussian with correlation ρ. Predictability is generated by an independent stationary AR(1) process,

$$X_{t+1} = \phi X_t + \sqrt{1-\phi^2}\,\eta_{t+1}, \qquad \eta_t \sim \mathcal{N}(0,1),$$

with

$$\mathbb{E}[X_t] = 0, \qquad \mathrm{Var}(X_t) = 1.$$

X_t is initialized from its stationary distribution and is independent of the volatility process and contemporaneous return innovation. Neither X_t nor ν_t is observable by the trading agents.

Let $\lambda \geq 0$ denote the target signal-to-noise ratio and define $s_\lambda = \sqrt{\frac{\lambda}{1+\lambda}}, n_\lambda = \sqrt{\frac{1}{1+\lambda}}$.

Log returns are generated as

$$r_t = s_\lambda \sqrt{\nu_t \Delta t}\, X_t - \frac{1}{2}\nu_t \Delta t + n_\lambda \sqrt{\nu_t \Delta t}\, \varepsilon_t,$$

where ε_t is independent standardized noise. Prices start at p_0 and, for $t = 0, \ldots, T-1$, evolve according to

$$p_{t+1} = p_t \exp(r_t).$$

By construction,

$$\frac{s_\lambda^2}{n_\lambda^2} = \lambda, \qquad s_\lambda^2 + n_\lambda^2 = 1.$$

Thus, increasing λ reallocates a fixed contemporaneous innovation variance from unpredictable noise to a persistent predictable component, rather than mechanically increasing return volatility. At $\lambda = 0$, the predictable component vanishes.

Remark 1 (From Latent SNR to Decision Confusion). The theory in Section 4 is stated in terms of action gaps, observation noise, drift, and sample size rather than the simulator parameter λ. Here, λ controls the decision-relevant information reaching the agent through a fixed observation pipeline: increasing λ shifts innovation variance from unpredictable noise toward the persistent latent predictor X_t, which agents observe only indirectly through price-derived features. As λ increases, these features tend to provide stronger action-ranking information, widening effective action gaps relative to ranking noise and thereby reducing $\mathrm{DCI}_{N,a}(s)$. Thus, sweeping λ moves the environment across different confusion regimes without changing the feature map or action interface, tracing the learned-versus-fixed crossover rather than introducing a separate theoretical quantity. Appendix E.8 provides the qualitative mapping and empirical proxies.

All observable features are constructed only from prices and the trading clock; latent predictors, latent volatility, and trading volume are never exposed to the agents.

A.2 Hyperparameters

Table 3 reports the main simulator, learning, and RRL/DC–HRL hyperparameters used in the controlled simulation and, where noted, in the real-data experiments. Hyperparameter values are fixed across SNR levels unless explicitly varied as an experimental condition.

A.3 Methods and Metrics

Architectures. All methods share the same observable features, inventory grid, micro-action set $\mathcal{A}^{\mathrm{lo}} = \{-1, 0, +1\}$, and after-cost reward; they differ only in where the decision mapping is learned.

- **End-to-end RL:** DQN and PPO map the observable state directly to micro actions.
- **Hierarchical RL:** A high-level policy selects a low-level policy every B micro-steps. We evaluate all four high/low pairings: DQN–DQN, DQN–PPO, PPO–DQN, and PPO–PPO.
- **Basal-action hierarchical variants:** BA–DQN and BA–PPO use the same hierarchical interface, but their lower-level policies are constructed as basal actions, each of which repeatedly executes a single action from the action space $\mathcal{A}^{\mathrm{lo}}$.
- **Single-cue rule:** The single-cue rule is a fixed rule.
- **RRL:** The low-level library is a fixed pool of single-cue rules constructed on the training partition (Appendix A.4); only a high-level PPO selector over $\mathcal{A}^{\mathrm{hi}}$ is learned.
- **Oracle:** Oracle $\mathrm{sign}(X_t)$ trades on the sign of the latent predictor (unavailable to agents) and provides an upper bound.

Metrics.

- **Total return:** terminal portfolio value relative to the initial wealth on the held-out test partition.

Table 3: Hyperparameters used in the controlled simulation and real-data experiments.

Component	Hyperparameter	Value
Simulator		
	Bar interval	5 seconds
	Trading horizon	243 trading days (1 year)
	Initial price p_0	3000
	Heston mean reversion κ	2
	Long-run variance θ	0.04
	Volatility of variance ξ	0.3
	Return–variance correlation ρ	-0.7
	Predictor persistence ϕ^*	0.9885
RL and HRL		
	Train/test split	$6:4$
	Episode length	5 trading days
	End-to-end RL training episodes	200
	Hierarchical high-level training episodes	100
	Hierarchical low-level training episodes	100
	Transaction-cost rates c in empirical study	15×10^{-5}
	High-level horizon B	60 bars (5 minutes)
RRL		
	Candidate cues	Observable feature set
	Rule type	Single-cue rules
	Maximum strategy-pool size K_h	200
	Active-disagreement threshold τ_D	0.05

*Computed as $\phi = \exp\big(-\ln 2 \cdot \Delta t/\tau_{1/2}\big)$.

- **Turnover:** mean absolute position change per micro-step, $|\Delta q_t|$, reported in units of 10^{-3}.
- **Maximum Drawdown (MDD)** captures the largest peak-to-trough decline in portfolio value during the trading period, reflecting the worst realized loss.
- **Annualized Volatility (AVOL)** quantifies the variability of returns and serves as a proxy for risk. Given a sequence of returns $r = (r_1, r_2, \ldots, r_T)$, the annualized volatility is defined as $AVOL = \sigma[r] \cdot \sqrt{N_{\text{year}}}$, where $\sigma[\cdot]$ denotes the standard deviation and N_{year} is the number of time steps in one year.
- **Annualized Sharpe Ratio (ASR)** evaluates risk-adjusted performance by comparing expected return against total volatility: $ASR = \frac{\mathbb{E}[r]}{\sigma[r]} \cdot \sqrt{N_{\text{year}}}$, where $\mathbb{E}[\cdot]$ denotes expectation.
- **Annualized Calmar Ratio (ACR)** relates return to downside risk by normalizing annualized return with respect to maximum drawdown Young (1991): $ACR = \frac{\mathbb{E}[r]}{MDD} \cdot N_{\text{year}}$.
- **Annualized Sortino Ratio (ASoR)** refines risk measurement by considering only downside fluctuations Sortino & Price (1994). It is defined as $ASoR = \frac{\mathbb{E}[r]}{DD} \cdot \sqrt{N_{\text{year}}}$, where DD denotes the downside deviation computed from negative returns only.
- **Decision alignment** $\text{corr}(w, X)$**:** Pearson correlation between the agent's position weight w_t and the latent predictor X_t on the test partition, where $w_t = q_t/q_{\max}$ is the inventory expressed as a fraction of the maximum position. Because agents never observe X_t, this quantity measures how much of the decision-relevant latent structure is recovered through the fixed price-based observation pipeline.

A.4 Algorithm Implementation Details

Model implementations and observable states. DQN and PPO follow the algorithms of Mnih et al. (2015) and Schulman et al. (2017), respectively. Our DQN and PPO implementations, including the hierarchical high- and low-level variants, are adapted from the EarnHFT codebase `https://github.com/TradeMaster-NTU/EarnHFT`.

Table 4: Observable features used in the controlled simulation. SMA, $\bar{\ell}_{t,h}$, $\mathrm{sd}_{t,h}$, and RSI are defined in the text above.

Feature	Meaning	Definition
$r^{(1)}$	1-bar return	$\ell_t - \ell_{t-1}$
$r^{(5)}$	5-bar return	$\ell_t - \ell_{t-5}$
$r^{(20)}$	20-bar return	$\ell_t - \ell_{t-20}$
$M^{(5,20)}$	Short MA spread	$\mathrm{SMA}_5(p)/\mathrm{SMA}_{20}(p) - 1$
$M^{(20,60)}$	Long MA spread	$\mathrm{SMA}_{20}(p)/\mathrm{SMA}_{60}(p) - 1$
$Z^{(20)}$	20-bar price z-score	$(\ell_t - \bar{\ell}_{t,20})/\mathrm{sd}_{t,20}$
$Z^{(60)}$	60-bar price z-score	$(\ell_t - \bar{\ell}_{t,60})/\mathrm{sd}_{t,60}$
$\mathrm{RV}^{(20)}$	20-bar realized volatility	$\sqrt{\sum_{j=0}^{19}(r_{t-j}^{(1)})^2}$
$\mathrm{RV}^{(60)}$	60-bar realized volatility	$\sqrt{\sum_{j=0}^{59}(r_{t-j}^{(1)})^2}$
$\mathrm{RSI}^{(14)}$	Relative strength index	see text

In the controlled simulation, the observable state is fixed across architectures. We first rank causal price-based features by

$$\left|\widehat{\mathrm{corr}}(X_t, f_t^{(m)})\right|$$

on the training partition of a single reference path with $\lambda = 1$, where X_t is the latent predictor and $f_t^{(m)}$ is computed only from prices observed up to t. Here $\widehat{\mathrm{corr}}$ denotes the sample Pearson correlation on the training partition:

$$\widehat{\mathrm{corr}}(X, f) = \frac{\sum_{t\in\mathcal{T}_{\mathrm{tr}}}(X_t - \bar{X})(f_t - \bar{f})}{\sqrt{\sum_{t\in\mathcal{T}_{\mathrm{tr}}}(X_t - \bar{X})^2 \sum_{t\in\mathcal{T}_{\mathrm{tr}}}(f_t - \bar{f})^2}}, \tag{8}$$

with $\bar{X}$ and $\bar{f}$ the corresponding sample means. After removing redundant features, we freeze and reuse the top ten for every SNR, prediction horizon, environment seed, and architecture.

The state contains

$$\left[r^{(1)}, r^{(5)}, r^{(20)}, M^{(5,20)}, M^{(20,60)}, Z^{(20)}, Z^{(60)}, \mathrm{RV}^{(20)}, \mathrm{RV}^{(60)}, \mathrm{RSI}^{(14)}\right],$$

together with the current position-grid index. Here $r^{(h)}$ denotes the h-bar cumulative log return with $\ell_t = \log p_t$, M moving-average spreads, Z rolling log-price z-scores, and RV realized volatility. For the moving-average and z-score features, we write

$$\mathrm{SMA}_n(p)_t = \frac{1}{n}\sum_{j=0}^{n-1} p_{t-j}, \quad \bar{\ell}_{t,h} = \frac{1}{h}\sum_{j=0}^{h-1}\ell_{t-j}, \quad \mathrm{sd}_{t,h} = \sqrt{\frac{1}{h-1}\sum_{j=0}^{h-1}(\ell_{t-j} - \bar{\ell}_{t,h})^2}, \tag{9}$$

and the 14-bar relative-strength index is

$$\Delta p_t = p_t - p_{t-1}, \qquad U_t = \max(\Delta p_t, 0), \qquad D_t = \max(-\Delta p_t, 0), \tag{10}$$

$$\mathrm{RSI}_t^{(14)} = 100 - \frac{100}{1 + \overline{U}_{t,14}/(\overline{D}_{t,14} + \varepsilon)}, \tag{11}$$

where $\overline{U}_{t,14}$ and $\overline{D}_{t,14}$ are simple 14-bar rolling means of U and D, and $\varepsilon > 0$ is a small numerical stabilizer. Table 4 lists the resulting features; SMA, rolling z-score, and RSI terms follow the definitions above. No observable feature uses the latent predictor X_t, latent variance ν_t, or trading volume.

Unlike the controlled simulation, the real-data experiments use model-specific observable states following the corresponding reference implementations and feature-selection procedures. Feature construction, normalization, and selection are restricted to the training data. The controlled experiment therefore fixes the observable representation to isolate differences in decision architecture, whereas the real-data experiment follows architecture-specific feature protocols to provide a best-practice comparison under realistic market conditions.

Action Space. In the trading instantiation, this action space $\mathcal{A}$ is realized at three nested layers: a five-level inventory grid

$$\mathcal{A}^{\text{grid}} = \{0, 1, 2, 3, 4\} \quad \text{with} \quad q_t = \frac{i_t}{4} q_{\max} \quad \text{where} \quad i_t \in \mathcal{A}^{grid},$$

a micro action set $\mathcal{A}^{\text{lo}} = \{-1, 0, +1\}$ that moves the grid index by at most one level with boundary clipping, and a high-level selector $\mathcal{A}^{\text{hi}} = \{1, \ldots, K\}$ over a library $\mathcal{H} = \{h_1, \ldots, h_K\}$ of low-level policies or rules. End-to-end RL learns a mapping into $\mathcal{A}^{\text{lo}}$; hierarchical methods learn a mapping into $\mathcal{A}^{\text{hi}}$ while the selected low-level polices which act in $\mathcal{A}^{\text{lo}}$. Section 4 treats $\mathcal{A}$ as this generic finite placeholder. All architectures share these nested action layers.

Reward. The micro-step reward is after-cost absolute P&L on the position held over the bar,

$$R_{t+1} = q_t(p_{t+1} - p_t) - cp_t|\Delta q_t|,$$

which instantiates the mark-to-market identity of the preliminaries with the scalar simulated price p_t in place of the close p_t^c. The transaction-cost settings used in the controlled simulation and real-data experiments are reported separately in the hyperparameter table.

Hierarchical models select a low-level policy every B micro-steps, which remains active over the corresponding macro interval. The high-level reward is

$$R_\tau^H = \sum_{j=0}^{B-1} R_{\tau B+j+1}.$$

Thus, architectures share the same market interaction, inventory constraints, and micro-level objective and differ only in which layer of $\mathcal{A}$ is learned.

Rule library. The low-level libraries of RRL and DC–HRL consist of fixed single-cue rules $h_k = (f^{(m_k)}, \Theta_k, \mathcal{D}_k)$ as in the preliminaries, following the one-reason decision principle (Gigerenzer & Goldstein, 1996; Martignon et al., 2008) and specialized to two thresholds and the micro-action set $\mathcal{A}^{\text{lo}}$. For each feature, candidate thresholds are estimated from the training partition using empirical quantiles and $\mu \pm j\,\text{sd}$ levels for $j \in \{0, 1, 2, 3\}$. Every ordered threshold pair $(\vartheta_{k,1}, \vartheta_{k,2})$ partitions the feature into three intervals, and all 3! permutations of $\{-1, 0, +1\}$ across these intervals are enumerated:

$$h_k(f_t^{(m_k)}) = \begin{cases} a_1, & f_t^{(m_k)} < \vartheta_{k,1}, \\ a_2, & \vartheta_{k,1} \le f_t^{(m_k)} < \vartheta_{k,2}, \\ a_3, & f_t^{(m_k)} \ge \vartheta_{k,2}, \end{cases} \qquad (a_1, a_2, a_3) \in \text{Perm}(-1, 0, +1).$$

In DC-HRL, candidate rules are evaluated on the training partition and ranked within each feature by the 75th percentile of their rolling-window return distribution. Candidates are then greedily filtered by *active disagreement*. Let $\mathcal{T}_{\text{tr}}$ denote the set of time indices in that training partition, and let s_t be the observable state at time $t \in \mathcal{T}_{\text{tr}}$. For two rules h and h', write $a_t = h(s_t)$ and $a'_t = h'(s_t)$ for the micro-actions they induce in $\mathcal{A}^{\text{lo}} = \{-1, 0, +1\}$, where 0 is the hold action. Their active disagreement is the fraction of such times on which the two actions differ, counting only times at which at least one rule does not hold:

$$\text{AD}(h, h') = \frac{\#\{t \in \mathcal{T}_{\text{tr}} : a_t \ne a'_t,\ \max(|a_t|, |a'_t|) > 0\}}{\#\{t \in \mathcal{T}_{\text{tr}} : \max(|a_t|, |a'_t|) > 0\}}, \tag{12}$$

where $\text{AD}(h, h')$ denotes this active disagreement, $\#\{\cdot\}$ is the number of elements in the indicated set, $|\cdot|$ is absolute value, and $\max$ returns the larger of its two arguments. Thus $\max(|a_t|, |a'_t|) > 0$ holds when at least one of a_t and a'_t differs from the hold action 0. If the denominator is empty, we set $\text{AD}(h, h') = 0$. A candidate is discarded if its active disagreement with any previously retained rule is below the threshold $\tau_D = 0.05$ reported in Table 3. At most 20 rules are retained for each feature. Once constructed, all low-level rules remain fixed and only the high-level selector is learned.

Controlled-simulation training and evaluation. Each simulated path is split chronologically into 60% training and 40% testing data. Five-day training episodes are sampled exclusively from the training partition. End-to-end DQN and PPO are trained on 200 sampled episodes, whereas hierarchical high-level policies use 100 episodes. Each learned architecture is retrained independently for every SNR and transaction-cost condition and is evaluated using its final training checkpoint.

Real-data training and evaluation. For the real-data experiments, the dataset is split into 60% training and 40% testing data. Each architecture is evaluated under the same trading constraints, with the real-market transaction-cost setting reported in the hyperparameter table. DC–HRL libraries are constructed independently for each market using its training data and remain fixed during subsequent high-level learning. Reported performance is computed exclusively on the chronologically held-out test period.

B Intuitive Demonstration of Learn versus Fix-rule

Theorem 1 compares a sandwich on learned ranking regret with the ecological regret of a fixed rule. Figure 2 makes the same comparison visible in a supervised analogue that isolates estimation error from approximation bias. An expressive network is re-estimated on each finite noisy sample, while a restricted linear rule is held fixed across all noise levels. Both predictors are scored against the same noise-free latent function f^*, so the comparison is recoverability of structure rather than in-sample fit to noisy targets.

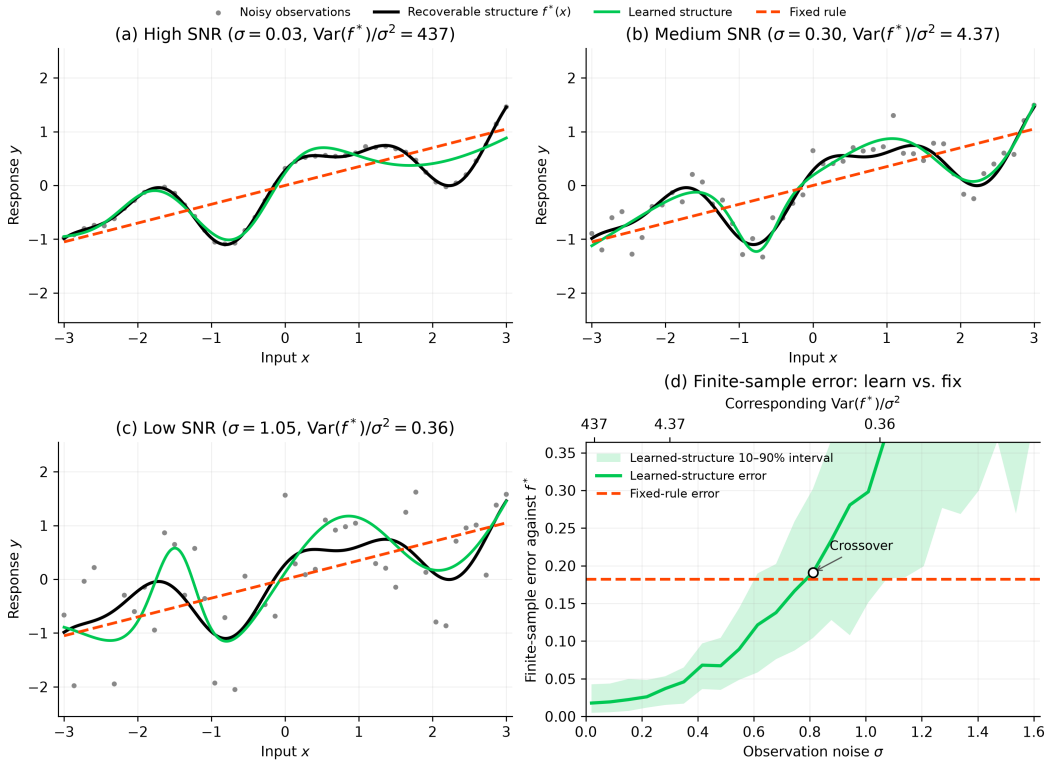


Figure 2: A illustration of the finite-sample trade-off between learning additional structure and fixing a restricted rule. Noisy observations follow $y = f^*(x) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. An expressive network is re-fit on each finite sample, while a misspecified linear rule is held fixed across noise levels. (a–c): As σ rises, the learned fit recovers f^* at high SNR, remains usable at medium SNR, and follows sample idiosyncrasies at low SNR, whereas the fixed rule does not change. (d) Mean finite-sample error of the learned fit (solid; 10–90% interval over 50 re-fits) increases with σ and crosses the constant error of the fixed rule.

At high SNR, the learned fit recovers the nonlinear oscillations of f^* and the restricted rule's bias is visible. At medium SNR, the learned curve still tracks the recoverable structure despite scattered observations. At low SNR, additional flexibility mainly follows idiosyncrasies of the finite sample, while the fixed rule is unchanged because it is not re-estimated. Panel (d) summarizes the same protocol over a noise grid: learned error rises with σ, the fixed-rule error is constant, and the two curves cross. Left of the crossover, even a misspecified restricted rule is worse than learning; right of it, learning additional structure increases finite-sample error relative to the fixed rule. This is the

operational content of Theorem 1, with σ standing in for ranking noise and the linear rule standing in for a biased but unestimated decision class.

C HRL VERSUS DC–HRL: PARAMETRIC LEARNING VERSUS FIXED-POLICY SCREENING

HRL and DC–HRL both use the training partition to produce an executable policy. In DC–HRL, the low-level library is not specified a priori: candidate thresholds are estimated from training quantiles and $\mu \pm j\,\mathrm{sd}$ levels, and quality and diversity screening are performed on the same training partition (Appendix A.4). Macroscopically, both architectures therefore use data for strategy construction.

The distinction lies in *how* data enter the low-level mapping. HRL estimates a parametric state–action map within a high-capacity function class, typically by temporal-difference or policy-gradient updates at both levels. DC–HRL instead performs a one-shot discrete screen over a finite, pre-enumerated class of single-cue rules: each retained rule's feature–threshold–decision table is fixed once constructed, and training data decide only which candidates are kept or discarded. After screening, the library is frozen, and only the high-level selector is optimized by reinforcement learning.

These two forms of data use have different statistical consequences for decision confusion (Section 4). Parametric low-level learning continually re-estimates which micro action is preferred, so finite-sample idiosyncrasies can inflate estimation error as the hypothesis class grows. Screening restricts the effective hypothesis class to a small set of frozen rules: selection bias remains possible—a rule that looks strong on the training partition may be retained by chance—but that bias is not equivalent to re-estimating micro-level action rankings from noisy observations. Quality screening removes weak candidates; diversity screening via active disagreement removes near-duplicates, preventing the estimation burden from being shifted into an oversized high-level action space.

The empirical comparison between HRL and DC–HRL should therefore be read as a comparison of data-use regimes—parametric estimation of a flexible low-level map versus discrete selection among frozen restricted rules—rather than as a claim that one architecture uses data and the other does not. DC–HRL operationalizes the learned-versus-fixed crossover of Section 4 by placing parametric learning at the coarser selector, where recoverable signal may justify adaptation, while suppressing parametric estimation of the micro-level action map.

D MORE EXPERIMENT RESULTS

This section complements and reports the full mean±std tables omitted from the main text. The results use 6 environment seeds and 6 method seeds per cell ($n = 36$). Table 5 and Table 6 report mean±standard deviation over environment and method seeds. In the total-return and Sharpe tables, pink, green, and blue cells mark the best, second-best, and third-best method within each SNR row excluding Oracle.

Table 5: Total Return (%) on the controlled Heston simulation at TC$= 0$. Cells are mean $\pm$ sample standard deviation over 6 environment seeds and 6 method seeds ($n = 36$; Oracle uses one value per environment, $n = 6$). Pink, green, and blue mark the best, second-best, and third-best method in each SNR row excluding Oracle.

SNR	DQN	PPO	DQN–DQN	DQN–PPO	PPO–DQN	PPO–PPO	Δ-DQN	Δ-PPO	FFH–PPO	Oracle
0	-2.48±6.56	-3.74±6.65	10.08±7.95	6.92±3.25	15.66±8.32	15.17±7.56	16.83±7.47	18.74±3.16	**13.88±6.24**	1.13±7.04
0.01	14.77±15.19	18.86±25.47	2.79±13.66	6.43±4.59	75.62±28.06	95.40±25.06	84.96±21.84	99.13±13.18	**106.1±24.9**	567.4±98.3
0.05	144.1±47.6	94.89±52.22	16.89±45.98	17.16±17.58	421.4±78.3	428.9±82.0	341.9±59.8	341.2±60.7	**432.6±77.0**	1336.7±250.6
0.1	261.8±60.6	168.5±68.8	11.60±25.75	23.58±21.99	665.1±141.6	680.3±123.8	531.5±106.5	548.7±102.8	**707.9±126.3**	1944.9±376.1
0.25	495.0±113.8	324.9±86.5	204.5±336.1	182.9±322.2	1248.2±241.7	1260.8±232.7	900.4±169.3	966.7±183.5	**1242.0±212.2**	3166.3±637.8
0.5	769.8±166.2	536.4±129.9	597.3±641.5	620.2±656.8	1865.0±338.6	1890.7±360.3	1293.4±237.1	1410.0±277.7	**1831.0±329.0**	4452.7±922.3
1	922.1±303.5	763.0±201.4	1311.2±1076.0	1306.8±1070.2	2621.9±498.9	2641.2±473.1	1773.6±352.4	1944.7±388.3	**2518.5±423.4**	5947.0±1260.0
2	1231.2±338.0	973.2±269.7	1669.8±1233.0	1717.6±1252.3	3319.7±627.9	3332.1±612.1	2361.6±436.2	2475.3±492.3	**3211.5±632.9**	7386.5±1591.1
4	1374.9±398.9	1176.3±356.5	2332.7±1430.3	2329.9±1402.9	3901.3±745.2	3910.9±730.6	2804.4±532.9	2900.6±583.0	**3808.7±680.4**	8510.2±1853.6

Figure 3 reports the median Sharpe ratio omitted from the main-text panel in Section 5.2. Bars are medians over 6 environment seeds and 6 method seeds for the same method subset as Figure 1a. At $\lambda = 0$, every method is near zero, so no architecture shows a spurious risk-adjusted advantage when the latent predictor is absent. Over the low-SNR range (approximately $\lambda \leq 0.25$), RRL remains

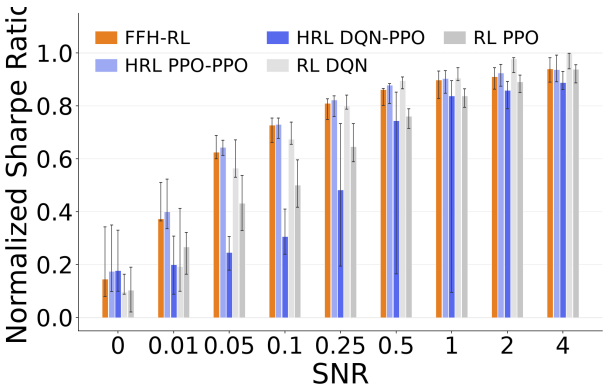


Figure 3: Median Sharpe ratio across SNR levels in the controlled Heston simulation at $TC = 0$. Bars are medians over 6 environment seeds and 6 method seeds for the same method subset as Figure 1a. At $\lambda = 0$, all methods are near zero. Over low SNR ($\lambda \leq 0.25$), RRL remains close to PPO–PPO, while end-to-end DQN catches up only at high SNR ($\lambda \geq 1$). Complete mean±std values are in Table 6.

Table 6: Sharpe Ratio on the controlled Heston simulation at TC$= 0$. Cells are mean $\pm$ sample standard deviation over 6 environment seeds and 6 method seeds ($n = 36$; Oracle uses one value per environment, $n = 6$). Pink, green, and blue mark the best, second-best, and third-best method in each SNR row excluding Oracle.

SNR	DQN	PPO	DQN–DQN	DQN–PPO	PPO–DQN	PPO–PPO	Δ-DQN	Δ-PPO	FFH–PPO	Oracle
0	-0.22±1.77	-0.19±1.80	1.94±1.34	2.59±1.11	2.50±1.31	2.39±1.17	3.33±1.80	2.62±0.60	**1.96±0.88**	0.29±1.04
0.01	5.14±5.64	5.34±7.54	-0.04±3.56	2.20±1.37	10.99±4.70	15.20±4.51	13.62±3.63	15.46±0.60	**15.11±3.16**	52.88±3.47
0.05	43.75±15.62	19.16±11.60	1.44±14.19	6.57±7.85	58.03±7.17	58.37±7.42	45.94±3.71	46.01±3.22	**55.81±3.73**	102.2±8.0
0.1	77.61±11.56	32.88±13.58	8.17±18.06	12.86±17.17	85.66±10.91	83.94±11.07	66.69±6.19	66.78±4.15	**83.49±8.34**	133.2±11.1
0.25	141.2±24.4	63.36±14.24	37.49±48.56	41.14±46.69	130.7±9.4	132.8±10.1	101.1±9.7	99.90±6.43	**123.2±10.0**	184.5±16.6
0.5	205.8±41.4	107.2±21.4	83.84±82.28	87.55±82.34	161.6±12.9	170.5±15.4	132.3±13.7	128.0±10.2	**159.8±15.9**	229.1±22.0
1	235.7±53.2	155.7±31.7	128.0±98.2	125.6±97.0	202.2±18.6	206.5±17.0	162.1±13.6	155.9±13.7	**195.8±27.5**	274.1±28.6
2	301.0±65.3	204.0±40.2	151.9±91.9	152.6±92.5	230.9±18.5	233.9±18.8	185.1±17.6	181.1±17.1	**224.9±26.4**	313.6±35.5
4	328.4±76.3	254.7±56.0	194.6±77.6	194.6±77.7	257.7±25.8	259.9±26.3	201.2±19.7	197.3±18.9	**255.1±27.5**	343.1±41.6

close to PPO–PPO. End-to-end DQN becomes competitive only once $\lambda \geq 1$. Table 6 reports the corresponding mean±standard deviation over environment and method seeds.

Table 7: Heston SNR f_{10} simulation, TC$= 0$: turnover mean±std by SNR. Official results use 6 environments $\times$ 6 seeds ($n = 36$). Values are $\times 10^{-3}$.

Method	0	0.01	0.05	0.1	0.25	0.5	1	2	4
DQN	$1.256_{\pm 0.996}$	$0.532_{\pm 0.802}$	$0.982_{\pm 0.934}$	$0.910_{\pm 0.921}$	$0.877_{\pm 0.745}$	$0.969_{\pm 0.597}$	$0.630_{\pm 0.545}$	$0.826_{\pm 0.611}$	$0.663_{\pm 0.326}$
PPO	$1.042_{\pm 1.287}$	$0.816_{\pm 1.045}$	$0.261_{\pm 0.196}$	$0.188_{\pm 0.152}$	$0.170_{\pm 0.081}$	$0.181_{\pm 0.064}$	$0.173_{\pm 0.042}$	$0.158_{\pm 0.023}$	$0.146_{\pm 0.017}$
DQN-DQN	$2.115_{\pm 0.329}$	$1.952_{\pm 0.364}$	$2.057_{\pm 0.353}$	$1.722_{\pm 0.516}$	$2.085_{\pm 0.350}$	$1.413_{\pm 0.572}$	$1.305_{\pm 0.604}$	$1.319_{\pm 0.663}$	$1.187_{\pm 0.557}$
DQN-PPO	$1.982_{\pm 0.277}$	$2.205_{\pm 0.123}$	$2.224_{\pm 0.156}$	$1.917_{\pm 0.502}$	$2.157_{\pm 0.377}$	$1.472_{\pm 0.611}$	$1.294_{\pm 0.630}$	$1.143_{\pm 0.769}$	$1.045_{\pm 0.579}$
delta-DQN	$0.028_{\pm 0.029}$	$0.058_{\pm 0.022}$	$0.054_{\pm 0.011}$	$0.051_{\pm 0.010}$	$0.049_{\pm 0.008}$	$0.050_{\pm 0.006}$	$0.051_{\pm 0.005}$	$0.051_{\pm 0.003}$	$0.052_{\pm 0.004}$
delta-PPO	$0.019_{\pm 0.021}$	$0.070_{\pm 0.003}$	$0.063_{\pm 0.002}$	$0.059_{\pm 0.001}$	$0.056_{\pm 0.001}$	$0.055_{\pm 0.001}$	$0.055_{\pm 0.001}$	$0.054_{\pm 0.001}$	$0.054_{\pm 0.001}$
PPO-DQN	$0.748_{\pm 0.825}$	$0.738_{\pm 0.814}$	$0.344_{\pm 0.417}$	$0.182_{\pm 0.108}$	$0.336_{\pm 0.238}$	$0.306_{\pm 0.138}$	$0.389_{\pm 0.280}$	$0.292_{\pm 0.132}$	$0.269_{\pm 0.082}$
PPO-PPO	$0.610_{\pm 0.769}$	$0.726_{\pm 0.720}$	$0.298_{\pm 0.415}$	$0.208_{\pm 0.127}$	$0.273_{\pm 0.142}$	$0.293_{\pm 0.107}$	$0.358_{\pm 0.158}$	$0.320_{\pm 0.170}$	$0.280_{\pm 0.092}$
FFH-PPO	$0.399_{\pm 0.162}$	$0.229_{\pm 0.080}$	$0.216_{\pm 0.061}$	$0.222_{\pm 0.082}$	$0.196_{\pm 0.077}$	$0.219_{\pm 0.069}$	$0.252_{\pm 0.119}$	$0.218_{\pm 0.130}$	$0.177_{\pm 0.064}$

E Theory Details and Proofs

Assumption 1 (Estimation and regularity conditions)**.** [label=i)] For the current decision time t:

1. The learner observes a noisy next state $\widetilde{S}_{t+1} = S_{t+1} + e_t$, with $\mathbb{E}[e_t \mid S_t, A_t] = 0$ and $\mathbb{E}[\|e_t\|^2 \mid S_t, A_t] \leq \sigma_e^2$.

2. For every action a, $Q_t^*(\cdot, a)$ is L_Q-Lipschitz in the state.

3. Current action values are estimated from the most recent N locally unbiased value observations. Their pairwise noise terms are conditionally mean-zero and form a martingale-difference sequence. The learned decision rule selects the action with the largest estimated action value.

4. Environmental drift is bounded by $\sup_{s,a} |Q^*_{t+1}(s,a) - Q^*_t(s,a)| \le \delta$.

5. Conditional on the current state, the averaged pairwise estimation error is Gaussian.

E.1 Noisy Observations and Bellman Targets

Part (i) of Assumption 1 states that the learner observes $\widetilde{S}_{t+1} = S_{t+1} + e_t$. The associated clean Bellman target is

$$Y_t = R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q^*_{t+1}(S_{t+1}, a'), \tag{13}$$

and the target induced by the noisy observation is

$$\widetilde{Y}_t = R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q^*_{t+1}(\widetilde{S}_{t+1}, a'). \tag{14}$$

Part (ii) states that every $Q^*_{t+1}(\cdot, a')$ is L_Q-Lipschitz, so their pointwise maximum is L_Q-Lipschitz. Therefore

$$|\widetilde{Y}_t - Y_t| \le \gamma L_Q \|e_t\|. \tag{15}$$

Squaring and taking the conditional expectation gives

$$\mathbb{E}\Big[(\widetilde{Y}_t - Y_t)^2 \mid S_t, A_t\Big] \le \gamma^2 L_Q^2 \, \sigma_e^2. \tag{16}$$

This bound records only the scale of the target perturbation. Zero-mean state noise does not imply $\mathbb{E}[\widetilde{Y}_t - Y_t \mid S_t, A_t] = 0$ when Q^* is nonlinear.

E.2 Finite-Sample Action-Value Estimation

Let the most recent N locally unbiased observations of action a, in the sense of part (iii) of Assumption 1, be

$$Z_{t-j,a}(s) = Q^*_{t-j}(s,a) + \omega_{t-j,a}(s), \qquad j = 0, \dots, N-1, \tag{17}$$

where $\mathbb{E}[\omega_{t-j,a}(s) \mid s] = 0$. The local average is

$$\widehat{Q}_{t,N}(s,a) = \frac{1}{N} \sum_{j=0}^{N-1} Z_{t-j,a}(s). \tag{18}$$

Substituting equation 17 yields

$$\widehat{Q}_{t,N}(s,a) = Q^*_t(s,a) + b_{N,a}(s) + \bar{\omega}_{N,a}(s), \tag{19}$$

where the staleness bias and the sampling error are

$$b_{N,a}(s) = \frac{1}{N} \sum_{j=0}^{N-1} \big(Q^*_{t-j}(s,a) - Q^*_t(s,a)\big), \qquad \bar{\omega}_{N,a}(s) = \frac{1}{N} \sum_{j=0}^{N-1} \omega_{t-j,a}(s). \tag{20}$$

For $a \neq a^*_t(s)$, define the one-observation pairwise noise

$$z_{t-j,a}(s) = \omega_{t-j,a^*_t(s)}(s) - \omega_{t-j,a}(s). \tag{21}$$

Then the normalized pairwise sensitivity $\chi_a(s)$ by the average one-observation variance is:

$$\frac{1}{N} \sum_{j=0}^{N-1} \mathrm{Var}\big(z_{t-j,a}(s) \mid s\big) = \gamma^2 L_Q^2 \sigma_e^2 \, \chi_a(s). \tag{22}$$

The factor $\chi_a(s)$ absorbs covariance between the two action-value estimates, state–action coverage, and estimator sensitivity. It is not implied as an equality by equation 16; that display only sets the noise scale that $\chi_a(s)$ renormalizes.

By part (iii), $\{z_{t-j,a}(s)\}_{j=0}^{N-1}$ is a conditional martingale-difference sequence, so cross-covariances vanish. The averaged pairwise error

$$\zeta_{N,a}(s) = \frac{1}{N}\sum_{j=0}^{N-1} z_{t-j,a}(s) \tag{23}$$

therefore satisfies

$$\mathrm{Var}\big(\zeta_{N,a}(s) \mid s\big) = \frac{1}{N^2}\sum_{j=0}^{N-1}\mathrm{Var}\big(z_{t-j,a}(s) \mid s\big) = \frac{\gamma^2 L_Q^2 \sigma_e^2\, \chi_a(s)}{N}. \tag{24}$$

E.3 Drift and Estimated Action Gaps

Under part (iii) of Assumption 1, action values are estimated by averaging the most recent N locally unbiased observations. The estimated pairwise gap then decomposes as

$$\widehat{\Delta}_{t,N,a}(s) = \Delta_{t,a}(s) + \beta_{N,a}(s) + \zeta_{N,a}(s), \tag{25}$$

where $\beta_{N,a}(s)$ is the pairwise staleness bias induced by drift and $\zeta_{N,a}(s)$ is the averaged pairwise sampling noise. From equation 24,

$$\mathrm{Var}\big(\zeta_{N,a}(s) \mid s\big) = \frac{\gamma^2 L_Q^2 \sigma_e^2\, \chi_a(s)}{N}. \tag{26}$$

The remainder of this subsection bounds $\beta_{N,a}(s)$, identifies the tracking MSE $\mathcal{E}_{N,a}(s) = \mathrm{Var}(\zeta_{N,a}(s) \mid s) + \beta_{N,a}(s)^2$, and obtains the Decision Confusion Index:

$$\mathrm{DCI}_{N,a}(s) = \frac{\sqrt{\mathrm{Var}(\zeta_{N,a}(s) \mid s) + \beta_{N,a}^2(s)}}{\Delta_{t,a}(s)}. \tag{27}$$

Its denominator is the true decision signal, while the numerator measures tracking uncertainty from finite-sample noise and staleness. Hence a larger $\mathrm{DCI}_{N,a}(s)$ indicates that the two actions are more difficult to distinguish reliably.

Part (iv) of Assumption 1 gives $|Q_{t+1}^*(s,a) - Q_t^*(s,a)| \le \delta$. Repeating the triangle inequality,

$$|Q_{t-j}^*(s,a) - Q_t^*(s,a)| \le j\delta. \tag{28}$$

Hence

$$|b_{N,a}(s)| \le \frac{1}{N}\sum_{j=0}^{N-1} j\delta = \frac{\delta(N-1)}{2}. \tag{29}$$

The pairwise staleness bias $\beta_{N,a}(s) = b_{N,a_t^*(s)}(s) - b_{N,a}(s)$ then satisfies

$$|\beta_{N,a}(s)| \le \delta(N-1). \tag{30}$$

Combining equation 19 and equation 23,

$$\widehat{\Delta}_{t,N,a}(s) = \widehat{Q}_{t,N}(s, a_t^*(s)) - \widehat{Q}_{t,N}(s,a) = \Delta_{t,a}(s) + \beta_{N,a}(s) + \zeta_{N,a}(s). \tag{31}$$

Since $\mathbb{E}[\zeta_{N,a}(s) \mid s] = 0$ and $\beta_{N,a}(s)$ is deterministic given s,

$$\mathcal{E}_{N,a}(s) := \mathbb{E}\Big[\big(\widehat{\Delta}_{t,N,a}(s) - \Delta_{t,a}(s)\big)^2 \mid s\Big] = \mathrm{Var}\big(\zeta_{N,a}(s) \mid s\big) + \beta_{N,a}(s)^2. \tag{32}$$

The Decision Confusion Index equation 4 is the pairwise ranking-difficulty index

$$\mathrm{DCI}_{N,a}(s) = \frac{\sqrt{\mathcal{E}_{N,a}(s)}}{\Delta_{t,a}(s)}.$$

The bias bound equation 30 yields the conservative upper bound

$$\mathcal{E}_{N,a}(s) \le \overline{\mathcal{E}}_{N,a}(s) := \frac{\gamma^2 L_Q^2 \sigma_e^2\, \chi_a(s)}{N} + \delta^2(N-1)^2. \tag{33}$$

The first term decreases with N, whereas the second increases with the temporal span of the historical window. When $\delta = 0$, larger N reduces tracking error. When $\delta > 0$, additional observations can inflate $\mathrm{DCI}_{N,a}(s)$ through staleness.

Through $\sigma_{N,a}(s) = \sqrt{\mathrm{Var}(\zeta_{N,a}(s) \mid s)}$, a large DCI places the ratio $\sigma_{N,a}(s)/\Delta_{t,a}(s)$ near a regime in which the Gaussian pairwise reversal probability of the next subsection approaches $1/2$, widening the learned-regret sandwich.

E.4 Pairwise Reversal Probability

Let $\sigma_{N,a}(s) = \sqrt{\mathrm{Var}(\zeta_{N,a}(s) \mid s)}$ and $d_N = \delta(N-1)$. By part (v) of Assumption 1, $\zeta_{N,a}(s)/\sigma_{N,a}(s)$ is standard normal given s. For the greedy rule in part (iii), the pairwise reversal event $\widehat{Q}_{t,N}(s,a) \geq \widehat{Q}_{t,N}(s, a_t^*(s))$ is exactly $\widehat{\Delta}_{t,N,a}(s) \leq 0$, or

$$\zeta_{N,a}(s) \leq -\Delta_{t,a}(s) - \beta_{N,a}(s). \tag{34}$$

Therefore

$$\psi_{N,a}(s) = \Phi\left(\frac{-\Delta_{t,a}(s) - \beta_{N,a}(s)}{\sigma_{N,a}(s)}\right), \tag{35}$$

where Φ is the standard normal distribution function. Because $-d_N \leq \beta_{N,a}(s) \leq d_N$ and Φ is increasing,

$$\psi^-_{N,a}(s) = \Phi\left(\frac{-\Delta_{t,a}(s) - d_N}{\sigma_{N,a}(s)}\right), \tag{36}$$

$$\psi^+_{N,a}(s) = \Phi\left(\frac{-\Delta_{t,a}(s) + d_N}{\sigma_{N,a}(s)}\right), \tag{37}$$

and $\psi^-_{N,a}(s) \leq \psi_{N,a}(s) \leq \psi^+_{N,a}(s)$. The lower envelope corresponds to the largest admissible gap bias $\beta_{N,a}(s) = d_N$, and the upper envelope to $\beta_{N,a}(s) = -d_N$.

E.5 Multi-Action Decision Error and Learned Regret

Let $\widehat{\pi}(s) = \arg\max_{a \in \mathcal{A}} \widehat{Q}_{t,N}(s,a)$. Under the conditional Gaussian errors in Assumption 1, $\widehat{\pi}(s) \neq a_t^*(s)$ if and only if at least one suboptimal action exceeds $a_t^*(s)$. Writing E_a for that pairwise event, then we have

$$p_{\mathrm{err}}(s) = \Pr\left(\bigcup_{a \neq a_t^*(s)} E_a\right). \tag{38}$$

A union is at least as likely as its most probable component, and the union bound caps the probability by one, so

$$\max_{a \neq a_t^*(s)} \psi^-_{N,a}(s) \leq p_{\mathrm{err}}(s) \leq \min\left\{1, \sum_{a \neq a_t^*(s)} \psi^+_{N,a}(s)\right\}. \tag{39}$$

Then, the exact regret of a learned policy is

$$\mathcal{R} = \mathbb{E}_S\left[\sum_{a \neq a_t^*(S)} \Delta_{t,a}(S)\, \widehat{p}_{N,a}(S)\right], \tag{40}$$

where $\widehat{p}_{N,a}(s) = \Pr(\widehat{\pi}(s) = a \mid s)$. Any incorrect decision loses at least $\Delta_{\min,t}(s) = \min_{a \neq a_t^*(s)} \Delta_{t,a}(s)$, and therefore

$$\mathcal{R} \geq \mathbb{E}_S[\Delta_{\min,t}(S)\, p_{\mathrm{err}}(S)] \geq \mathbb{E}_S\left[\Delta_{\min,t}(S) \max_{a \neq a_t^*(S)} \psi^-_{N,a}(S)\right] =: \underline{\mathcal{R}}. \tag{41}$$

Conversely, selecting a particular suboptimal action a requires $\widehat{Q}_{t,N}(s,a) \geq \widehat{Q}_{t,N}(s, a_t^*(s))$. Thus $\widehat{p}_{N,a}(s) \leq \psi_{N,a}(s) \leq \psi^+_{N,a}(s)$, and

$$\mathcal{R} \leq \mathbb{E}_S\left[\sum_{a \neq a_t^*(S)} \Delta_{t,a}(S)\, \psi^+_{N,a}(S)\right] =: \overline{\mathcal{R}}. \tag{42}$$

E.6 Fixed-Rule Ecological Regret

The regret of a fixed rule $h_k : \mathcal{S} \to \mathcal{A}$ is,

$$\mathcal{R}_k^{\text{fix}} = \mathbb{E}_S[Q_t^*(S, a_t^*(S)) - Q_t^*(S, h_k(S))]\,. \tag{43}$$

Group states by the action the rule selects. Define the usage rate $u_{k,a} = \Pr(h_k(S) = a)$, the conditional validity $\alpha_{k,a} = \Pr(a_t^*(S) = a \mid h_k(S) = a)$, and the conditional mean loss

$$\overline{\Delta}_{k,a}^{\text{vw}} = \mathbb{E}[Q_t^*(S, a_t^*(S)) - Q_t^*(S, a) \mid h_k(S) = a,\ a_t^*(S) \neq a]\,. \tag{44}$$

If $u_{k,a} = 0$, or if the rule never errs when it selects a, the corresponding summand is defined to be zero, so that the conditional expectation is not formed on a null event. On the remaining actions, recommending a contributes usage $u_{k,a}$, error probability $1 - \alpha_{k,a}$, and conditional mean regret $\overline{\Delta}_{k,a}^{\text{vw}}$. Summing over a gives

$$\mathcal{R}_k^{\text{fix}} = \sum_{a \in \mathcal{A}} u_{k,a}(1 - \alpha_{k,a})\,\overline{\Delta}_{k,a}^{\text{vw}}. \tag{45}$$

E.7 Architecture Dominance

Lemma 3 yields the sandwich equation 7 under Assumption 1. Comparing its endpoints with the ecological regret equation 5 of a fixed rule h_k gives the three exhaustive cases of Theorem 1: $\overline{\mathcal{R}} < \mathcal{R}_k^{\text{fix}}$ implies that the learned structure has lower expected regret; $\underline{\mathcal{R}} > \mathcal{R}_k^{\text{fix}}$ implies that the fixed rule has lower expected regret; and $\underline{\mathcal{R}} \leq \mathcal{R}_k^{\text{fix}} \leq \overline{\mathcal{R}}$ is indeterminate from pairwise information alone. Thus the decision problem separates into a learning-dominant region, a fixed-structure-dominant region, and an indeterminate region. Learned decision quality is jointly governed by the current action gap $\Delta_{t,a}(s)$, observation noise σ_e, sample size N, pairwise sensitivity $\chi_a(s)$, and environmental drift δ. These quantities determine $\mathrm{DCI}_{N,a}(s)$ and, through the pairwise reversal probabilities, the regret bounds of the learned structure. A fixed rule is advantageous only when its ecological regret equation 5 is small relative to those bounds.

E.8 Connecting λ to DCI

Remark 2 (Qualitative map). λ does not enter Theorem 1 directly. In the simulator of Appendix A.1, the return innovation is

$$r_t = s_\lambda \sqrt{\nu_t \Delta t}\, X_t - \frac{1}{2}\nu_t \Delta t + n_\lambda \sqrt{\nu_t \Delta t}\, \varepsilon_t, \qquad \frac{s_\lambda^2}{n_\lambda^2} = \lambda.$$

Increasing λ shifts return innovation from unpredictable noise toward the persistent predictive component while leaving the feature and action interfaces fixed. If this increases the decision-relevant information transmitted through the observable features, effective action gaps become larger relative to ranking uncertainty, moving the environment toward lower-confusion regimes. The map is qualitative: the preceding bounds do not imply that $\mathrm{DCI}_{N,a}(s)$ is strictly decreasing in λ.

Remark 3 (Optional empirical proxies). When diagnostics are desired, the following training-partition proxies can be reported alongside architecture comparisons; none is required by Theorem 1.

1. *Feature leakage:* $\max_m \left|\widehat{\mathrm{corr}}(X_t, f_t^{(m)})\right|$ on the training split.
2. *Gap proxy:* for a held-out oracle or a strong teacher policy, the mean pairwise return advantage of the preferred micro-action over alternatives, as a surrogate for $\Delta_{t,a}$.
3. *Confusion proxy:* the ratio of a rolling value-estimate standard error to that gap proxy, mirroring the form of $\mathrm{DCI}_{N,a}$.

These quantities are a way to read a λ-grid as a sweep of decision-confusion regimes. They do not substitute for the regret bounds, and they are not claimed as estimators of $\mathrm{DCI}_{N,a}(s)$.