Portfolio Reinforcement Learning Under Transaction Costs: A Modular Research Framework

Noah Donovan

Abstract

We study portfolio-level reinforcement learning (RL) with explicit transaction costs under a walk-forward evaluation protocol. Using daily SPY/QQQ data from 2015–2024 and a causal feature set (returns and rolling volatility), we compare a minimal PPO baseline against simple portfolio baselines. We report risk-adjusted metrics and exposure diagnostics under embargoed splits. The equal-weight baseline achieves positive Sharpe, while the minimal PPO baseline underperforms with substantially higher turnover. These results are intended as a reproducible, conservative baseline rather than a claim of profitability.

1 Introduction

Portfolio RL for trading requires strict causality, explicit transaction costs, and leakage-resistant evaluation. We present a modular framework with deterministic simulation and walk-forward evaluation, suitable for reproducible research and skeptical review.

Contributions. (i) A modular financial RL library with explicit cost modeling and deterministic execution; (ii) a leakage-safe feature pipeline and walk-forward evaluation with embargo; (iii) baseline agents and a minimal PPO scaffold as a conservative research baseline.

2 Related Work

We position this framework relative to prior work in financial RL, portfolio management, and robust backtesting, emphasizing prevention of lookahead bias, explicit cost modeling, and evaluation under walk-forward protocols.

3 System Architecture

The framework is decomposed into data, features, environment, execution, reward, agents, and experts. Each module exposes a strict interface to avoid cross-layer leakage.

3.1 Data and Features

Data sources expose time-indexed OHLCV bars. Feature transformers are causal: feature vectors at time t use only data up to t. We include causal returns and rolling volatility with explicit handling of insufficient history.

3.2 Environment and Execution

The portfolio environment executes target-weight actions at time t using a deterministic execution model with proportional commission and slippage. Rewards are realized using mark-to-market value at $t + 1$ and include explicit cost penalties.

3.3 Agents and PPO Baseline

Agents output portfolio target weights. We provide baseline agents (cash-only, equal-weight, hold) and a minimal PPO implementation with a linear Gaussian policy and value head, designed for reproducibility rather than performance.

4 Experimental Protocol

We adopt walk-forward evaluation with an embargo gap to reduce leakage. Splits are defined by a training window, an embargo, and a forward test window. Metrics include cumulative return, max drawdown, realized volatility, Sharpe, Sortino, and Calmar ratios, plus exposure and turnover diagnostics. We also report bootstrap confidence intervals (CIs) for mean Sharpe and cumulative return across splits.

5 Data and Features

We use daily OHLCV data for `SPY` and `QQQ` from 2015-01-02 to 2024-12-31 (2516 trading days), obtained from the Stooq public dataset. We align assets by the intersection of available trading dates and use raw OHLCV values without additional corporate action adjustments. The feature set is the concatenation of 1-day causal returns and 10-day rolling volatility.

6 Environment and Execution Model

The portfolio environment executes target-weight actions at time t using the current bar price with proportional commission and slippage (5 bps each). Rewards are log returns with explicit transaction cost penalties.

7 Methods

We evaluate baseline agents (cash-only and equal-weight) and a minimal PPO policy with a linear Gaussian policy and value head. The PPO training is intentionally lightweight (5 iterations, rollout length 64) to provide a conservative baseline rather than a performance-optimized result.

8 Results

Summary of findings. On this dataset, the equal-weight baseline achieves positive Sharpe with low turnover; the minimal PPO baseline underperforms with substantially higher turnover and modest risk-adjusted metrics.
Bootstrap CIs for mean Sharpe and mean cumulative return across splits are reported in the experiment summaries and can be reproduced from the exported JSON files.

Table 1: Walk-forward performance summary (mean across splits).

Agent	Cum. Ret.	Max DD	Sharpe	Sortino	Calmar	Mean Turnover
Cash-only	0.000	0.000	0.00	0.00	0.00	0.0000
Equal-weight	0.178	0.155	1.29	1.72	2.94	0.0061
PPO (minimal)	-0.008	0.147	0.30	0.34	0.76	0.5767

9 Discussion

The results highlight the importance of explicit cost modeling, leakage-safe features, and conservative evaluation. The minimal PPO baseline trades frequently and underperforms a simple equal-weight portfolio, underscoring that stronger training and richer models are required before drawing performance conclusions.

10 Limitations and Future Work

Future work includes richer execution models, alternative reward formulations, and more expressive policy classes. We emphasize the need to validate improvements under walk-forward protocols with embargo and realistic transaction costs.

11 Conclusion

We introduced a modular, deterministic financial RL framework with explicit transaction costs and a walk-forward evaluation protocol. The system is designed for reproducible research and provides baseline agents and a minimal PPO scaffold. The current results serve as a conservative baseline for future, more extensive experimentation.