A Modular Research Framework for Portfolio Reinforcement Learning Under Transaction Costs

Anonymous

January 31, 2026

Abstract

We present a modular, research-grade framework for portfolio-level reinforcement learning with explicit transaction costs, deterministic market simulation, and walk-forward evaluation. The system separates data, features, environment, execution, reward, agents, and experts via strict interfaces to reduce leakage risk and improve reproducibility. We describe the architecture, execution-cost modeling, and evaluation protocol, and provide baseline agents and a minimal PPO implementation for end-to-end experiments. Empirical results are reported on real market data with walk-forward splits and an embargo to minimize leakage.

1 Introduction

Reinforcement learning for trading requires careful handling of causality, transaction costs, and evaluation to avoid overly optimistic conclusions. We focus on portfolio-level decision making under explicit trading costs and deterministic simulation, with clear module boundaries to support rigorous experimentation.

Contributions. (i) A modular financial RL research library with explicit transaction 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 for reproducible research.

2 Related Work

We position this framework relative to prior work in financial RL, portfolio management, and robust backtesting. We emphasize 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 Evaluation 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.

5 Experiments

This section describes datasets, preprocessing, and experimental setup. Results use real market data and walk-forward splits with embargo.

5.1 Datasets and Preprocessing

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 the raw OHLCV values without additional corporate action adjustments. The aligned dataset is converted to a multivariate array for the environment.

5.2 Training and Evaluation Setup

We evaluate using walk-forward splits with a 504-day training window (˜2 years), a 5-day embargo, and a 252-day test window (˜1 year), advancing by 252 days per split. The execution model uses proportional commission and slippage (5 bps each). Features are the concatenation of 1-day causal returns and 10-day rolling volatility. The PPO baseline uses a linear Gaussian policy with 5 training iterations and rollout length 64 per split. All results are averaged across splits.

5.3 Results

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	-0.008	0.147	0.30	0.34	0.76	0.5767

6 Discussion

We discuss the importance of explicit cost modeling, leakage-safe features, and walk-forward evaluation. We also outline limitations of the baseline PPO and directions for more realistic execution modeling.

7 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.

8 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.