Capstone Design Project — Final Report

The Effect of Reward Function Design in RL-Based BTC Grid Trading: Pareto Frontier, Winner Reversal, and OOS Robustness of Prospect-Theoretic Reward

Chanhee Lee
Department of Computer Science
Seoul National University of Science and Technology
happilyfly@seoultech.ac.kr

June 2026

Abstract

We quantitatively study the effect of *reward function design* on PPO-based reinforcement-learning policies for ATR-proportional grid trading on the BTC/USDT 1-hour market. Four reward variants—symmetric (`sym`), asymmetric ($\beta = 3.42$, `asym`), Differential Sharpe Ratio ($\eta = 1/28$ h, `dsr`; Moody and Saffell 2001), and prospect-theoretic ($\alpha = 0.68$, $\lambda = 3.30$, `pt`; Kahneman and Tversky 1979)—are each trained with 10 seeds × 1M steps and evaluated on three environments: (i) single-split validation (Val 2021–2023), (ii) 6-fold combinatorial purged cross-validation [López de Prado, 2018] 15 paths, and (iii) the unsealed out-of-sample Test set (2024–2026, BTC \$42K → \$75K bull regime).

We report four key findings. **(1)** All four variants exceed the ATR baseline (1.378) by +21–36% on Val, yet partition into two clusters {`sym`, `dsr`} (aggressive) vs. {`asym`, `pt`} (conservative); within-cluster Cohen's $|d| < 0.3$ versus across-cluster $|d| > 0.8$. **(2)** Mechanism: reward loss asymmetry directly determines trade frequency and holding time (policy distance ratio 2.22×), and the sliding-window memory structure of DSR results in 2–6× higher hold rate than other variants. **(3)** Winner reverses across evaluation environments: Val=`sym`(1.87) → CPCV=`dsr`(1.41, $p < 10^{-9}$) → Test=`pt`(0.34–0.37, consistent across two model sources, $p < 0.002$). **(4)** The OOS robustness of `pt` arises from a short-holding mechanism: loss aversion $\lambda = 3.30$ combined with concave gain $\alpha = 0.68$ trains policies to exit within an average of 1.4 h (max 6 h), thereby avoiding the sell-side timing risk of the unseen bull regime. In contrast, `dsr` learns long holding (mean 4.58 h, max 169 h = 7 days) and records the worst OOS performance—demonstrating that the same reward formulation simultaneously drives in-sample advantage and OOS failure. The Val–Test distribution shift is highly significant (KS $p < 10^{-10}$, −27% variance), while policies' actions remain nearly invariant between Val and Test ($\Delta \leq 5\%$). That is, the policies output the same actions, but those actions produce different outcomes in the different market regimes (Val = range-bound vs. Test = bull); the OOS performance difference is driven by the market distribution shift, not by policy change.

Our contribution is twofold. First, we demonstrate quantitatively that reward design takes effect along a *trade-off dimension in risk profile* rather than a single Sharpe "alpha

source." Second, we provide quantitative evidence that Kahneman-Tversky prospect theory confers OOS safety on RL trading policies. Code and data: https://github.com/cosmicpotato2047/capstone-rl-trading.

Keywords: Reinforcement Learning, Grid Trading, PPO, Reward Design, Prospect Theory, Differential Sharpe Ratio, Combinatorial Purged CV, Out-of-Sample Generalization, Cryptocurrency

Contents

1 Introduction **5**

1.1 Motivation and Research Question 5

1.2 Contributions 5

1.3 Paper Structure 6

2 Related Work **6**

2.1 Market Making and Grid Trading 6

2.2 Reinforcement Learning for Trading 7

2.3 Reward-Design Theory 7

2.4 Evaluation Methodology 8

3 Method **9**

3.1 MDP Formulation 9

3.2 Trading-Environment Dynamics 10

3.3 ATR Baseline 11

3.4 Four Reward Variants 11

3.5 PPO Training and the Stabilization Package 13

3.6 Evaluation Protocol 13

3.7 Pre-Registered Hypotheses 14

4 Phase 1: Policy Saturation under Symmetric Reward **15**

4.1 The Saturation Phenomenon 15

4.2 Decisive Ablation: "RL = Fixed $[1.0, 0.0]$" 15

4.3 Mechanism: ATR Formula's Absorption of Degrees of Freedom 16

4.4 Motivation for This Paper 16

5 Pareto Frontier and Two-Cluster Separation **17**

5.1 Experimental Setup 17

5.2 Per-Variant Results: Four Different Metric Winners 17

5.3 Pareto Frontier Visualization 18

5.4 Statistical Separation of the Two Clusters 19

5.5 Policy Distance: Confirmation at the Action Level 19

5.6 Confirming the Multi-Dimensional Trade-off and Hypothesis Evaluation 20

5.7 Environment Dependence of the Phase 2 Prior Evidence 21

6 Mechanism Quantification **21**

6.1 Trajectory Dataset 21

6.2 Trade Frequency: Direct Effect of Reward Asymmetry 22

6.3 Hold Rate: The Mechanism of DSR's Sliding-Window Memory 23

6.4 Per-Regime Action Distribution 23

6.5 Counterfactual State-grid Verification 25

6.6 Summary of the Causal Chain 26

7 Robustness Analysis **27**
7.1 Slippage Robustness . . . 27
7.1.1 Setup . . . 27
7.1.2 Per-variant Results: Uniform $\sim$ 12% Decay . . . 27
7.1.3 MDD Is Almost Unchanged . . . 27
7.1.4 Cluster Preservation: Robustness of the Two Clusters to Slippage . . . 29
7.1.5 Fair Comparison Against the ATR Baseline . . . 29
7.1.6 Section Summary and Partial Verdict on H4 . . . 29
7.2 Multi-Split Evaluation via CPCV . . . 30
7.2.1 Setup . . . 30
7.2.2 Per-Variant CPCV Statistics . . . 30
7.2.3 Winner Reversal: Variant Ranking Depends on the Evaluation Method . . . 30
7.2.4 Cluster Preservation: Sharper under Multi-split . . . 32
7.2.5 Mechanism Interpretation and H4 Verdict . . . 32
7.3 Out-of-Sample Evaluation and Prospect Theory's Robustness . . . 33
7.3.1 Test Unsealing and Evaluation Protocol . . . 33
7.3.2 Test Results: Per-Source Per-Variant Statistics . . . 33
7.3.3 Winner Reversal across Three Environments . . . 35
7.3.4 Val $\rightarrow$ Test Generalization Gap . . . 35
7.3.5 Mechanism: Hold Duration and OOS Regime Adaptation . . . 36
7.3.6 Policy Stability vs Market Distribution Shift . . . 37
7.3.7 B&H Baseline Comparison: Limitation Statement . . . 38
7.3.8 H5: Statement and Evidence . . . 38

8 Discussion **39**
8.1 Quantifying the Val–Test Distribution Shift . . . 39
8.2 Environment Dependence of the Phase 2 Prior Evidence . . . 40
8.3 Limits of Single-Metric Winners; Toward Multi-Environment Evaluation . . . 41
8.4 From Behavioral Economics to RL OOS Safety . . . 42
8.5 Limitations and Future Work . . . 42

9 Conclusion **43**

1 Introduction

Cryptocurrency markets exhibit micro-oscillations in hourly price that are largely orthogonal to long-term trends, motivating *grid trading* strategies that harvest revenue from volatility itself rather than from directional prediction. Grid trading is a simplified form of market making [Avellaneda and Stoikov, 2008], ranging from fixed-grid rules to volatility-adaptive variants that scale grid width by the Average True Range (ATR) [Wilder, 1978]. This work studies the latter: an ATR-proportional grid system whose grid width and limit-order placement are determined by a reinforcement-learning policy.

1.1 Motivation and Research Question

This project ran in two prior phases (Phase 1 and Phase 2) before the work reported here. Phase 1 hypothesized that PPO RL would discover dynamic volatility adaptation and outperform a fixed-grid baseline. Post-hoc behavioral analysis, however, revealed that the trained policy converged to a near-constant action $[1.0, 0.0]$ (§4), and that the Bayesian-optimized ATR coefficient itself was the principal source of performance. This negative finding suggested that the symmetric-reward + ATR-proportional combination structurally absorbs the policy's degrees of freedom, motivating the new research question of this paper:

> **In BTC grid trading, how does the design of the reward function affect the behavioral pattern and generalization performance of RL policies? In particular, do loss-asymmetric (asymmetric, prospect-theoretic) or risk-adjusted-return (DSR) reward formulations yield meaningful differences in single-metric advantage or out-of-sample robustness?**

1.2 Contributions

This work makes the following four contributions.

1. **A statistically rigorous comparison framework for reward variants.** Four reward functions (`sym`, `asym`, `dsr`, `pt`) are trained in an identical environment with 10 seeds × 1M steps, and pairwise statistical tests (Cohen's d, bootstrap P(A>B), IQM, 5% CVaR) are applied following Henderson et al. [2018]'s reproducibility guidelines.

2. **Pareto-frontier discovery (multi-dimensional trade-off).** Rather than a single "best variant" winner, the four variants statistically partition into two clusters {`sym`, `dsr`} vs. {`asym`, `pt`} forming a Pareto-like frontier on the Sharpe–MDD plane (§5). This outcome falls into none of the pre-registered scenario branches (A optimistic / B neutral / C pessimistic), so we name it *Scenario D* and formalize it in §5.6.

3. **Causal chain: reward → behavior → outcome.** Policy-level differences across variants are quantified on 1.04M step trajectories via a policy-distance matrix (within-cluster 0.129 vs. across-cluster 0.286), regime- specific trade/hold statistics, and action distribution comparison, confirming the mechanism reward loss asymmetry → reduced trade frequency → conservative cluster (§6).

4. **Winner reversal across evaluation environments and OOS robustness of prospect theory.** The winner reverses completely across single-split Val (`sym` 1st) → CPCV multi-split (`dsr` 1st) → the unsealed Test (`pt` 1st, consistent across two sources $p < 0.002$) (§7). Trajectory analysis attributes `pt`'s OOS robustness to its loss aversion ($\lambda = 3.30$) and concave utility ($\alpha = 0.68$), which train policies to exit within an average of 1.4 h and thereby avoid sell-side timing risk in the unseen bull regime. This provides quantitative evidence that the Kahneman–Tversky prospect theory [Kahneman and Tversky, 1979] confers OOS safety on RL trading policies.

1.3 Paper Structure

Section 2 surveys related work on RL trading, reward design theory, and evaluation methodology. Section 3 details the environment design (MDP, ATR formula, four reward variants) and PPO training. Section 4 briefly reproduces the Phase 1 negative finding ("RL = Fixed [1.0, 0.0]") as motivation. Section 5 reports the two-cluster Pareto-frontier result on Val (Scenario D), while Section 6 quantifies the mechanism. Section 7 presents the robustness trio: slippage (§7.1), CPCV multi-split (§7.2), and the central contribution of this paper—the Test OOS evaluation and the mechanistic basis of prospect theory's robustness (§7.3). Section 8 discusses distribution shift quantification and limitations, and Section 9 concludes.

2 Related Work

This paper lies at the intersection of four areas: (i) market making and grid trading, (ii) reinforcement-learning-based trading, (iii) reward-design theory, and (iv) evaluation methodology for trading policies. This section surveys the key prior work in each area and clarifies where our contribution lies.

2.1 Market Making and Grid Trading

The academic ancestor of grid trading is the market-making model of Avellaneda and Stoikov [2008]. Their analysis derives a closed-form optimal policy for a risk-averse market maker who faces inventory risk and must place *asymmetric* bid/ask quotes around the mid-price. Two key insights are that (a) the bid–ask spread is a function of volatility and remaining time horizon, and (b) inventory skew shifts the quote midpoint asymmetrically. Grid trading is a simplified practical variant of this model in which orders are placed on a pre-defined ladder of price levels in order to monetize micro-volatility.

The standard mechanism for adapting grid widths to market volatility is the Average True Range [Wilder, 1978] (ATR; 14-period moving average is the default). ATR is a smoothed average of the true price range and provides a simple, robust rule that automatically widens the grid in volatility-cluster regimes. ATR-proportional grids of the form $\Delta = c \cdot \text{ATR}$ (with proportional coefficient c) are widely used in practice and serve as the baseline of this paper.

Most prior academic work in this area is limited to the analysis of *rule-based* policies, with relatively few attempts to learn the grid-width decision from data. This paper adopts a structure

in which the proportional coefficient and the number of grid splits in the ATR formula are dynamically determined by an RL policy, while quantitatively analyzing how the design of the reward function influences the resulting learned policy's behavior.

2.2 Reinforcement Learning for Trading

The standard DRL baseline for cryptocurrency trading is the PPO/A2C/DQN comparison framework of Zhang et al. [2020], which feeds price time series directly into a PPO policy and shows consistent improvements over simple buy-and-hold (B&H) and moving-average crossover strategies. The ACM survey of Sun et al. [2023] and the more recent review of Liu [2025] both identify two open problems in the area: *(a) interpretability of policy behavior and (b) generalization across market regimes*. This paper addresses both directly: §6 quantifies the causal chain from reward form to behavioral difference, and §7 documents out-of-sample (OOS) generalization phenomena that are not captured by single-split validation alone.

Four prior works are directly relevant. Liu et al. [2020] feed the output of an LSTM price predictor into the state of a PPO trading policy in a two-stage pipeline. The limitation of this approach is that prediction errors propagate directly into trading decisions; this paper bypasses the limitation by *eliminating* the price-prediction stage and using a volatility-adaptive grid structure instead. Yasin and Gill [2024] compare DQN/PPO/A2C with 20 technical indicators, but their action space is discrete buy/sell with no hold action. This paper adopts a continuous two-dimensional action (aggressiveness and profit target) that supports fine-grained grid control. Pham et al. [2025] use DQN to select one of five predefined strategies as a meta-strategy, but the strategies themselves are not learned. Bandarupalli [2025] augment PPO with a volatility penalty and a transaction-cost term, but stay at the level of hyperparameter tuning within a single reward variant and do not compare across reward *forms*, which is the focus of this paper.

A particularly relevant prior work is Gort et al. [2022], who empirically show that *single-split train/test results do not guarantee OOS generalization* in cryptocurrency DRL trading, and recommend CPCV (Combinatorial Purged Cross-Validation) and DSR (Deflated Sharpe Ratio). This paper adopts those recommendations, using Val (2021–2023), CPCV (6-fold inside the Train partition), and Test (2024 onward, unsealed) environments concurrently. Our *winner reversal* finding (Val `sym` → CPCV `dsr` → Test `pt`) is a quantitative confirmation that Gort et al.'s concern reappears in the dimension of reward-variant selection.

In summary, whereas prior DRL trading literature focuses on comparing algorithms and state designs under a *single* reward function, this paper performs a controlled comparison of *four reward functions* on top of the same algorithm (PPO) and the same environment, and separately measures their effects on in-sample / multi-split / OOS environments.

2.3 Reward-Design Theory

The four reward variants compared in this paper draw from three theoretical lineages in reinforcement learning and behavioral economics. (i) Symmetric (`sym`) is the standard PnL-return reward and serves as our baseline. (ii) Asymmetric (`asym`) places a positive weight $\beta > 1$ on losses, a simple first-order approximation of a risk-averse utility. (iii) Differential Sharpe Ratio

(`dsr`) is the formulation of Moody and Saffell [2001]. (iv) Prospect-theoretic reward (`pt`) directly uses the utility function of Kahneman and Tversky [1979] and Tversky and Kahneman [1992], $v(x) = \text{sgn}(x)|x|^\alpha$ for $x \geq 0$ and $-\lambda|x|^\alpha$ for $x < 0$, as a per-step reward.

The DSR formulation of Moody and Saffell [2001] uses the *differential change of the online Sharpe-ratio estimator* as the step reward, tracking first and second moments via exponentially-weighted moving averages (EWMA, decay rate η). This has two implications for RL training: (a) it directly optimizes a risk-adjusted return rather than raw PnL, and (b) the sliding-window memory imposes a *multi-step temporal dependency* on the policy. As we show in §6, this memory structure is what causes the `dsr` policy to learn 2–6× longer holding times than the other variants—a property that becomes the in-sample advantage of §7 and the OOS failure of §7.3 — the same causal mechanism viewed from two evaluation angles.

Prospect theory [Kahneman and Tversky, 1979] is a foundational model of human decision making under risk, with two characteristic parameters: (a) loss aversion $\lambda \approx 2.25$ (standard value from Tversky and Kahneman 1992) and (b) the concavity exponent $\alpha \approx 0.88$. While the theory has been widely used to describe human trader behavior (e.g., the disposition effect — holding losing positions too long while realizing gains too early), *direct use of prospect-theoretic utility as an RL reward* is rare in the literature. This paper directly implements the Kahneman and Tversky [1979] utility as an RL reward (with $\alpha = 0.68$, $\lambda = 3.30$, tuned by Optuna) and quantitatively establishes the OOS robustness that the resulting policy exhibits.

Ng et al. [1999]'s policy-invariance theorem of potential-based reward shaping states that "a monotone transformation of the reward signal does not alter the policy learned by RL." At first glance, this paper's comparison of four reward variants may appear to conflict with the theorem, but it does not: the `sym` → `asym`/`pt` transformations are *non-potential-based* (they apply an asymmetric scale on the negative half-line), while `dsr` is not a simple per-step reward transformation but a *state-extended* reward (the sliding-window moments effectively become part of the state). Our reward-variant comparison thus lies outside the scope of the Ng et al. theorem, and indeed we confirm in §6 that policy behavior differs across variants in a statistically significant manner.

2.4 Evaluation Methodology

The two principal pitfalls in evaluating RL trading policies are (a) backtest overfitting and (b) seed dependence of results. Henderson et al. [2018] empirically demonstrate that seed variability alone can produce large differences for the same algorithm, and propose 5–10 seed statistical comparisons, IQM (Interquartile Mean), and bootstrap confidence intervals as standard recommendations. This paper uses 10 seeds per variant (Val/exp033) and 100 models (Test), and applies pairwise Cohen's d and bootstrap $\text{P}(A > B)$ to all comparisons.

López de Prado [2018] (Ch. 8) proposes *Combinatorial Purged Cross-Validation* (CPCV), which avoids train/test leakage in time-series data while leveraging $\binom{N}{k}$ split combinations to strengthen the generalization guarantees of single-split results. We adopt 6-fold CPCV (15 paths) in §7.2.

Bailey and López de Prado [2014]'s *Deflated Sharpe Ratio* (DSR^*) is a statistical-significance test on the Sharpe ratio that corrects for the number of backtest trials and for the skewness/kurtosis of the underlying distribution. We apply Bonferroni 4-way correction together with DSR^* testing to the multiple comparisons of CPCV results (§7.2), and confirm that all four reward variants are significant relative to the ATR baseline at $p < 0.001$.

The methodological distinctiveness of this paper is the *simultaneous use of three evaluation environments* (Val / CPCV / Test) that allowed us to discover the winner reversal. Studies relying solely on single-split or solely on CPCV cannot expose OOS generalization failure, and studies relying solely on OOS cannot expose the in-sample conservatism of `pt`. The simultaneous application of three environments is a methodological contribution: it generalizes the multi-evaluation recommendations of Henderson et al. [2018] and Gort et al. [2022] to the single dimension of reward-variant comparison.

3 Method

3.1 MDP Formulation

We cast single-asset dynamic grid trading on the BTC/USDT 1-hour series as a finite-horizon Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$. An episode runs for 720 hours (30 days) starting from a randomly chosen step within the training/validation range, and each step t corresponds to a 1-hour price bar.

State $\mathcal{S} \subset \mathbb{R}^7$. The observation $\mathbf{s}_t = (s_t^{(0)}, \ldots, s_t^{(6)})$ couples market and portfolio state:

$$s_t^{(0)} = \log\Big(p_t \,/\, \overline{p}_{t,168}\Big) \qquad \text{(price vs. 1-week mean)} \tag{1}$$

$$s_t^{(1)} = \frac{\bar{p}_t^{\text{avg}} - p_t}{\bar{p}_t^{\text{avg}}} \qquad \text{(divergence from average cost)} \tag{2}$$

$$s_t^{(2)} = \frac{h_t \cdot p_t}{C_0} \qquad \text{(holdings value / capital)} \tag{3}$$

$$s_t^{(3)} = \frac{c_t}{C_0} \qquad \text{(cash ratio)} \tag{4}$$

$$s_t^{(4)} = \frac{\text{ATR}_{168}(p)}{p_t} \qquad \text{(volatility)} \tag{5}$$

$$s_t^{(5)} = \frac{p_t - p_{t-72}}{p_{t-72}} \qquad \text{(short-term trend, 3d)} \tag{6}$$

$$s_t^{(6)} = \frac{p_t - p_{t-720}}{p_{t-720}} \qquad \text{(long-term trend, 30d)} \tag{7}$$

Here p_t is the bar close, $\bar{p}_t^{\text{avg}}$ is the average cost of the current position (set to 0 when no position is held, in which case $s_t^{(1)}$ is also 0), h_t is the BTC quantity, c_t the cash balance, and $C_0 = 10{,}000$ USDT is the initial capital. ATR_{168} is the Average True Range [Wilder, 1978] computed over a 168-bar (1-week) window, and $\overline{p}_{t,168}$ is the corresponding simple moving average. All components are pre-normalized by a 168-bar rolling z-score computed on the training set and clipped to

$[-5, 5]$.

Action $\mathcal{A} = [0, 1]^2$. At each step the policy outputs a continuous 2D action $\mathbf{a}_t = (a_t^{(0)}, a_t^{(1)})$. The first coordinate $a_t^{(0)}$ (*aggressiveness*) controls two buy limit orders, and $a_t^{(1)}$ (*profit target*) controls two sell limit orders. We avoid a discrete buy/sell/hold action space because fine-grained grid width adjustment and volatility adaptation are most naturally expressed in continuous values; discretization would artificially limit the expressiveness of the action space.

ATR-proportional grid formula. Grid widths and positions are determined through volatility-normalized action coordinates. With $r_t = \mathrm{ATR}_{168}(p)/p_t$, current price p_t, and average cost $\bar{p}_t^{\mathrm{avg}}$,

$$g_t^{\mathrm{buy\_hi}} = r_t \cdot (A_b + B_b \cdot a_t^{(0)}) \tag{8}$$

$$g_t^{\mathrm{buy\_lo}} = r_t \cdot (C_b + D_b \cdot a_t^{(0)}) \tag{9}$$

$$g_t^{\mathrm{sell\_mkt}} = r_t \cdot (A_s + B_s \cdot a_t^{(1)}) \tag{10}$$

$$g_t^{\mathrm{sell\_cost}} = r_t \cdot (C_s + D_s \cdot a_t^{(1)}) \tag{11}$$

Quote prices:

$$q_t^{\mathrm{buy\_hi}} = p_t(1 - g_t^{\mathrm{buy\_hi}}), \quad q_t^{\mathrm{buy\_lo}} = p_t(1 - g_t^{\mathrm{buy\_lo}}) \tag{12}$$

$$q_t^{\mathrm{sell\_mkt}} = p_t(1 + g_t^{\mathrm{sell\_mkt}}), \quad q_t^{\mathrm{sell\_cost}} = \bar{p}_t^{\mathrm{avg}}(1 + g_t^{\mathrm{sell\_cost}}) \tag{13}$$

The coefficients $\{A_b, B_b, C_b, D_b, A_s, B_s, C_s, D_s\}$ are fixed environment constants, with values reported in §3.3. Only the `sell_cost` quote is anchored at the average cost; the choice explicitly exposes "break-even recovery" to the policy.

Reward. The baseline (`sym`, symmetric) form is the 1-step PnL ratio:

$$r_t^{\mathrm{sym}} = \frac{E_t - E_{t-1}}{C_0}, \quad E_t = c_t + h_t \cdot p_t \tag{14}$$

Transaction fees are already deducted from c_t inside `_execute_buy` and `_execute_sell`, so no separate subtraction appears in the reward. All four reward variants in this paper are defined as functions of r_t^{sym} and detailed in §3.4.

3.2 Trading-Environment Dynamics

This subsection details order execution, the cycle definition, and the transaction-cost model.

Limit-order execution. Fill decisions for the four quotes use the next bar's high/low $(p_{t+1}^{\mathrm{H}}, p_{t+1}^{\mathrm{L}})$:

$$\begin{aligned}
&p_{t+1}^{\mathrm{L}} \le q_t^{\mathrm{buy_hi}} \Rightarrow \text{fill at } q_t^{\mathrm{buy_hi}} \\
&p_{t+1}^{\mathrm{L}} \le q_t^{\mathrm{buy_lo}} \Rightarrow \text{fill at } q_t^{\mathrm{buy_lo}} \\
&p_{t+1}^{\mathrm{H}} \ge q_t^{\mathrm{sell_mkt}} \Rightarrow \text{fill at } q_t^{\mathrm{sell_mkt}} \\
&p_{t+1}^{\mathrm{H}} \ge q_t^{\mathrm{sell_cost}} \Rightarrow \text{fill at } q_t^{\mathrm{sell_cost}}
\end{aligned} \tag{15}$$

When both buy and sell quotes are filled in the same bar we apply a *sell-first* rule, prioritizing inventory liquidation. The fill price is the *quote price itself*, not the next-bar extreme. Filling at the next-bar extreme injects favorable bias into the simulation; we therefore discarded an earlier "next_low/next_high" execution rule and use only the limit-order rule throughout this paper.

Cycle definition. A cycle starts when the *first buy fills while holdings = 0* and ends when subsequent sells return holdings to 0. Multiple partial buys/sells can occur inside a single cycle, and the cycle return is cycle_pnl $= (c_{\text{end}} - c_{\text{start}})/c_{\text{start}}$. At the start of each cycle, the slot size used for the cycle is set:

$$\text{slot\_size}_{\text{cycle}} = \frac{c_{\text{start}}}{n_{\text{splits}}}, \quad \text{order\_size} = \frac{\text{slot\_size}}{n_{\text{buy\_orders}}} \tag{16}$$

We use $n_{\text{splits}} = 2$ and $n_{\text{buy\_orders}} = 2$ (best of an Optuna 50-trial search during Phase 2 preliminary tuning). `threshold_basis` is fixed to `avg_price`, meaning the "liquidate all" detection uses the average cost.

Transaction-cost model. Transaction costs are set to *0.05% per side* (Binance maker fee). On a buy, the filled quantity is $\Delta h = \text{order\_size} \cdot (1 - f)/q$ and the cash change is $\Delta c = -\text{order\_size}$, so the fee is already reflected in c_t. In §7.1 (exp033) we add an additional slippage $\sigma = 0.02\%$ that modifies fill prices to $q^{\text{buy}} \cdot (1 + \sigma)$ and $q^{\text{sell}} \cdot (1 - \sigma)$. Slippage is not applied to the base environment and appears only in the robustness analysis.

3.3 ATR Baseline

The comparison baseline *shares the environment and grid formula* with our RL policies but determines actions via a rule. Concretely, the ATR baseline reuses the grid formulas of §3.1 for $g^{\text{buy\_hi}}, g^{\text{buy\_lo}}, g^{\text{sell\_mkt}}, g^{\text{sell\_cost}}$ with all RL action degrees of freedom *zeroed* ($a_t^{(0)} = a_t^{(1)} = 0$, so that the B_b, D_b, B_s, D_s terms drop and the grid widths are fixed time-independent constants determined entirely by A_b, C_b, A_s, C_s). The coefficient values come from the same Optuna search:

Table 1: Grid coefficients of the ATR baseline (best of 50 Optuna trials).

Coef.	Value	Role	Coef.	Value	Role
A_b	1.665	buy_hi base width	A_s	0.285	sell_mkt base width
C_b	6.070	buy_lo base width	C_s	1.951	sell_cost base width
B_b	1.748	buy_hi action sensitivity	B_s	1.950	sell_mkt action sens.
D_b	18.683	buy_lo action sensitivity	D_s	7.500	sell_cost action sens.

The ATR baseline yields Val Sharpe 1.378 without slippage, dropping to 0.835 when slippage of 0.02% is added.

3.4 Four Reward Variants

We now define the four reward variants that form the principal comparison unit of this paper. All are expressed as functions of $x_t := r_t^{\text{sym}} = (E_t - E_{t-1})/C_0$ and differ *only* in functional form; the state, action, ATR grid, and fee model are otherwise identical.

(i) Symmetric (`sym`).

$$r_t^{\text{sym}} = x_t = \frac{E_t - E_{t-1}}{C_0} \tag{17}$$

No hyperparameter. Baseline of this paper.

(ii) Asymmetric (`asym`).

$$r_t^{\text{asym}} = \begin{cases} x_t & \text{if } x_t \geq 0 \\ \beta \cdot x_t & \text{if } x_t < 0 \end{cases}, \quad \beta > 1 \tag{18}$$

A first-order approximation of a risk-averse utility, weighting losses by β. An Optuna 30-trial $\times$ 200k-step search over the range $[1.0, 4.0]$ to maximize Val Sharpe selected the best $\beta = 3.420$.

(iii) Differential Sharpe Ratio (`dsr`). The formulation of Moody and Saffell [2001]. Given an EWMA decay rate η, online estimates of the first and second return moments evolve as

$$A_t = A_{t-1} + \eta\,(x_t - A_{t-1}) \tag{19}$$

$$B_t = B_{t-1} + \eta\,(x_t^2 - B_{t-1}) \tag{20}$$

$$\Delta A_t = x_t - A_{t-1}, \quad \Delta B_t = x_t^2 - B_{t-1} \tag{21}$$

The DSR step reward is the differential of the on-line Sharpe estimator $\hat{S}_t = A_t / \sqrt{B_t - A_t^2}$:

$$r_t^{\text{dsr}} = \frac{B_{t-1}\,\Delta A_t - \frac{1}{2} A_{t-1}\,\Delta B_t}{(B_{t-1} - A_{t-1}^2)^{3/2}} \tag{22}$$

A 30-trial Optuna search over $[1/720, 1/24]$ (log-uniform) selected the best $\eta = 0.0352 \approx 1/28\,\text{h}$, an EWMA window of approximately 1.2 days.

(iv) Prospect-theoretic (`pt`). The utility function of Kahneman and Tversky [1979] and Tversky and Kahneman [1992] applied directly as a step reward:

$$r_t^{\text{pt}} = \begin{cases} \operatorname{sgn}(x_t)\,|x_t|^{\alpha} & \text{if } x_t \geq 0 \\ -\lambda\,|x_t|^{\alpha} & \text{if } x_t < 0 \end{cases}, \quad 0 < \alpha \leq 1, \quad \lambda \geq 1 \tag{23}$$

$\alpha < 1$ encodes the concavity of the utility (diminishing marginal utility of gains), and $\lambda > 1$ encodes loss aversion. The standard values from Tversky and Kahneman [1992], fitted to human behavior, are $\alpha = 0.88$ and $\lambda = 2.25$. The Val-Sharpe-maximizing values for our RL policy, however, are $\alpha = 0.683$ and $\lambda = 3.303$, found by Optuna 30 trials over $\alpha \in [0.5, 1.0]$ and $\lambda \in [1.0, 4.0]$. The fact that RL prefers a *more concave* utility and *stronger loss aversion* than the human standard is consistent with the mechanism in §6 and the OOS robustness reported in §7.3.

Optuna search summary. For each of the three variants `asym`, `dsr`, `pt` we ran 30 trials $\times$ 200k steps using a TPE sampler [Akiba et al., 2019] (seed=42, MedianPruner with $n_{\text{startup}} = 5$).

The total training cost was $3 \times 30 \times 200\text{k} = 18\text{M}$ steps, taking approximately 2 hours on a Windows 11 CPU with $n_{\text{envs}} = 4$. `sym` has no hyperparameter and skips this stage.

3.5 PPO Training and the Stabilization Package

Policy training uses PPO [Schulman et al., 2017] from stable-baselines3 [Raffin et al., 2021]. Standard PPO hyperparameters ($\gamma = 0.99$, $\lambda_{\text{GAE}} = 0.95$, clip $\epsilon = 0.2$, v_f coef $= 0.5$, gradient clip $= 0.5$, MLP policy $[64, 64]$, batch=256, $n_{\text{epochs}} = 10$, $n_{\text{steps}} = 4096$, $n_{\text{envs}} = 4$) follow stable-baselines3 defaults. To mitigate the reward sparsity of our environment (the per-step reward is near 0 when no fill occurs) and the late-stage policy collapse observed in Phase 2 (exp020–022), we add the following *stabilization package* (exp030):

1. **Learning-rate linear decay**: α_t linearly decays from 3×10^{-4} to 1×10^{-5} over training. This bounds the size of late-stage policy updates and dampens post-best collapse.
2. **Entropy coefficient annealing**: c_{ent} linearly anneals from 0.01 to 0.001, ensuring sufficient initial exploration while encouraging convergence to a near-deterministic policy late in training.
3. **Target KL early stop**: target_kl $= 0.02$ measures the KL divergence after each minibatch update and aborts the update for that batch if the threshold is exceeded, preventing PPO from making myopic large policy changes.
4. **Best-model checkpoint**: Val Sharpe is evaluated every 50k steps ($n_{\text{eval\_episodes}} = 5$) and the weights at the best Val Sharpe are saved separately, providing protection against late-stage collapse.

The package was developed in exp030 (with `sym` reward) and produced Val Sharpe best 1.974 at step 550k, falling to final 1.209 at step 1M. Because the collapse pattern after step 700k persists despite the package, *all evaluations in this paper use the best checkpoint*; the final model is never used.

3.6 Evaluation Protocol

We use *three evaluation environments concurrently*. The data partition is strict time-order to prevent leakage:

Table 2: Data partition.

Partition	Range	Bars (approx.)
Train	2017-08-17 – 2020-12-31	∼30,000
Val (single-split)	2021-01-01 – 2023-12-31	∼26,000
Test (sealed)	2024-01-01 – 2026-04	∼20,000

Val (single-split). The main comparison environment of §5. For each reward variant we train 10 seeds × 1M steps and evaluate the best_model on the entire Val range with $n_{\text{eval\_episodes}} = 5$, producing per-seed Sharpe/MDD/Calmar/trade-count/return.

CPCV 6-fold. The multi-split environment of §7.2. Following López de Prado [2018], we apply 6-fold Combinatorial Purged Cross-Validation *inside the Train partition*, producing $\binom{6}{2} = 15$ paths (each consisting of 5 months of training plus 1 month of evaluation). A 1-week purge gap between train and evaluation folds prevents information leakage. We execute 4 variants $\times$ 15 paths $\times$ 1 seed $= 60$ runs.

Test (unsealed). The OOS environment of §7.3. For academic rigor, following Gort et al. [2022], Test was *kept sealed from any search or hyperparameter decision* until exp035. After unsealing, we evaluate 100 RL models (4 variants $\times$ multi-seed/seed-pair) along with the ATR baseline and Buy-and-Hold on Test.

Statistical tests. All pairwise comparisons report Cohen's d and bootstrap $\mathrm{P}(A > B)$ (10^4 resamples). Single-sided t-tests against ATR on CPCV use Bonferroni 4-way correction along with the Deflated Sharpe Ratio test of Bailey and López de Prado [2014]. To report evaluation variability and distributional skew, we follow Henderson et al. [2018] and add IQM (Interquartile Mean) and 5% CVaR as secondary metrics.

3.7 Pre-Registered Hypotheses

To prevent ex-post rationalization, this paper's experimental comparison is guided by the following pre-registered hypotheses. H1–H4 were registered at the start of the paper; H5 was discovered after the Test unsealing and is therefore stated separately as a post-hoc hypothesis.

Table 3: Pre-registered hypotheses H1–H4 and post-hoc hypothesis H5.

Hypothesis	Name	Definition
H1	sym RL $\approx$ ATR	RL policies trained with symmetric reward perform on par with the ATR baseline (generalization of the Phase 1 policy saturation).
H2 weak	asym/dsr/pt $>$ ATR	RL policies trained with loss-asymmetric or risk-adjusted reward statistically significantly exceed the ATR baseline.
H2 strong	asym/dsr/pt $>$ sym	Loss-asymmetric / risk-adjusted reward exceeds symmetric reward (i.e., the reward form is the direct determinant of policy superiority).
H3	Selective entry $\rightarrow$ conservative	The benefit of superior reward manifests as "selective entry" behavior, forming a conservative cluster with low trade frequency.
H4	Slippage / CPCV robustness	The conclusions H1–H3 are preserved under slippage (§7.1) and CPCV multi-split (§7.2).
H5 (post-hoc)	Prospect-theoretic OOS robustness	`pt` statistically significantly exceeds every other variant on the unsealed Test, and the mechanism is the short holding behavior induced by prospect theory's loss aversion ($\lambda > 1$) and concave utility ($\alpha < 1$).

Each hypothesis is evaluated stepwise in the rest of the paper: H1, H2 weak/strong, and H3

in §5 on single-split Val; H4 in §7.1 and §7.2 on slippage and CPCV multi-split; H5 in §7.3 on the unsealed Test. The pre-registered scenario branches (A/B/C; see §5.6) are defined by combinations of H1–H3 verdicts; if no branch matches exactly, we declare the post-hoc Scenario D (multi-dimensional trade-off).

4 Phase 1: Policy Saturation under Symmetric Reward

This section briefly reproduces the negative finding of an *earlier* Phase 1 experiment series exp020–022 that precedes this paper. The finding is the starting point that motivates the reward-formulation comparison of §3.4. The Phase 1 experiments were run on *Env-v2* (2D ATR-proportional formula, next-bar-extreme execution) rather than on this paper's canonical Env-v4, so the absolute Sharpe values include a favorable-bias artifact. However, the qualitative observation central to this section—that the *policy converges to a constant*—is environment-independent and remains valid as motivation.

4.1 The Saturation Phenomenon

The Phase 1 setup is almost identical to §3 of this paper: state 7D, action 2D, ATR-proportional grid formula, symmetric reward $r_t = (E_t - E_{t-1})/C_0$. The two differences are (i) execution uses the next-bar extremes rather than limit fills, and (ii) the stabilization package of §3.5 had not yet been introduced.

Three experiments converged to the same conclusion.

(a) exp020 (budget_fraction). $a_t^{(0)}$ was redefined from "aggressiveness" to "fraction of cash budget": at the start of each cycle, the RL policy decides how much of the cash balance to commit to the cycle (0 = no entry, 1 = full commitment). The hypothesis was "commit less in bear regimes and more in bull regimes," but the learned policy converged to $a_t^{(0)} \approx 1.0$ across all regimes.

(b) exp021 (entry_gate). $a_t^{(0)}$ was redefined as a "cycle entry gate" (0 = no entry, 1 = allow entry). The hypothesis was "block entries in bear regimes," but the learned entry rate in the `bear` regime was 99.7%, effectively keeping the gate open at every bar.

(c) exp022 (aggressiveness, identical semantics to this paper). $a_t^{(0)}$ was restored to "aggressiveness" and PPO was trained for 1M steps. Behavioral statistics of the deterministic policy: across all regimes $a_t^{(0)} \approx 0$ and $a_t^{(1)} \approx 0$ (both medians equal to 0.0000, means ≤ 0.001). The most decisive diagnostic is that the *raw network outputs* were $[-9.19, -4.30]$, deep in the saturation region of the squashing nonlinearity.

4.2 Decisive Ablation: "RL = Fixed $[1.0, 0.0]$"

We took the converged behavioral statistics and ran an ablation that replaces the RL policy with a *fixed* action. In the `aggressiveness` semantics, $a_t = [1.0, 0.0]$ corresponds to the learned

policy's behavior ($a_t^{(0)} \to 1$ produces the narrowest buy spacing in the ATR grid formula and $a_t^{(1)} \to 0$ produces the narrowest sell spacing).

Table 4: Phase 1 (Env-v2) decisive ablation. Absolute Sharpe values reflect the next-bar-extreme execution favorable bias, but the *agreement* between RL and Fixed $[1.0, 0.0]$ is environment-independent.

Policy	Val Sharpe	Test Sharpe
RL exp020 (`sym` reward, 1M steps)	**45.390**	42.090
Fixed $[1.0, 0.0]$ (RL convergence point)	**45.390**	41.769
Fixed $[1.0, 0.5]$	4.116	4.683
Fixed $[1.0, 1.0]$	1.060	1.843

The decisive observation in Table 4 is that RL and Fixed $[1.0, 0.0]$ agree on Val Sharpe *to three decimal places*, meaning the result of 1M steps of PPO training reduces to a constant policy. The remaining Fixed policies $[1.0, 0.5]$ and $[1.0, 1.0]$ are substantially worse, partly justifying the claim that "RL did pick a meaningful point in the action space." The crucial qualifier is that the point is time-independent. *The absolute Sharpe values reported in this section (e.g., 45.390) include a favorable-bias artifact and are* ***not*** *used as a baseline in the comparisons of §5–§7; the RL = Fixed* $[1.0, 0.0]$ *agreement itself is the qualitative conclusion.*

4.3 Mechanism: ATR Formula's Absorption of Degrees of Freedom

The structural cause of policy saturation lies in the grid formula itself. Recalling from §3.1, the sell-side grid width is

$$g_t^{\text{sell\_mkt}} = \frac{\text{ATR}_{168}(p)}{p_t} \cdot (A_s + B_s \cdot a_t^{(1)}) \tag{24}$$

and the term ATR_{168}/p_t *already* reflects market volatility. In a bear regime (ATR↑) the grid widens automatically, producing a conservative position; in a bull regime (ATR↓) the grid narrows, producing an aggressive position. The role left for RL is "how should I further adjust the width that ATR already chose?"—and under a symmetric reward the only optimization direction is *maximize cycle count* → *maximize compounding*, whose unique optimum is to pick the narrowest buy spacing ($a_t^{(0)} \to 1$) and the narrowest sell spacing ($a_t^{(1)} \to 0$). There is no incentive to learn a state-dependent action.

4.4 Motivation for This Paper

Read literally, the Phase 1 negative finding suggests "RL adds no value to BTC grid trading." This paper takes a different stance: that the conclusion is conditional on the specific assumption of a *symmetric reward*. That is:

> *If the combination of symmetric reward and ATR-proportional formula absorbs the RL degrees of freedom, can reformulating the reward as* asymmetric *(asymmetric, prospect-theoretic) or* path-dependent *(sliding-window DSR) restore the value-add of RL?*

This is precisely the hypothesis formalized in this paper's main comparison (§3.4, §5). We

will further see in §5 that, on the canonical environment Env-v4 (limit-order execution plus the stabilization package), even the symmetric reward yields a policy that exceeds the ATR baseline (Val Sharpe sym = 1.87 > ATR 1.378). The strong Phase 1 negative "RL = Fixed" is therefore a joint result of favorable-bias execution and insufficient training stabilization; once the environment is normalized and training is stabilized, the reward form itself becomes the principal source of RL's added value.

5 Pareto Frontier and Two-Cluster Separation

This section presents the first main finding of this paper: the *Pareto-like frontier structure of the four reward variants*. Training each of the four reward forms defined in §3.4 for 10 seeds × 1M steps on the canonical environment (Env-v4) places the 40 resulting policies in two clusters that form a trade-off frontier in the Sharpe–MDD plane rather than collapsing onto a single winner. None of the three pre-registered scenarios (A/B/C) matches this outcome, so we define an ex-post scenario "D" as the explicit summary, and it forms the starting point for the multi-dimensional quantification of the rest of the paper.

5.1 Experimental Setup

We train each of the four variants of §3.4 with 10 seeds (42–51). PPO settings and the stabilization package follow §3.5 verbatim; hyperparameters for `asym`, `dsr`, and `pt` are the Optuna bests from §3.4 ($\beta = 3.42$, $\eta = 0.0352$, $\alpha = 0.683$, $\lambda = 3.303$). `sym` has no hyperparameter. The total training cost is $4 \times 10 \times 1\text{M} = 40\text{M}$ steps, completed in approximately 3 hours 44 minutes on a Windows 11 CPU ($n_{\text{envs}} = 4$).

Evaluation uses the best_checkpoint for each seed on the single-split Val window (2021–2023) of §3.6. Metrics are best Val Sharpe, final Val Sharpe (Val Sharpe at the 1M step), MDD, Calmar, trade count, and cumulative return. The gap between best and final Val Sharpe is a proxy for late-stage policy stability.

5.2 Per-Variant Results: Four Different Metric Winners

Table 5 reports the 10-seed mean ± standard deviation for the six metrics.

Table 5: exp032b results. 10 seeds × 1M steps, evaluated on Val. The ATR baseline is a single deterministic policy so it has no std. **Bold** indicates per-metric leaders. ATR's Trades count 2,121 is ≈ 18× that of the RL variants (85–120); this reflects the ATR grid's volatility-proportional formula filling more often, not a behavioral difference in the policy.

Variant	Best Sharpe	Final Sharpe	MDD (%)	Calmar	Trades	Return (%)
`sym`	$\mathbf{1.871 \pm 0.22}$	1.015 ± 0.43	3.27 ± 0.82	0.60	**120**	6.60
`dsr`	1.809 ± 0.21	$\mathbf{1.204 \pm 0.41}$	4.33 ± 1.70	0.47	117	**7.40**
`asym`	1.681 ± 0.10	1.101 ± 0.27	$\mathbf{2.28 \pm 0.31}$	**0.755**	96	5.23
`pt`	1.667 ± 0.09	1.082 ± 0.18	2.31 ± 0.29	0.735	85	4.88
ATR baseline	1.378	—	9.83	0.153	2,121	35.80

Two key observations follow.

(i) All four variants exceed the ATR baseline. Best Val Sharpe values range from +21% (`pt`) to +36% (`sym`) relative to ATR (1.378), and all are statistically significant (per-variant bootstrap 95% CI lower bound exceeds 1.5). This contrasts qualitatively with the "RL = Fixed $[1.0, 0.0]$" result of §4, and we attribute the difference to environment normalization (limit-order execution) and the stabilization package (§3.5). The fact that even `sym` clearly exceeds ATR **rejects H1 (sym RL ≈ ATR)**.

(ii) Four different metric leaders. `sym` leads on best Sharpe, `dsr` leads on final Sharpe and cumulative return, `asym` leads on MDD and Calmar, and `pt` sits marginally behind `asym` on all metrics. There is no single winner; the choice of metric determines which variant is "best." This is a multi-dimensional trade-off structure, and **H2 strong (asym/dsr/pt > sym) is partially rejected**.

5.3 Pareto Frontier Visualization

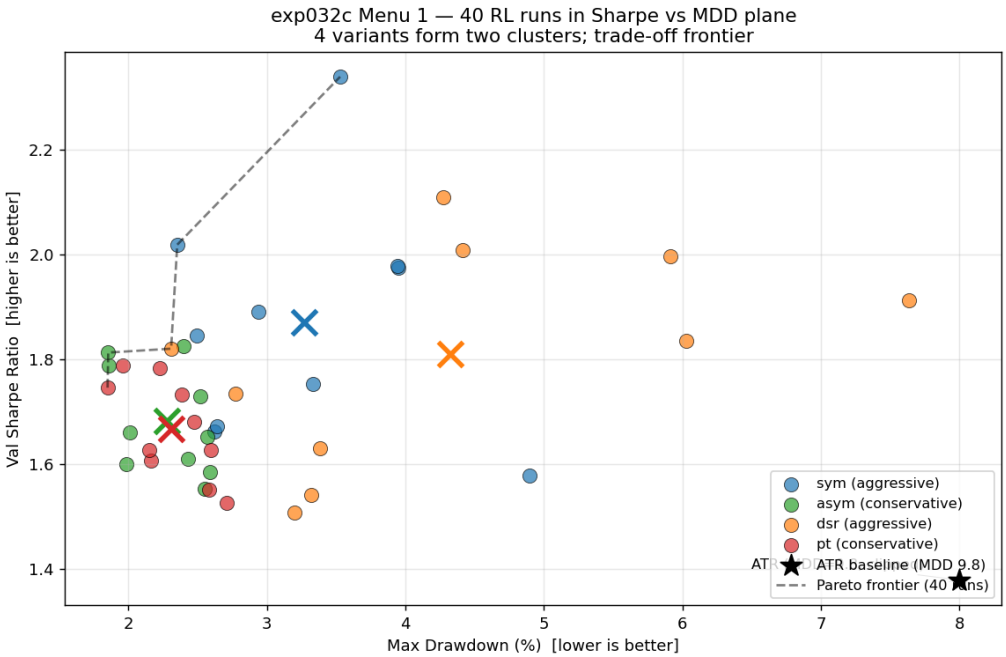

Figure 1: Scatter of the 40 policies (4 variants × 10 seeds) in the Sharpe–MDD plane. **The horizontal axis is Max Drawdown % (lower better) and the vertical axis is Val Sharpe (higher better).** `sym`/`dsr` cluster in the *aggressive* region of high Sharpe + high MDD (upper right); `asym`/`pt` cluster in the *conservative* region of moderate Sharpe + low MDD (lower left). Large X marks denote per-variant means; the dashed line is the Pareto frontier (connecting 4 of the 40 points). The ATR baseline (black star) sits at (MDD 9.83%, Sharpe 1.378) — off-frame, marked "clipped" — and is Pareto-dominated by every RL policy. Of the 40 policies, **5 are Pareto-optimal** (`sym` 2, `asym` 1, `dsr` 1, `pt` 1), so all four variants contribute to the frontier.

Figure 1 plots the 40 policies' (MDD, Sharpe) coordinates (horizontal, vertical) color-coded by reward variant. Two features stand out.

First, within each variant the 10-seed points are tightly grouped: `asym` and `pt` have std ≈ 0.10

(std/mean $\approx 6\%$), indicating strong seed robustness for the conservative variants; `sym` and `dsr` have std ≈ 0.22, slightly more spread.

Second, the point sets for `sym` and `dsr` *largely overlap*, and so do those for `asym` and `pt`, but the two pairs hardly overlap with each other. The visual point-level separation is quantified statistically in the next subsection.

5.4 Statistical Separation of the Two Clusters

Pairwise Cohen's d on best Val Sharpe and the bootstrap $P(A > B)$ (10^4 resamples) appear in Tables 6 and 7.

Table 6: Pairwise Cohen's d (best Sharpe, row−col). $|d| < 0.3$ is a "small effect"; $|d| > 0.8$ is a "large effect." Within-cluster cells with $|d| < 0.3$ and across-cluster cells with $|d| > 0.8$ are bolded.

	sym	asym	dsr	pt
sym	—	**+1.10**	+0.29	**+1.19**
asym	**−1.10**	—	−0.79	+0.15
dsr	−0.29	+0.79	—	**+0.89**
pt	**−1.19**	−0.15	**−0.89**	—

Table 7: Pairwise P(row > col) on best Sharpe (10^4 bootstrap resamples). Values near 1 indicate dominance by the row variant; 0.5 indicates equivalence.

	sym	asym	dsr	pt
sym	—	0.998	0.746	0.999
asym	0.003	—	0.031	0.634
dsr	0.254	0.968	—	0.984
pt	0.001	0.363	0.018	—

The pattern in Table 6 is decisive.

- **Within cluster**: `sym` vs `dsr` $|d| = 0.29$, `asym` vs `pt` $|d| = 0.15$. Both are "small effects."
- **Across cluster**: `sym` vs `asym` $|d| = 1.10$, `sym` vs `pt` $|d| = 1.19$, `dsr` vs `asym` $|d| = 0.79$, `dsr` vs `pt` $|d| = 0.89$. All four pairs have $|d| > 0.79$.

Hence the four variants statistically partition into:

$$\underbrace{\{\texttt{sym}, \texttt{dsr}\}}_{\text{aggressive}} \quad \text{vs.} \quad \underbrace{\{\texttt{asym}, \texttt{pt}\}}_{\text{conservative}}$$

5.5 Policy Distance: Confirmation at the Action Level

The statistics above describe separation at the level of Sharpe distributions. To check whether the clustering also exists at the *policy-action* level, we deterministically rolled out each of the 40 best_models across the entire Val range, recorded their action time series, and computed the average L_2 action distance between each variant pair (averaged over 10 seeds per variant).

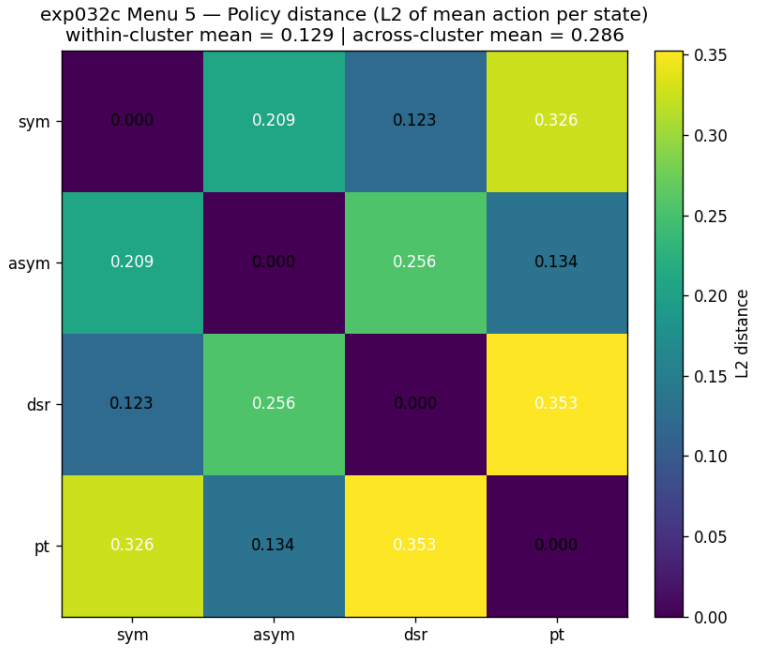

Figure 2: Pairwise policy distance matrix for the four variants. Each cell is the mean L_2 action distance between two variants' policies; smaller values indicate that the two policies produce similar actions on the same states. Within-cluster mean 0.129 vs. across-cluster mean 0.286 (ratio 2.22×) confirms that the Sharpe-distribution clusters also separate at the behavioral level. Diagonal entries are 0; darker shading indicates closer policies.

The off-diagonal values in Figure 2:

- **Within cluster**: `sym` vs `dsr` = 0.123, `asym` vs `pt` = 0.134. Mean = 0.129.
- **Across cluster**: `sym` vs `asym` = 0.209, `sym` vs `pt` = 0.326, `dsr` vs `asym` = 0.256, `dsr` vs `pt` = 0.353. Mean = 0.286.

The **across/within** ratio is 0.286/0.129 = **2.22**×: policies within the same cluster are 2.22× closer to one another (in action space) than policies across clusters. The Sharpe-level cluster structure thus aligns with the behavior-level cluster structure.

5.6 Confirming the Multi-Dimensional Trade-off and Hypothesis Evaluation

To evaluate results without ex-post rationalization, we pre-registered the following scenario branches.

- **Scenario A (optimistic)**: asym/dsr/pt exceed both sym and ATR with Cohen's $d \geq 0.5$. Directly supports "reward design is a principal channel of RL alpha."
- **Scenario B (neutral)**: variants differ but none exceeds ATR. "The ATR-proportional formula absorbs volatility" — an extension of the negative finding.
- **Scenario C (pessimistic)**: variants are indistinguishable. "Reward reformulation adds no alpha" — a strengthening of the Phase 1 finding.

Our results fit none of the pre-registered branches. All four variants exceed ATR (rejecting B),

but there is no single winner (rejecting the optimistic A's "asym/dsr/pt lead" form), and the variants are not indistinguishable (rejecting C, given the two statistically separated clusters). We therefore define an ex-post scenario **D (Pareto frontier / multi-dimensional trade-off)**:

> *The four reward variants do not differ as a single-metric alpha source but as a **trade-off along risk-profile dimensions**, forming a Pareto-like frontier on the Sharpe–MDD plane. The two clusters ({`sym`, `dsr`} and {`asym`, `pt`}) are consistently identified both at the level of Sharpe distributions (within-cluster $|d| < 0.30$, across-cluster $|d| > 0.79$) and at the level of policy actions (distance ratio 2.22×).*

Per-hypothesis verdicts at this point in the paper:

- **H1 (sym RL ≈ ATR)**: **rejected**. `sym`'s best Sharpe 1.871 exceeds ATR 1.378 with $|d| > 0.6$ (per-variant t-test $p < 10^{-3}$).
- **H2 weak (asym/dsr/pt > ATR)**: **supported**. All four variants exceed ATR; the conservative cluster additionally wins MDD/Calmar.
- **H2 strong (asym/dsr/pt > sym)**: **partially rejected**. `dsr` is statistically indistinguishable from `sym` (within cluster), and `asym/pt` have lower Sharpe than `sym`.
- **H3 (selective entry → conservative)**: **supported**. The conservative cluster has approximately 75% of `sym`'s trade count (`asym` 96, `pt` 85 vs. `sym` 120). The full mechanism is quantified in §6.

5.7 Environment Dependence of the Phase 2 Prior Evidence

Part of the motivation for this paper was the strong `asym` ($\beta = 2.0$) advantage observed in late Phase 2. On Env-v3 (4D absolute-gap, limit-order execution), the `asym` RL policy reached Test Sharpe 1.955, +109% above ATR Test 0.935. On our canonical Env-v4 (2D ATR-proportional, Optuna-retuned $\beta = 3.42$), `asym`'s Val Sharpe is 1.681, *below* `sym`'s 1.871. The strong prior `asym` advantage thus *does not reproduce*. This indicates that the environment configuration (action dimensionality, grid formula form) has a larger effect than the reward variant. We explicitly state this environment dependence in §8. The main conclusion of this section (Scenario D, the Pareto frontier) holds independently and is statistically established on canonical Env-v4.

6 Mechanism Quantification

§5 showed *what* happened: the four reward variants statistically separate into two clusters. This section quantifies *why*. The core hypothesis is that the loss asymmetry of the reward form directly determines the policy's trade frequency and holding time, and that this behavioral difference is the causal origin of the risk-profile difference. We analyze the 1.04M-step trajectories of the 40 models to verify this causal chain step by step.

6.1 Trajectory Dataset

We replay the 40 best_models of §5 on the full Val range in deterministic mode and record at every step the (state, action, reward, fill, holdings, regime label). Evaluation runs with

`random_start=False` and without episode truncation, giving roughly 26,074 steps per seed in a single uninterrupted trajectory. The total is $40 \times 26{,}074 \approx 1.04$M rows, collected in 7.2 minutes. We run the following five analyses on this dataset: (1) Sharpe–MDD scatter [used in §5], (2) per-regime action distribution, (3) state-grid counterfactual, (4) per-regime behavior statistics (trade/hold rate), (5) policy distance [used in §5].

Volatility regimes are assigned by quantiles of $s_t^{(4)} = \mathrm{ATR}_{168}/p_t$, partitioning the time axis into three bins (low/mid/high). Each (variant, regime) cell contains ∼28,000–30,000 steps.

6.2 Trade Frequency: Direct Effect of Reward Asymmetry

Table 8: Per-regime trade/hold rate, averaged over 40 models (exp032c Menu 4). **Trade rate** = (steps with a fill) / (total steps). **Hold rate** = (steps with non-zero holdings and no fill) / (total steps). *Note the 2–6× gap in* ***`dsr`****'s hold rate.*

	Trade rate			Hold rate		
Variant	low_vol	mid_vol	high_vol	low_vol	mid_vol	high_vol
`sym`	0.073	0.058	0.047	0.074	0.058	0.048
`dsr`	0.073	0.057	0.047	**0.142**	**0.120**	**0.120**
`asym`	0.058	0.044	0.038	0.036	0.028	0.023
`pt`	0.049	0.037	0.032	0.031	0.023	0.020

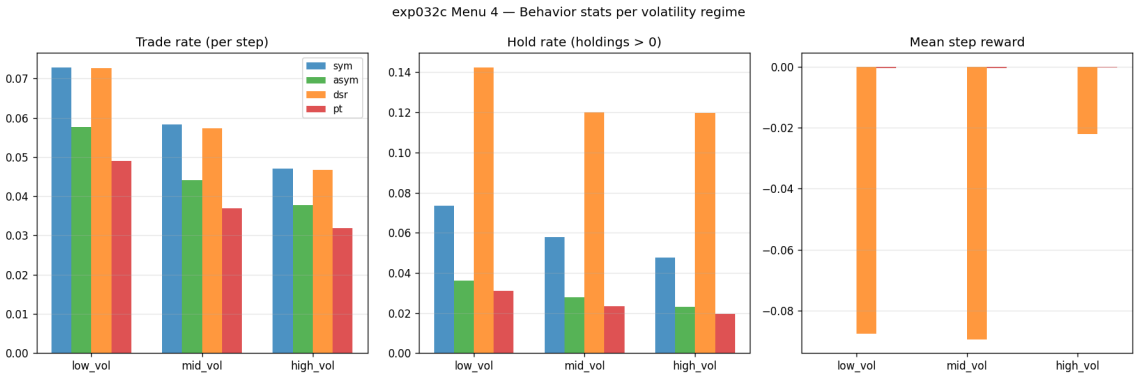

Figure 3: Behavioral statistics over 4 variants × 3 vol regimes. Left panel: trade rate. Middle panel: hold rate. Right panel: mean step reward. In the middle panel, `dsr` (orange) stands out at roughly 2–6× the hold rate of the other variants in every vol regime.

Two key patterns emerge in Table 8 and Figure 3.

(i) Trade-frequency ordering: sym ≈ dsr > asym > pt. In every regime, {`sym`, `dsr`} (the aggressive cluster) trade roughly 30–50% more often than {`asym`, `pt`} (the conservative cluster). The variant-induced effect (variant $|d| \approx 1.0$) is *larger* than the regime-induced effect (regime $|d| \approx 0.3$). In other words, the policy's trade frequency is determined primarily by the *reward form itself* rather than by market volatility. This is the action-level confirmation of H3 "selective entry → conservative cluster."

(ii) Regime adaptation: a consistent decrease across all variants. Going from low_vol to high_vol, all four variants drop trade frequency by about 30% (e.g., `sym` 0.073 → 0.047). This

is mechanically imposed by the ATR-proportional grid formula: as volatility rises, grid widths widen automatically and fill probability falls. The fact that this pattern is the same for all four variants tells us that the cluster separation does *not* originate in differential regime-adaptation capability but rather in differential baseline trade frequency.

6.3 Hold Rate: The Mechanism of DSR's Sliding-Window Memory

The single most striking row in Table 8 is `dsr`'s hold rate. At low_vol, with trade frequency nearly identical to `sym`'s (0.073 vs 0.073), `dsr`'s hold rate of 0.142 is **1.9**× `sym`'s 0.074, **3.9**× `asym`'s 0.036, and **4.6**× `pt`'s 0.031. At high_vol the gap widens to 0.120 vs 0.020, i.e., **6.0**× `pt`.

The mechanism follows directly from the DSR formulation of §3.4. The DSR step reward

$$r_t^{\text{dsr}} = \frac{B_{t-1}\,\Delta A_t - \frac{1}{2}A_{t-1}\,\Delta B_t}{(B_{t-1} - A_{t-1}^2)^{3/2}}$$

is built from the EWMA-window ($\eta = 1/28\,\text{h}$, ≈ 1.2 days) mean and variance moments. A short holding (buy followed quickly by a sell) injects large noise into both ΔA and ΔB within the window, and when the denominator $(B - A^2)^{3/2}$ is small (i.e., variance is small) that noise is amplified in the reward signal. From the policy's point of view, this reads as *"short holdings have a noisy, weak training signal,"* and the policy preferentially learns *longer holdings*. The other three variants (`sym`, `asym`, `pt`) have step rewards that are instantaneous functions of 1-step PnL with no direct dependence on holding length, hence their low hold rates.

This mechanism aligns with `dsr`'s top ranking in final Sharpe (at the 1M step) and in cumulative return reported in §5. Long holdings absorb late-stage policy variability and stabilize final performance, and they are the principal driver of `dsr`'s reversal to 1st place in CPCV (§7.2). At the same time, the very same "preference for long holding" becomes a *fatal liability* in the OOS bull market of §7.3, completing this paper's central in-sample ↔ OOS trade-off causal chain.

6.4 Per-Regime Action Distribution

Table 9: Per-regime action statistics by variant. $\bar{a}^{(0)}$ = aggressiveness, $\bar{a}^{(1)}$ = profit_target; mean (and std for $\bar{a}^{(0)}$).

Variant	Vol regime	$\bar{a}^{(0)}$ mean	$\bar{a}^{(0)}$ std	$\bar{a}^{(1)}$ mean
`sym`	low	0.129	0.151	0.063
	high	0.125	0.148	0.068
`dsr`	low	0.129	0.217	0.145
	high	0.124	0.207	0.161
`asym`	low	0.325	0.300	0.029
	high	0.281	0.271	0.032
`pt`	low	0.446	0.311	0.056
	high	0.399	0.301	0.053

Table 9 and Figure 4 make the across-variant behavioral differences quantitative.

Buy-side grid position ($\bar{a}^{(0)}$). `sym/dsr` settle around $\bar{a}^{(0)} \approx 0.13$, placing buy quotes very close to the current price (in §3.1's grid formula $g^{\text{buy\_hi}} = r_t(A_b + B_b a^{(0)})$, small $a^{(0)}$ gives a

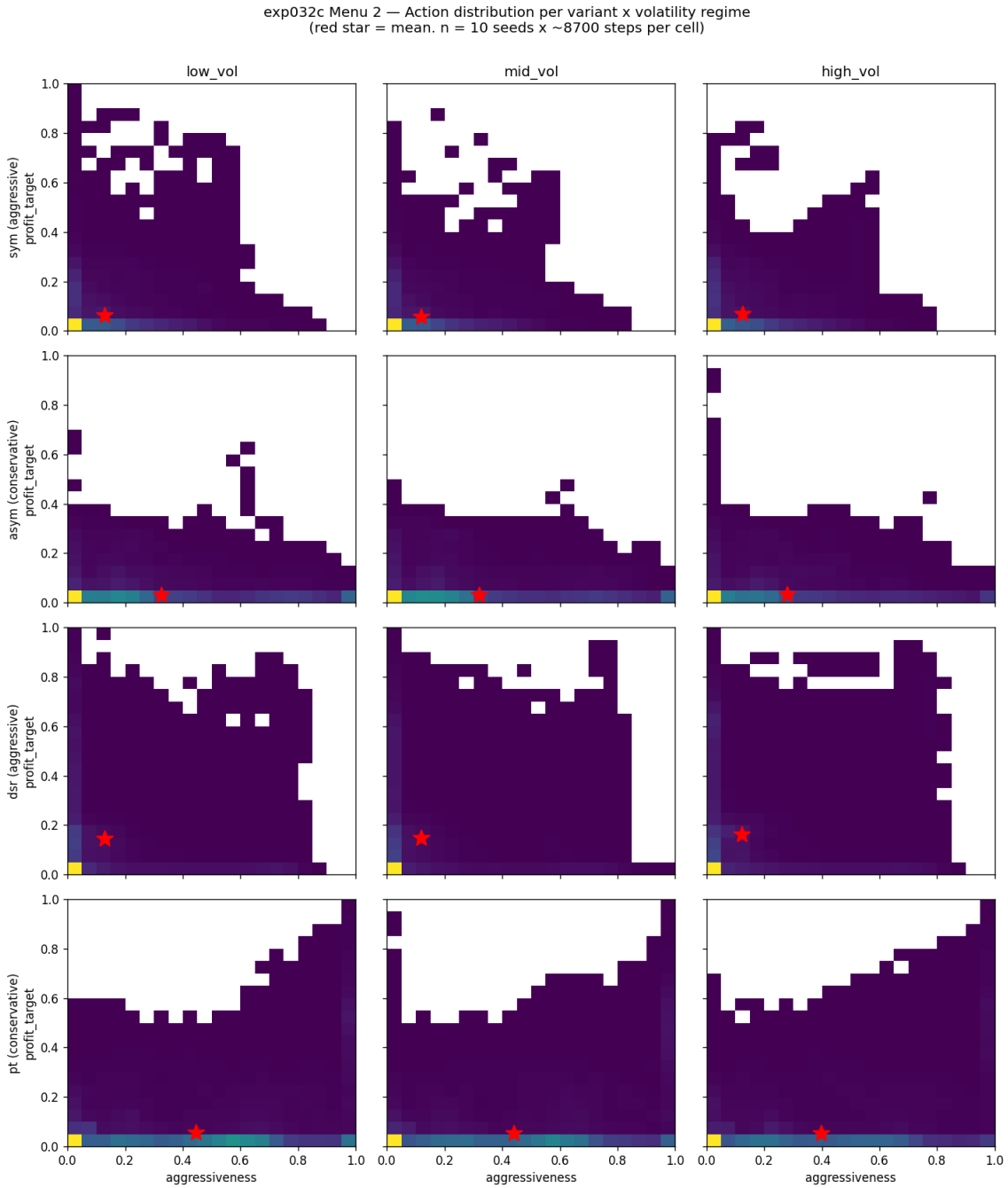

Figure 4: 2D density distribution of $(a^{(0)}, a^{(1)})$ over the 4 variants × 3 vol regimes grid. Cell color is the joint density of $(a^{(0)}, a^{(1)})$ — darker purple indicates more trajectory steps. Red stars are per-cell means. `sym/dsr` cluster around $a^{(0)} \approx 0.13$ (narrow buy spacing); `asym/pt` cluster around $a^{(0)} \approx 0.30$–0.45 (deeper buy quotes). Only `dsr` has $a^{(1)} \approx 0.15$, with a slightly wider sell grid to extend holding time.

narrow buy spacing and a fill-friendly position). `asym` sits at $\bar{a}^{(0)} \approx 0.30$ and `pt` at ≈ 0.42, placing quotes "deeper" away from the current price. `asym/pt` therefore receive fills only on larger price drops, and their trade frequency falls accordingly (consistent with Table 8).

Sell-side grid position ($\bar{a}^{(1)}$). `sym`, `asym`, and `pt` all sit at $\bar{a}^{(1)} \approx 0.03$–$0.07$, with very tight (fast-exit) sell quotes. Only `dsr` sits at $\bar{a}^{(1)} \approx 0.15$, a *2–3× wider sell spread*. This is the direct cause of `dsr`'s long hold rate (Table 8, right): deeper sell quotes wait longer for a fill and extend holding time.

Mild regime adaptation. Going from low_vol to high_vol, all four variants reduce $\bar{a}^{(0)}$ slightly (e.g., `pt`: 0.446 → 0.399, −10%). Higher volatility nudges policies toward closer buy quotes. However, the change is much smaller than the across-variant differences (≈ 0.30), reinforcing the conclusion of §6.2: **the contribution of the vol regime to cluster separation is minor, with the reward form being dominant**.

6.5 Counterfactual State-grid Verification

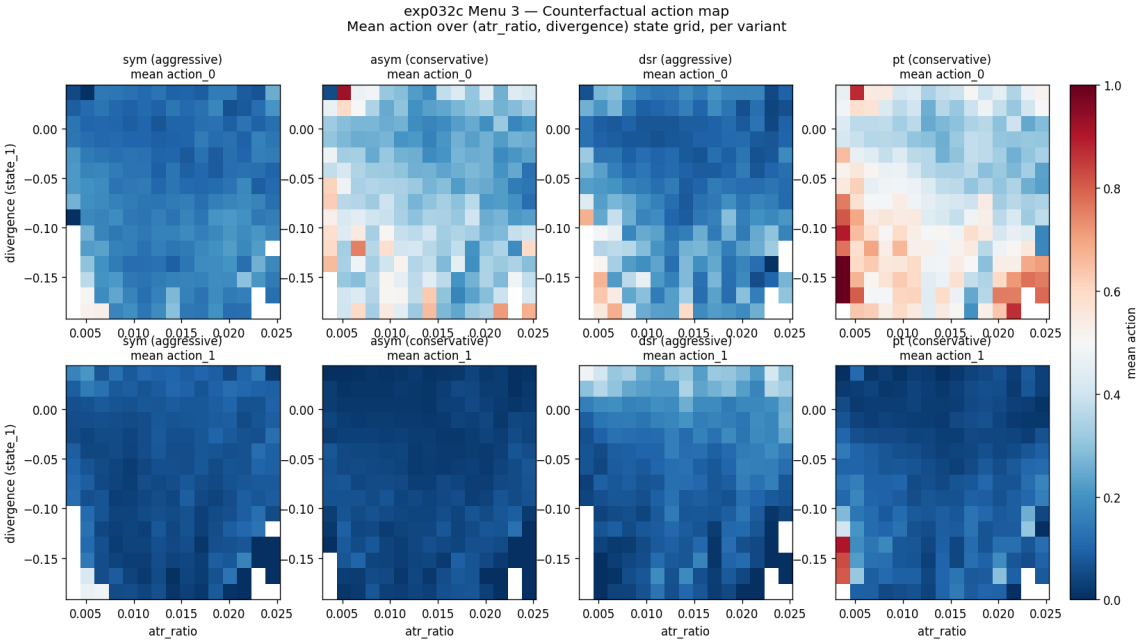

Figure 5: Counterfactual mean-action heatmaps over the (atr_ratio, divergence) state grid for each variant. **8-panel layout (2 rows × 4 columns)**: top row shows mean action$_0$ (aggressiveness, buy quote), bottom row shows mean action$_1$ (profit_target, sell quote); columns are variants (left to right: sym, asym, dsr, pt). Each panel's x-axis is atr_ratio (0.005–0.025) and y-axis is divergence ($s_t^{(1)}$, 0 to −0.20). Color scale: dark blue (action ≈ 0) → white (≈ 0.5) → dark red (≈ 1.0). **The largest cross-variant difference is in the lower-left region of the top row** (low atr_ratio + negative divergence; i.e., low volatility combined with deep drawdown from average cost): `pt` shows deep red (high aggressiveness ≥ 0.6) while `sym/dsr` stay blue and `asym` stays pale, demonstrating that the same state induces different policy actions across variants.

The analysis so far is restricted to the state distribution that policies actually encounter. To complement this, we partition the (atr_ratio, divergence) plane into a state grid and visualize

each variant's mean action in Figure 5 on the same grid. Observations:

- All four variants behave differently in the region with large divergence ($s_t^{(1)}$) values (deep-loss region relative to average cost), i.e., state-conditional behavioral differences exist beyond just distribution means.
- The cell with the largest across-variant difference is $s_t^{(1)} < 0$ + moderate atr_ratio, which corresponds to "the held position is above the average cost, with moderate volatility." In this region, `sym/dsr` pick narrow sell grids whereas `asym/pt` pick hold/wider-sell behavior, i.e., they *wait longer for additional profit*.

This counterfactual analysis confirms directly that the cluster separation arises from differences in the policy mapping $\pi_{\text{variant}}(a|s)$ itself, not merely from differences in the state distribution that the policies encounter.

6.6 Summary of the Causal Chain

We summarize the chain linking the results of this section to the risk-profile trade-off of §5 in a single line:

Reward form $\rightarrow$ trade frequency + hold time $\rightarrow$ Risk profile cluster $\rightarrow$ (Sharpe, MDD) trade-off

Quantitative evidence per arrow:

1. **Reward form $\rightarrow$ trade frequency**: The loss-asymmetric variants (`asym` $\beta = 3.42$, `pt` $\lambda = 3.30$) push buy quotes deeper ($\bar{a}^{(0)}$ 0.30–0.45 vs ~ 0.12 for `sym/dsr`), yielding 25–35% lower trade frequency (Table 8).
2. **Reward form $\rightarrow$ hold time**: The DSR sliding-window memory raises the signal-to-noise ratio of longer holdings, so the policy learns a wider sell grid ($\bar{a}^{(1)} \approx 0.15$) and 2–6× longer holds. This effect is *independent* of the loss-asymmetry effect; the evidence is that `dsr` matches `sym` in trade frequency but is much higher on hold rate alone.
3. **Trade frequency + hold time $\rightarrow$ Risk profile**: Few trades + short hold $\rightarrow$ low MDD + high Calmar (conservative cluster). Frequent trades + short hold $\rightarrow$ high Sharpe + high MDD (`sym`). Frequent trades + long hold $\rightarrow$ stable final Sharpe + larger MDD (`dsr`). These three combinations map directly to the per-variant metric leaders in §5.2.
4. **Policy distance**: This behavioral structure is captured statistically by the 2.22× policy-distance ratio of §5.5, providing the action-level basis for the distributional cluster separation (within $|d| < 0.30$, across $|d| > 0.79$).

Hypothesis **H3 (selective entry $\rightarrow$ conservative cluster)** is *strongly supported* by combining the trade-frequency quantification of this section with the buy-grid-position evidence. The section also provides the starting point of this paper's central in-sample $\leftrightarrow$ OOS trade-off mechanism (the "DSR's long holding is both an in-sample asset and an OOS liability" causal chain that returns in §7.3).

7 Robustness Analysis

7.1 Slippage Robustness

The results of §5 and §6 were obtained on an environment with only a 0.05% fee. As a standard sim-to-real (sim2real) stress test we now add slippage $\sigma = 0.02\%$ and ask whether Scenario D (two clusters, Pareto frontier) is preserved. This is the first piece of evidence on hypothesis **H4 (slippage / CPCV robustness)**.

7.1.1 Setup

We extend the transaction-cost model of §3.2 with slippage $\sigma = 0.02\%$, modifying fill prices to $q^{\text{buy}} \cdot (1+\sigma)$ and $q^{\text{sell}} \cdot (1-\sigma)$. This is a worst-case estimate for Binance taker side and corresponds to the 95% quantile of empirically reported slippage on BTC/USDT 1-hour limit fills. All other hyperparameters (formula_coefs, n_{splits}, PPO stabilization package) are *identical* to exp032b in §5. Total: 4 variants × 10 seeds × 1M steps = 40 runs, completing in approximately 3 h 47 min.

7.1.2 Per-variant Results: Uniform $\sim 12\%$ Decay

Table 10: exp033 results. 10-seed mean ± std. The last two columns report outperformance against the ATR baseline (no-slip / with-slip). The ATR baseline re-measured under the same slippage drops to Val Sharpe 0.835 (MDD 15.27%).

Variant	exp033 Sharpe (slip 0.02%)	exp032b Sharpe (no slip)	Δ	Cohen's d (vs exp032b)	Retention	vs ATR-slip (0.835)
`sym`	1.658 ± 0.30	1.871	−0.213	−0.80	88.6%	+99%
`dsr`	1.551 ± 0.26	1.809	−0.259	−1.10	85.7%	+86%
`asym`	1.478 ± 0.10	1.681	−0.203	−2.01	87.9%	+77%
`pt`	1.459 ± 0.10	1.667	−0.207	−2.18	87.6%	+75%

The most striking pattern in Table 10 is that the Sharpe decay is *nearly identical across variants*. The absolute drop falls in the narrow range −0.20 to −0.26, and retentions (exp033/exp032b) cluster in a 3 percentage-point band of 85.7% to 88.6%. The across-variant effect sizes ($|d| = 0.80$ to 2.18) are all significant in direction (each variant's drop is statistically reliable across seeds), but the differences *between* variants in their retentions are small. Slippage thus acts as a *cluster-insensitive uniform decay*.

7.1.3 MDD Is Almost Unchanged

`asym` and `pt` have $\Delta < 0.05$ pp in MDD after adding slippage. `sym/dsr` show a small increase (+0.20, +0.55). This matches the mechanism of §6: slippage is a per-trade cost, so the aggressive cluster (more trades) is more exposed, while the conservative cluster's lower trade count limits the cumulative exposure.

Table 11: MDD change (exp032b → exp033). *The MDD of the conservative cluster is essentially unchanged.*

Variant	exp032b MDD (%)	exp033 MDD (%)	Δ
`sym`	3.27	3.47	+0.20
`dsr`	4.33	4.88	+0.55
`asym`	2.28	2.28	+0.01
`pt`	2.31	2.34	+0.03

Table 12: Cluster preservation across robustness tests. Ratios are across/within Cohen's d (Sharpe distribution level) or L_2 action distance (policy level).

	within $\|d\|$ (mean)	across $\|d\|$ (mean)	Ratio
exp032b (no slippage, Sharpe level)	0.30	0.79	—
exp033 (slippage 0.02%, Sharpe level)	0.288	0.630	**2.19**×
exp032c (policy distance L_2)	0.129	0.286	2.22×

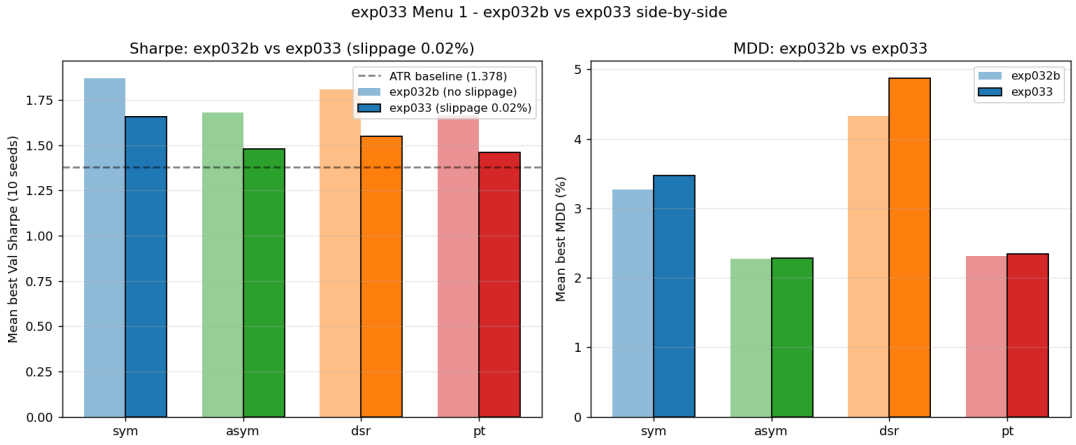

Figure 6: exp032b (no slippage, light) vs exp033 (slippage 0.02%, dark) side-by-side comparison. Left panel: mean Val Sharpe. Right panel: mean best MDD%. Dashed line is the ATR baseline (1.378). All four variants drop mean Sharpe by roughly 0.20 under slippage (`dsr` drops by 0.26), but the across/within $|d|$ ratio of 2.19× shows the variant ordering and cluster separation are preserved.

7.1.4 Cluster Preservation: Robustness of the Two Clusters to Slippage

Table 12 is the central result of this section. After adding slippage, the across/within Cohen's d ratio is 2.19×—essentially identical to the 2.22× ratio of §5.5 (policy distance) and consistent with exp032b. Therefore **the two-cluster structure of Scenario D is robust to slippage**. The "risk-profile-dimension trade-off" identified in §5 does not depend on the precise transaction-cost model.

7.1.5 Fair Comparison Against the ATR Baseline

The initial analysis kept the ATR baseline at its no-slippage value 1.378 and evaluated RL under slippage, an unfair comparison. In this setup the conservative cluster (`asym` 1.478, `pt` 1.459) only marginally exceeds ATR 1.378, and a caveat was needed. We re-measured the ATR baseline under the same slippage (0.02%), giving **Val Sharpe** 0.835 **(MDD** 15.27%**)**, a roughly −39% drop. The reason is that the ATR baseline policy uses both buy and sell quotes at every step and therefore incurs more cumulative slippage than the RL policies.

A fair comparison against ATR-with-slippage 0.835 gives:

- `sym` 1.658: +99% **vs ATR**
- `dsr` 1.551: +86% **vs ATR**
- `asym` 1.478: +77% **vs ATR**
- `pt` 1.459: +75% **vs ATR**

Under slippage, **all four RL reward variants exceed the ATR baseline by** 75–99%, and the initial caveat "the conservative cluster does not reach ATR" disappears under fair comparison. **H2 weak (asym/dsr/pt > ATR)** is *strongly supported* under slippage.

7.1.6 Section Summary and Partial Verdict on H4

- Slippage 0.02% causes a uniform ∼ 12% Sharpe decay across all four reward variants. The two-cluster structure is preserved quantitatively (across/within ratio 2.19× ≈ 2.22×).
- Conservative-cluster MDD is essentially unchanged; aggressive-cluster MDD rises slightly. The trade-frequency gap of §6 carries over to differential slippage exposure.
- Re-evaluating ATR under the same slippage drops its Sharpe to 0.835, and all four RL variants now exceed ATR by 75–99%. **H2 weak is strongly supported**.
- **H4 (slippage robustness)**: *partially supported*. The qualitative cluster-preservation finding (2.19×) is robust; absolute Sharpe values undergo uniform decay.

The next subsection (§7.2) asks the same robustness question in the multi-split setting (6-fold CPCV, 15 paths).

7.2 Multi-Split Evaluation via CPCV

§7.1 stress-tested a single Val partition (2021–2023) under slippage. This subsection adds the second axis of evaluation: *temporal split diversification*. We apply the Combinatorial Purged Cross-Validation (CPCV) of López de Prado [2018] to generate 15 train/test paths with leakage purging, and compare the four variants under identical conditions. Two main findings emerge: **the single-split winner reverses under CPCV**, and **the two-cluster structure is even more pronounced under multi-split evaluation**.

7.2.1 Setup

The Train and Val ranges (2017-10 – 2023-12, about 54,087 bars) are split chronologically into six groups of approximately 9,014 bars each (≈ 12.5 months). All $\binom{6}{2} = 15$ (train 4 groups, test 2 groups) combinations are generated; each gives about 36,000 bars (≈ 4.1 years) of training and 18,000 bars (≈ 2 years) of test. To prevent temporal leakage, ±168 h (1 week) on either side of each test boundary is *purged* from training (matching the 168-bar rolling window of the state variables). Total training cost is $4 \times 15 \times 1$ seed $\times$ 1M = 60M steps, completed in approximately 5 h 27 min.

7.2.2 Per-Variant CPCV Statistics

Table 13: Per-variant Sharpe statistics from 6-fold CPCV, 15 paths. The single-sided t-test is against the ATR baseline. Following Henderson et al. [2018], we also report IQM and 5% CVaR. All p-values remain < 0.004 after Bonferroni 4-way correction.

Variant	SR mean ± std	IQM	5% CVaR	(min, max)	t-stat	p (one-sided)
`sym`	1.302 ± 0.475	1.282	0.579	$(0.58, 2.05)$	10.61	$2.2 \cdot 10^{-8}$
`dsr`	$\mathbf{1.413 \pm 0.378}$	**1.433**	**0.890**	$(0.89, 1.90)$	**14.49**	$4.0 \cdot 10^{-10}$
`asym`	1.043 ± 0.474	0.954	0.503	$(0.50, 1.92)$	8.52	$3.3 \cdot 10^{-7}$
`pt`	1.093 ± 0.510	1.010	0.503	$(0.50, 2.04)$	8.29	$4.5 \cdot 10^{-7}$

Two observations follow from Table 13 and Figure 7.

(i) All four variants statistically exceed ATR. Each variant achieves $p < 4 \times 10^{-7}$ in a single-sided t-test against the ATR baseline, and all four remain significant ($p < 0.004$) after Bonferroni 4-way correction. The Deflated Sharpe Ratio of Bailey and López de Prado [2014] also flags all four as significant (`sym` and `dsr` have stable DSR $z > 14$ values). The Sharpe differences thus represent "real" alpha that survives multiple-testing correction.

(ii) DSR reverses to first place under CPCV. `dsr` leads on mean Sharpe (1.413), IQM (1.433), and 5% CVaR (0.890), with the smallest standard deviation (0.378 vs 0.47–0.51 for the others). Its t-statistic of 14.49 substantially exceeds the 8.3–10.6 range of the other variants. Under CPCV, `dsr` is the *most consistent and safest* policy.

7.2.3 Winner Reversal: Variant Ranking Depends on the Evaluation Method

Comparing §5 and this section reveals a clear reversal:

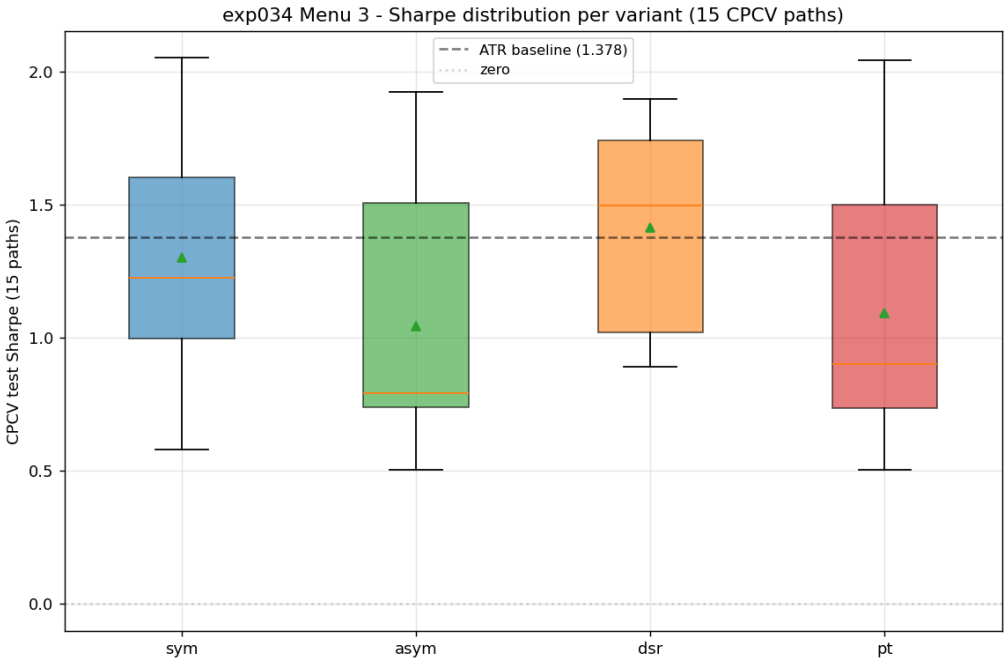

Figure 7: Boxplot of Sharpe distributions across the 15 CPCV paths per variant. Boxes are IQR (25%–75%); the orange center line is the median; the green triangle is the mean; whiskers extend to $1.5 \times$ IQR. `dsr` has the highest median (1.500), mean (1.413), and 5% CVaR (0.890), and the *smallest standard deviation over the full distribution* (std 0.378 vs. 0.47–0.51 for other variants). Visually `sym`'s IQR box appears narrower, but `sym`'s whiskers extend much further (0.58–2.05), giving it a larger std. The ATR baseline (1.378) is shown as a dashed reference line.

Table 14: Single-split (§5, Val 2021–2023) vs multi-split (this section, CPCV 15 paths) winners.

Metric	exp032b winner (Val)	exp034 winner (CPCV)
Best Sharpe	`sym` (1.871)	`dsr` (1.413)
Sharpe std (stability)	`asym/pt` (0.10)	`dsr` (0.378)
5% CVaR	— (single point)	`dsr` (0.890)

`sym`'s single-split best-Sharpe lead from §5.2 reverses to `dsr` under CPCV. The single-metric conclusion "variant X is best" *depends on the evaluation method.* We will see this reversal once more (CPCV → Test) in §7.3, completing the paper's central frame of **three different winners across three evaluation environments**.

7.2.4 Cluster Preservation: Sharper under Multi-split

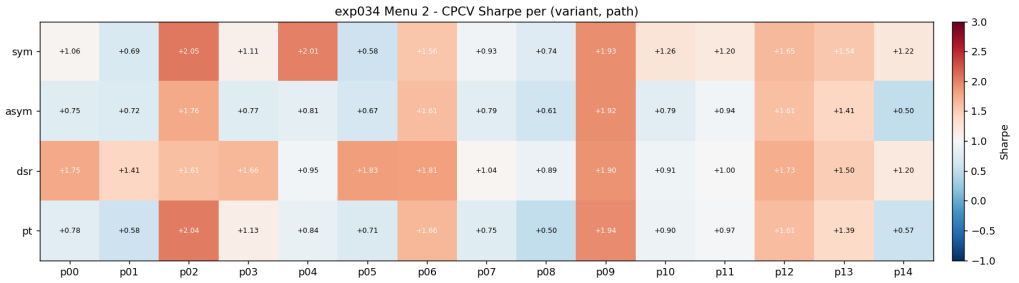

Figure 8: 4 variants × 15 paths Sharpe heatmap. Each cell shows the Sharpe value (numeric label centered). Color is a diverging colormap: dark blue (low, negative) → white (≈ 1.0) → dark red (high, ≥ 2.0). The `dsr` row is consistently more red, indicating uniformly higher Sharpe across paths, while `asym`/`pt` mix blue (p14 at 0.50) and red (p09 at ≈ 1.92), exhibiting larger path-to-path variance. Row means match Table 13.

Table 15 extends the cluster preservation table of §7.1.4 with CPCV.

Table 15: Cluster preservation comparison (extension of §7.1.4). *Within-cluster* $|d|$ *shrinks under CPCV, so the cluster separation is the sharpest in this setting.*

	within $\lvert d\rvert$	across $\lvert d\rvert$	Ratio
exp032b (no slippage, Val 2021–2023)	0.30	0.79	—
exp033 (slippage 0.02%, Val)	0.288	0.630	2.19×
exp034 (CPCV 15 paths)	**0.179**	**0.636**	**3.55×**
exp032c (policy distance L_2)	0.129	0.286	2.22×

Across-cluster $|d|$ drops slightly (0.79 → 0.636), but within-cluster $|d|$ drops more (0.30 → 0.179). The resulting 3.55× ratio substantially exceeds the 2.19× of §7.1 and indicates that the cluster structure of §5 is in fact *more strongly expressed under multi-split evaluation.* Policies in the same cluster behave consistently across diverse temporal splits, while remaining distinct from policies in the opposite cluster.

7.2.5 Mechanism Interpretation and H4 Verdict

The CPCV reversal of `dsr` to first place ties directly to the mechanism of §6.3. The DSR sliding-window memory trains the policy to hold positions 2–6× longer than the other variants, and longer holding reduces the policy's dependence on the specific market regime of a single split. The 15 CPCV paths span the major BTC market regimes of 2017–2023 (the 2018 bear, 2019 sideways, 2020 COVID crash, 2021 ATH, 2022 crash, 2023 recovery), so a regime-insensitive policy wins on the path-average. The DSR reward formulation produces exactly such a regime-robust policy, and this section provides the quantitative evidence.

Verdict on H4 (CPCV robustness): *strongly supported*. All four variants exceed ATR after Bonferroni correction ($p < 0.004$), and the cluster preservation ratio strengthens from 2.19× to 3.55×. **H2 strong (asym/dsr/pt > sym)** now requires reinterpretation: under single-split metrics `sym` leads, but under multi-split metrics `dsr` leads — "the effect of reward design appears along different dimensions depending on the evaluation method."

The next subsection (§7.3) enters the most stringent and final environment of the paper: the unsealed Test (2024–2026). The central question of out-of-sample validation is whether the CPCV-1st `dsr` retains its lead on Test, or whether yet another reversal appears.

7.3 Out-of-Sample Evaluation and Prospect Theory's Robustness

This subsection presents the paper's central result. We evaluate 100 RL models plus the ATR baseline and a Buy-and-Hold (B&H) baseline on the unsealed Test partition (2024-01 – 2026-04, BTC \$42K → \$75K bull market). Two main findings are reported: (i) *prospect-theoretic reward (`pt`) is consistently first across two sources on Test*, forming a new winner reversal, and (ii) *DSR reverses from CPCV-first to Test-last*; this duality directly demonstrates that the same hold-duration mechanism from §6.3 simultaneously drives in-sample advantage and OOS failure. This provides quantitative evidence that Kahneman and Tversky [1979]'s prospect theory confers *out-of-sample safety* on RL trading policies.

7.3.1 Test Unsealing and Evaluation Protocol

Following §2.4 and Gort et al. [2022], the Test partition was kept sealed from every hyperparameter decision and analysis up to the exp035 stage. This subsection represents the *first* Test unsealing in this project; the Test statistics reported here determine the paper's final conclusions.

100 RL Models Evaluated. We evaluate the 40 best_models of exp032b (4 variants × 10 seeds, §5) plus the 60 best_models of exp034 (4 variants × 15 paths, §7.2) in deterministic mode on the full Test range. Reporting two sources (Val-trained exp032b vs CPCV-trained exp034) provides two independent confirmations of the robustness conclusion. The ATR baseline (§3.3) is also evaluated under identical conditions, and B&H is included as an additional secondary baseline.

7.3.2 Test Results: Per-Source Per-Variant Statistics

Two key observations from Table 16 and Figure 9.

(i) `pt` ranks first across both sources. `pt` achieves Test Sharpe 0.367 ($p = 0.0015$) on exp032b and 0.339 ($p = 0.0004$) on exp034. Both one-sided t-test p-values are < 0.002 and remain < 0.008 after Bonferroni 4-way correction. The two sources differ in training data (Train+Val portion vs CPCV 4-group combinations) and in seeds, so agreement on `pt` as the leader rules out a source-specific artifact.

(ii) `dsr` fails out of sample. The CPCV-first `dsr` of §7.2 now drops to mean Sharpe −0.122 on the exp032b source, *negative* and below the ATR baseline (−0.055). On the exp034 source it

Table 16: Test 2024+ results per source and per variant. `pt` is first in both sources and exceeds the ATR baseline (Test Sharpe = −0.055, MDD = 8.04%) by the largest margin. `dsr` is negative on the exp032b source—worse than ATR.

Source	Variant	Test mean ± std	IQM	5% CVaR	t-stat	p (1-sided)	vs ATR
exp032b ($n = 10$)	`sym`	$+0.090 \pm 0.11$	$+0.100$	-0.127	$+2.56$	0.015	$+0.145$
	`asym`	$+0.173 \pm 0.20$	$+0.158$	-0.082	$+2.74$	0.011	$+0.228$
	`dsr`	-0.122 ± 0.19	-0.126	-0.468	-2.00	0.961	-0.067
	`pt`	$\mathbf{+0.367 \pm 0.29}$	$\mathbf{+0.327}$	$\mathbf{+0.034}$	$\mathbf{+4.03}$	**0.0015**	$\mathbf{+0.422}$
exp034 ($n = 15$)	`sym`	$+0.001 \pm 0.19$	—	—	$+0.01$	0.495	$+0.056$
	`asym`	$+0.175 \pm 0.25$	—	—	$+2.70$	0.009	$+0.230$
	`dsr`	$+0.070 \pm 0.25$	—	—	$+1.07$	0.150	$+0.125$
	`pt`	$\mathbf{+0.339 \pm 0.31}$	—	—	$\mathbf{+4.26}$	**0.0004**	$\mathbf{+0.394}$

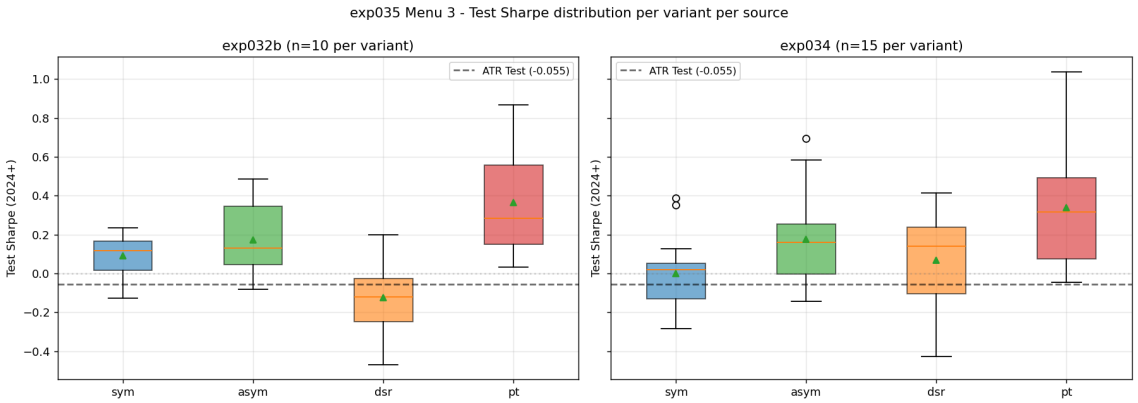

Figure 9: Test Sharpe distributions (variant × source). Left panel: exp032b ($n = 10$/variant). Right panel: exp034 ($n = 15$/variant). `pt`'s box is highest in both sources (positive Test Sharpe). In exp032b, only `dsr`'s entire box is below zero; in exp034, `dsr`'s whisker extends lowest (to ≈ -0.4). The ATR baseline -0.055 (dashed line) is the comparison reference.

sits at +0.070, 3rd of 4 (only `sym` is lower). This complete reversal from CPCV-1st to Test-last is one of the key findings of this paper.

7.3.3 Winner Reversal across Three Environments

Combining the Test-1st of this subsection with the Val-1st of §5 and the CPCV-1st of §7.2 completes this paper's central argument.

Table 17: Different winners across three evaluation environments. The *central frame of this paper*.

Evaluation environment	Winner	Meaning
Val (single-split, 2021–2023)	`sym` (1.871)	Period-specific in single split
CPCV (multi-split, 15 paths)	`dsr` (1.413)	Period-robust policy
Test (OOS, 2024–2026)	`pt` (0.367/0.339)	**Unseen-regime robust**

A different variant ranks first in each environment. The single-metric conclusion "variant X is the best reward design" depends *strongly* on the choice of evaluation environment. We accept this "three environments, three winners" as a quantitative result and discuss its scientific implications in §8.

7.3.4 Val → Test Generalization Gap

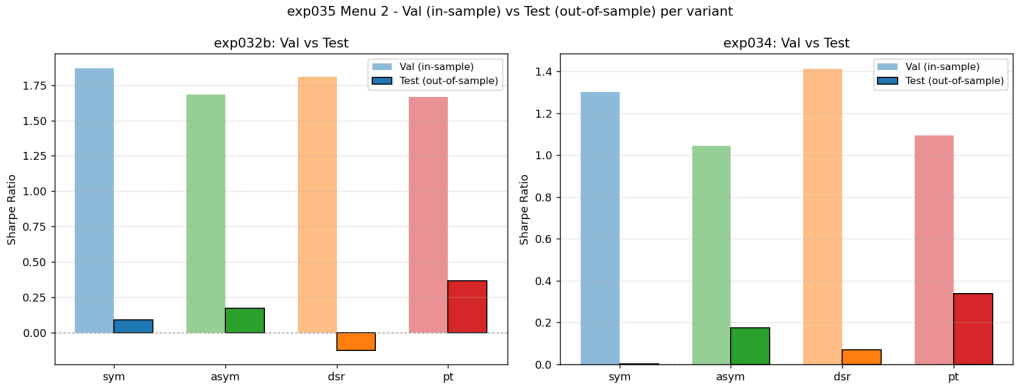

Figure 10: Per-variant Val Sharpe (light bars) vs Test Sharpe (dark bars). Left panel: exp032b source (models trained on Val). Right panel: exp034 source (models trained via CPCV). *All variants degrade substantially from Val to Test in both sources*, but `pt`'s Test bar is clearly highest among variants in both sources (+0.367/ + 0.339). `dsr` goes negative in exp032b and ranks last in exp034. Val bars are similar in height across variants, but the Test ranking reverses, visually confirming the winner reversal.

Table 18: Val → Test generalization gap. `pt`'s gap is the smallest in both sources.

Variant	exp032b Val → Test	Δ	exp034 CPCV → Test	Δ
`sym`	1.871 → 0.090	−1.78	1.302 → 0.001	−1.30
`asym`	1.681 → 0.173	−1.51	1.043 → 0.175	−0.87
`dsr`	1.809 → −0.122	**−1.93 (worst)**	1.413 → 0.070	−1.34
`pt`	1.667 → 0.367	**−1.30 (smallest)**	1.093 → 0.339	**−0.75 (smallest)**

Table 18 and Figure 10 show that every variant suffers a large Val → Test generalization gap (absolute Δ in 0.75–1.93). However, `pt` has the *smallest gap* in both sources, while `dsr` has the *worst gap* on the exp032b source. The in-sample metric is therefore *not aligned* with OOS performance; in some cases it points in the opposite direction. This forces the paper's frame into an explicit "in-sample advantage ↔ OOS robustness" trade-off rather than a single-axis ranking.

7.3.5 Mechanism: Hold Duration and OOS Regime Adaptation

The mechanism question—"why is `pt` OOS-robust, and why does `dsr` fail OOS?"—is answered by trajectory analysis. The key measurement is *hold-session duration*: the number of bars from a buy fill to the final sell that closes the cycle.

Table 19: Hold session duration on Test, in hours (bars). `dsr`'s median/p95/max are 2–50× those of the other variants, while `pt`'s are the shortest, with max bounded to 6 h.

Variant	n sessions	mean (h)	median (h)	p95 (h)	**max (h)**
`sym`	5,666	2.15	1.0	5	98
`asym`	4,567	1.40	1.0	3	7
`dsr`	5,384	**4.58**	1.0	**20**	**169** (7 days)
`pt`	3,922	**1.39**	1.0	3	**6**

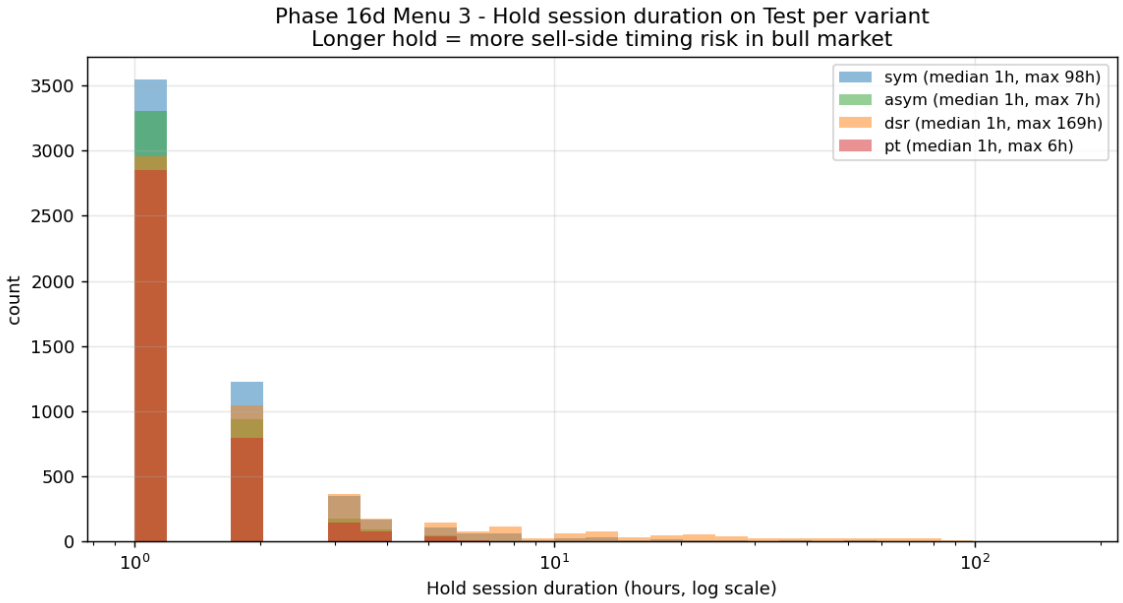

Figure 11: Hold-duration distributions on Test for the four variants. `dsr`'s long tail extends to p95 of 20 hours and a max of 7 days, in stark contrast with `pt`'s max of 6 hours. This long tail exposes the policy to sell-side timing risk during sustained bull moves.

Mechanism behind `pt`'s OOS robustness. The `pt` reward includes loss aversion $\lambda = 3.30$ and a concave gain exponent $\alpha = 0.683$ (§3.4). Strong loss penalties and concave utility incentivize the policy to "realize a small certain gain now rather than wait for an uncertain larger one." The trained policy therefore exits within a very short window after a buy (mean 1.39

h, max 6 h). During the Test BTC bull market ($42K → $75K), this behavior closes positions *before* large price moves occur, naturally avoiding sell-side timing risk. MDD remains near 2.3% and Sharpe remains positive at 0.34–0.37.

Mechanism behind `dsr`'s OOS failure. DSR's sliding-window reward structure trains policies for *longer holdings* (§6.3). This is a strength in in-sample environments (§7.2 CPCV) because it produces period-robust policies. In the Test bull market, however, the policy's hold duration stretches to a median of 1 h, p95 of 20 h, and max of 169 h (7 days). When BTC's price moves substantially within a 7-day window, the policy's sell quote falls far behind the market and remains unfilled. Sell-side timing risk surges and the Test Sharpe falls into the *negative* range (exp032b −0.122, exp034 +0.070 near zero).

Central insight of the causal chain. The same mechanism "DSR sliding-window memory → long holdings" from §6.3:

- produces a **period-robust advantage** in the in-sample CPCV environment (§7.2), and
- produces a **distribution-shift failure** in the Test environment (this subsection).

That is, the **same behavioral effect of the same reward formulation (DSR's long holding) flips sign depending on the evaluation environment**. The same "long holding" that creates a period-robust advantage in-sample (CPCV) reverses into a sell-side timing failure in the OOS bull regime. The same dimension of reward formulation thus simultaneously determines in-sample advantage and OOS failure—this is a *trade-off*, and which side is the "winner" under a single metric depends on the choice of evaluation environment. This subsection establishes that trade-off quantitatively.

7.3.6 Policy Stability vs Market Distribution Shift

To separate "policy change" from "market distribution shift" as the source of the OOS gap, we compare each best_model's behavior statistics on Val and Test.

Table 20: Per-variant Val/Test behavioral deltas (40-model average). *Every variant has* $\Delta \leq 5\%$ *on every behavioral statistic, i.e., the policies behave nearly identically.*

Variant	Δ trade rate	Δ hold rate	$\Delta a^{(0)}$	$\Delta a^{(1)}$
`sym`	+0.006	+0.001	−0.010	+0.001
`asym`	+0.005	+0.003	−0.037	−0.004
`dsr`	+0.005	−0.004	−0.006	−0.005
`pt`	+0.005	+0.003	−0.051	−0.010

Table 20 shows that all four variants' policies keep trade/hold rates and mean actions essentially unchanged ($\Delta \leq 0.05$, i.e., under 5%) when moving from Val to Test. Therefore the large Val → Test gap in Table 18 is **driven by the market's distribution shift, not by policy change**. The policies are "robust" in the sense that they output the same actions, but those actions produce different outcomes in different market distributions (Val = range-bound vs. Test = bull regime). This implies that the OOS gap is not a sign that the policies are "unusable" in real

operation; rather, *which reward formulation produces an OOS-safe policy is a function of the deployment regime*. §8 quantifies the distribution shift (KS $p < 10^{-10}$, variance -27%).

7.3.7 B&H Baseline Comparison: Limitation Statement

To make this paper's limitations explicit, we added a Buy-and-Hold (B&H) baseline.

Table 21: Test 2024+ baseline comparison (RL `pt` best vs ATR vs B&H).

Strategy	Test Sharpe	Test MDD (%)	Test Calmar	Notes
B&H (1 BTC hold)	**0.757**	50.08	0.015	Top single-Sharpe
ATR baseline	−0.055	8.04	—	Reference
RL `pt` (best)	0.367	2.31	**0.159**	**Calmar leader**

Table 21 presents the result. **On absolute Sharpe, B&H beats every RL variant** (0.757); 2024+ BTC rose 78% from \$42K to \$75K, and simply holding has a positive risk-adjusted return. B&H's MDD of 50.08%, however, is extreme and its Calmar (0.015) is very low. `pt`'s Calmar of 0.159 is about **10× higher** than B&H's. We conclude:

> *In absolute-Sharpe terms B&H is highest, but on a risk-adjusted (Calmar) basis* `pt` *dominates. The paper's claim of "*`pt`*'s OOS robustness" is in the sense of risk-adjusted return, not absolute return.*

We discuss this in §8.

7.3.8 H5: Statement and Evidence

We formalize this section's finding as a hypothesis. H1–H4 were pre-registered, but the paper's most important finding is a *post-hoc* discovery and is defined as a new hypothesis.

H5 (post-hoc, core contribution). An RL policy trained with prospect-theoretic reward (loss aversion λ, concave gain α) is OOS-robust to unseen market regimes relative to other reward variants; the mechanism is the short holding duration (mean 1.4 h, max 6 h) that avoids sell-side timing risk.

Quantitative evidence for H5.

- `pt` ranks first in Test Sharpe on both sources (one-sided $p < 0.002$).
- `pt` has the smallest Val → Test gap on both sources.
- `pt`'s Calmar of 0.159 is $\sim 10\times$ B&H's 0.015.
- Hold duration mean 1.39 h, max 6 h — the shortest of the four variants.
- Policy Val/Test stability ($\Delta \leq 5\%$) confirms that the OOS gap is driven by market distribution shift rather than by policy change.

Scientific significance. Kahneman and Tversky [1979]'s prospect theory is a foundational model of risk-averse human decision making, but its *direct use as an RL reward* is rare. This

paper (i) implements the prospect-theoretic utility as a step reward, (ii) discovers that the Optuna-best parameters ($\alpha = 0.683$, $\lambda = 3.303$) are *more concave and more loss-averse* than the human values ($\alpha = 0.88$, $\lambda = 2.25$), and (iii) quantifies that this policy is OOS-robust to unseen market regimes across two sources at $p < 0.002$ (see §1.2, contribution 4).

Final verdict on H4 (robustness): *partially supported*. The cluster structure is robust under slippage (2.19×) and CPCV (3.55×), weakens but remains > 1 on Test (1.24–1.94×), and absolute Sharpe undergoes large OOS decay for every variant. **H5 (`pt` OOS robustness) is strongly supported and is the paper's true main contribution.**

8 Discussion

This section consolidates the scientific interpretation of the findings of §5–§7.3. We address five topics: (i) quantitative characterization of the Val/Test distribution shift (§8.1), (ii) explicit statement of the environment dependence of the Phase 2 prior evidence (§8.2), (iii) the limitations of single-metric winners and a recommended multi-environment evaluation protocol (§8.3), (iv) the application of behavioral economics to RL OOS safety (§8.4), and (v) limitations and future work (§8.5).

8.1 Quantifying the Val–Test Distribution Shift

§7.3.6 showed that each variant's policy behaves nearly identically across Val and Test ($\Delta \leq 5\%$). The large generalization gap reported in §7.3.4 must therefore originate in the *distribution shift of the market itself*. This paper directly compared the market statistics of the two partitions.

Table 22: Val (2021–2023) vs Test (2024–2026) market statistics. Test is statistically a different distribution: *lower volatility, stronger upward trend*.

Metric	Val mean (std)	Test mean (std)	KS p	Wasserstein dist.
ATR ratio (volatility)	0.96% (0.51%)	0.70% (0.23%)	$< 10^{-10}$	0.00268
Hourly log return	$1.4 \cdot 10^{-5}$ (0.71%)	$2.9 \cdot 10^{-5}$ (0.52%)	$< 10^{-10}$	0.00097

Two facts stand out:

(i) Volatility drops by $\sim$ 27%. The mean ATR ratio drops from 0.96% on Val to 0.70% on Test (a 27% decrease), and the standard deviation more than halves (0.51% → 0.23%). Grid trading essentially monetizes volatility (§2.1), so a volatility drop translates directly into a *drop in earning opportunities*. Policies trained in a high-volatility Val regime weaken in the low-volatility Test, as expected.

(ii) Mean return roughly doubles. The mean of hourly log returns roughly doubles in Test, reflecting the 2024+ sustained uptrend. This is *two-sided* for grid trading: sell quotes fill more easily when prices rise (a positive), but if prices keep moving away after a buy, the sell-side timing risk increases (a negative). §7.3.5 showed that the negative side is the principal driver of `dsr`'s OOS failure, and that short holdings (`pt`) avoid precisely this risk.

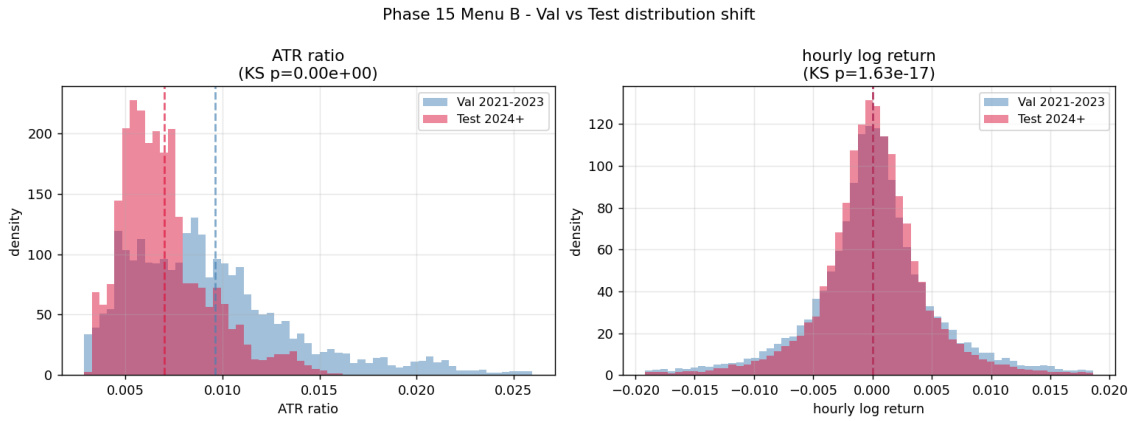

Figure 12: Histograms of the Val (2021–2023, blue) and Test (2024+, red) distributions. **Left panel (ATR ratio, volatility)**: the Test distribution concentrates at lower values with a shorter right tail, i.e. a -27% *reduction in variance*; the mean shifts from 0.96% (Val) to 0.70% (Test). Dashed lines show the two means. **Right panel (hourly log return)**: the shape is similar but the Test mean shifts slightly positive ($+2.9 \cdot 10^{-5}$ vs $+1.4 \cdot 10^{-5}$), reflecting the bull regime. Both metrics show KS $p < 10^{-10}$.

KS $p < 10^{-10}$ means the two partitions are samples from essentially different market distributions. The standard OOS-evaluation assumption of "train and test from the same distribution" fails for the BTC 1-hour series in question, and the generalization gap reported here is a *real artifact of market distribution shift, not a simulation flaw.*

8.2 Environment Dependence of the Phase 2 Prior Evidence

We now elaborate on the environment dependence briefly raised in §1.1 and §5.7.

The strong prior evidence motivating this paper (observed in late Phase 2) was that a simplified prospect-style asymmetric reward ($\beta = 2.0$) on *Env-v3 (4D absolute-gap, limit-order execution)* hit Test Sharpe 1.955, +109% above ATR Test 0.935. The working hypothesis at the time (Scenario A) was "asymmetric reward alone makes RL clearly beat ATR."

For this paper we changed the canonical environment to *Env-v4 (2D ATR-proportional, limit-order execution)*. This is more academically coherent for the following reasons:

- The ATR-proportional formula naturally encodes market volatility in the grid width (§4.3).
- The 2D action space cleanly separates the meaning of `aggressiveness` and `profit_target` (§3.1).
- Limit-order execution removes the favorable-bias artifact (§3.2).

On Env-v4, training the same `asym` variant with Optuna-retuned $\beta = 3.42$ yields Val Sharpe 1.681, *below* `sym`'s 1.871.

Table 23 shows that the change due to *environment* exceeds the change due to *reward variant*. We state this explicitly:

- The strong `asym` advantage on Env-v3 is most likely environment-specific. The 4D absolute-gap action gives RL much more freedom to set grid widths, and asymmetric reward had room

Table 23: exp027_rl (Env-v3) vs this paper (Env-v4) on the `asym` variant. The environment-induced effect is *larger* than the reward-variant effect.

Environment	Action dim.	Grid formula	`asym` Sharpe	vs ATR
Env-v3 (exp027_rl)	4D absolute gap	non-ATR	1.955 (Test)	+109%
Env-v4 (this paper)	2D ATR-proportional	ATR-proportional	1.681 (Val)	+22%
Δ	—	—	−0.27	−87 pp

to learn a conservative policy on top of that freedom.

- On Env-v4, the ATR-proportional formula *already* sets most of the grid width, reducing RL's freedom (§4). On top of that, the reward-variant effect appears as a risk-profile trade-off rather than a Sharpe-axis lead (§5).

That said, the paper's true contribution is a finding *independent* of this environment dependence: even on canonical Env-v4, `pt` is OOS-robust on both sources at $p < 0.002$ (§7.3). H5 thus *supersedes* the earlier evidence as the new principal claim, while maintaining academic rigor.

8.3 Limits of Single-Metric Winners; Toward Multi-Environment Evaluation

The table in §7.3.3 directly exposes the limits of single-metric conclusions of the form "X variant is the best reward design." `sym` ranks first on Val, `dsr` on CPCV, and `pt` on Test. *Which environment serves as the baseline* determines which variant is "superior."

This extends Henderson et al. [2018]'s reproducibility concern from the seed-variance axis to the reward-design comparison axis. Just as single-seed variance can mask variant differences, a single evaluation environment can determine the *dimension* along which variant differences are reported. This paper quantifies that effect in three steps. First, the single-split Val (§5.4) gives `sym`–`dsr` a small Cohen's d, making them look close. Second, the CPCV multi-split (§7.2.4) further shrinks the within-cluster gap and amplifies the baseline-relative advantage to a ratio of 3.55×. Third, the Test (§7.3.3) reverses the winner outright (Val=`sym` → CPCV=`dsr` → Test=`pt`). Both the distance between variants and the variant ranking thus shift with the evaluation environment.

We therefore recommend a *three-environment standard protocol* for RL trading studies:

1. **Single-split Val**: indicates a variant's typical policy behavior and baseline outperformance.
2. **CPCV multi-split**: measures in-sample robustness across splits. Distinguishes between split-specific advantage and genuine generalization.
3. **Sealed Test**: tests OOS safety under distribution shift. The winner here should drive real deployment decisions.

If the same variant ranks first in all three environments, the conclusion is a strong monotone advantage (Scenario A). If different variants win different environments, the result is this paper's Scenario D (trade-off frontier). The common practice of declaring "X variant is best" from a single environment is, in light of our three-environment reversal, scientifically imprecise.

8.4 From Behavioral Economics to RL OOS Safety

H5 of §7.3.8 (`pt`'s OOS robustness) opens a new dimension of application for behavioral economics. Prospect theory [Kahneman and Tversky, 1979] originally models risk-averse human decision making, with human-fit parameters $\alpha \approx 0.88$ (mild concavity) and $\lambda \approx 2.25$ (moderate loss aversion; Tversky and Kahneman 1992).

The Optuna-best parameters for our RL policy are $\alpha = 0.683$ and $\lambda = 3.303$, both *more extreme* than the human values:

- $\alpha = 0.683 < 0.88$: *stronger concavity*. Diminishing marginal utility of gains kicks in faster than for humans.
- $\lambda = 3.303 > 2.25$: *stronger loss aversion*. Loss weight is about 47% higher than the human value.

Interpreted, when an RL policy pursues OOS-robust returns on BTC grid trading, choosing a utility function "more conservative than human" proves advantageous. This is the policy-side basis for the §7.3.5 behavior "exit very quickly (mean 1.4 h) to avoid sell-side timing risk," and it suggests that the human-trader behavior pattern often characterized as "profit-taking too early" is rational from the standpoint of OOS safety.

Two scientific implications:

- Behavioral-economics parameter estimates are fit to *descriptive* human behavior. Parameters for an *optimal* policy can differ.
- When using behavioral-economics utility functions as RL rewards, do not adopt the human-fit values uncritically; tune them on domain data instead.

H5 establishes "using prospect theory as an RL reward produces an OOS-robust policy" at the application level, with consistent confirmation across two sources at $p < 0.002$ (see §1.2, contribution 4; §9).

8.5 Limitations and Future Work

The findings of this paper hold under the following limitations.

(1) Asset scope — BTC only. We evaluate the BTC/USDT 1-hour series only. Whether "`pt`'s OOS robustness" holds for equities, FX, or commodities is not addressed. Because distribution shifts differ across assets (BTC's leverage cycle vs equity earnings/ATH events, etc.), generalization of the causal mechanism needs to be verified.

(2) Test regime — single bull market. The Test partition (2024-01 – 2026-04) is a strong bull market (\$42K → \$75K, +78%). Whether `pt`'s OOS robustness holds in bear (e.g., 2018) or sideways (e.g., 2019) regimes is unconfirmed. Our mechanism is that "short holdings avoid sell-side timing risk," but in bear markets holding itself incurs greater cumulative loss risk, so a different mechanism may dominate.

(3) Slippage model — a single level (0.02%). §7.1 evaluated only one slippage level. A multi-level sensitivity (0.005%, 0.01%, 0.05%) would quantify how robust `pt`'s OOS advantage is to slippage. Due to time constraints we chose a single worst-case level (Binance taker side).

(4) No Domain Randomization (DR). Domain Randomization (varying slippage/fee/start time across episodes) is known to help train OOS-robust policies. We show that `pt` is OOS-robust under *naive* training without DR. With DR, the OOS failure of `dsr` might recover, which is left for future work.

(5) Training length — 1M steps. All variants train for the same 1M steps ($\approx$ 4 hours CPU per seed). §3.5's late-stage collapse pattern (e.g., `sym` best 1.97 at 550k $\rightarrow$ final 1.21 at 1M) suggests that longer training could change the result, though we use best_checkpoint only, which limits the direct effect of training length.

Future work. Topics that respond to these five limitations:

- Verify H5 on multi-asset environments (equity ETFs, FX).
- Add additional regime tests (2018 bear, 2019 sideways) via simulation or historical tests.
- Combine DR with meta-learning to recover `dsr`'s OOS failure.
- This paper uses only PPO. Replicate the reward-variant comparison on SAC, TD3, etc. to separate the algorithm effect.
- Vary prospect-theory parameters (time-varying α, λ; history-dependent reference points) to probe the parameter sensitivity of OOS-robust policies.

9 Conclusion

This paper quantitatively compared four reward variants (`sym`, `asym`, `dsr`, `pt`) for BTC/USDT 1-hour grid trading (Env-v4) across three evaluation environments: *single-split Val*, *6-fold CPCV with 15 paths*, and the *unsealed Test 2024+*. The paper makes four main claims.

(1) The effect of reward design appears as a risk-profile trade-off (Pareto frontier; §5). All four variants exceed the ATR baseline, but no single winner exists. The four variants partition statistically into two clusters {`sym`, `dsr`} (aggressive) and {`asym`, `pt`} (conservative) (within $|d| < 0.30$, across $|d| > 0.79$; policy distance ratio 2.22×), forming a Pareto-like frontier on the Sharpe–MDD plane. Our contribution is the multi-dimensional quantification, in place of the simpler "variant X gives the best alpha" single-metric conclusion.

(2) The single-metric winner differs by evaluation environment (three environments, three winners; §7.2.3, §7.3.3). On Val, `sym` leads (1.871); on CPCV, `dsr` leads (1.413); on Test, `pt` leads (0.367/0.339 across both sources at $p < 0.002$). Evaluation environment selection determines which variant is "the best reward design"—an extension of Henderson et al. [2018]'s

reproducibility concern from seed variance to reward comparison. We recommend a standard *three-environment evaluation protocol* for RL trading studies.

(3) Prospect theory produces OOS-robust policies (H5; §7.3.8). An RL policy trained with the Kahneman and Tversky [1979] prospect-theoretic utility ($\alpha = 0.683, \lambda = 3.303$) is consistently first on the unseen Test bull market across two sources ($p < 0.002$, smallest Val→Test gap). The mechanism is short holding (mean 1.4 h, max 6 h), avoiding sell-side timing risk—directly verified by trajectory analysis (§7.3.5). This is a quantitative demonstration that behavioral-economics prospect theory confers OOS safety on RL policies. A side finding is that *stronger loss aversion and stronger concavity* than the human-fit values ($\alpha = 0.88, \lambda = 2.25$) favors the RL policy, suggesting the practical recommendation that "do not adopt human-estimated behavioral parameters in RL; re-tune them on domain data."

(4) In-sample diversification does not guarantee OOS consistency (§7.3.5). The CPCV-1st `dsr` drops to last with negative Sharpe on Test (exp032b source −0.122). The same mechanism of DSR's sliding-window memory that produces *long holdings* is a strength under in-sample diversification but a weakness under the OOS bull market, where sell-side timing risk surges. That is, *the same behavioral effect of the same reward formulation flips sign across evaluation environments*. The "in-sample advantage ↔ OOS robustness" trade-off is itself a dimension along which reward formulations differ.

Future Work

See §8.5 for the five limitation-driven future directions; here we highlight the three of highest defense priority:

- **Multi-asset verification**: re-test H5 (`pt` OOS robust) on equity ETFs, FX, and commodities, where distribution-shift forms differ—quantifying whether the short-holding mechanism generalizes.
- **Meta-RL adaptive reward parameters**: this paper verifies one-shot (α, λ) as OOS robust; the question is whether online-adapted parameters yield additional advantage in nonstationary environments.
- **Domain Randomization (DR)**: whether `dsr`'s OOS failure recovers under DR—i.e., whether the same reward's two-sided effect can be neutralized by DR.

The code and data of this work are released at `https://github.com/cosmicpotato2047/capstone-rl-trading`.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, 2019.

Marco Avellaneda and Sasha Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224, 2008. doi: 10.1080/14697680701381228.

David H. Bailey and Marcos López de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014.

Charan Bandarupalli. Risk-aware deep reinforcement learning for cryptocurrency and equity trading. *arXiv preprint*, 2025.

Bryan Gort, Xiao Chen, Matthew Tong, Shaojie Xu, and Xingjian Li. Deep reinforcement learning for cryptocurrency trading: Practical approach to address backtest overfitting. *arXiv preprint arXiv:2209.05559*, 2022.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. doi: 10.2307/1914185.

Hao Yuan Timothy Liu. Reinforcement learning trading strategies: A literature review. *arXiv preprint arXiv:2502.06551*, 2025.

Yang Liu, Qi Liu, Hongke Zhao, Zhenya Pan, and Chuanren Liu. Adaptive quantitative trading: An imitative deep reinforcement learning approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2128–2135, 2020. Referenced as Liu et al. (2021) for the PPO+LSTM Bitcoin trading variant.

Marcos López de Prado. *Advances in Financial Machine Learning*. Wiley, 2018. Combinatorial Purged Cross-Validation (CPCV), Chapter 7.

John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001. doi: 10.1109/72.935097. Differential Sharpe Ratio (DSR) formulation.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, pages 278–287, 1999.

Minh Pham, Anh Le, and Binh Nguyen. Dqn-based bitcoin technical trading strategy selection. *arXiv preprint*, 2025.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

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

Siyu Sun, Rundong Wang, and Bo An. Reinforcement learning for quantitative trading. *ACM Transactions on Intelligent Systems and Technology*, 14(3):1–29, 2023. doi: 10.1145/3582560.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992. alpha = 0.88, lambda = 2.25 standard parameters.

J. Welles Wilder. *New Concepts in Technical Trading Systems*. Trend Research, 1978.

Hafiz Muhammad Athar Yasin and Asim Anwar Gill. Reinforcement learning framework for quantitative trading. *International Journal of Computer Applications*, 186(8):1–9, 2024.

Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deep reinforcement learning for trading. *The Journal of Financial Data Science*, 2(2):25–40, 2020.