Dynamic Portfolio Optimization with Proximal Policy Optimization (PPO)

Daniel Lee[1], Feolu Kolawole[1], Vedant Srinivas[1]

(leedan, flukol, vedants8)@stanford.edu

Department of Computer Science, Stanford University

Stanford Computer Science

Introduction

- Traditional portfolio strategies (60/40, equal weight) fail to adapt during regime shifts
- We frame daily asset allocation as a reinforcement learning problem using PPO
- Inputs: State vector containing market indicators and portfolio features
- Outputs: Continuous allocation weights across 10 ETFs + cash, with realistic trading constraints
- **Goal: learn adaptive allocation policies that outperform standard financial baselines**

Dataset & Features

Since the inputs are financial time series, we derive features from OHLCV data [1], technical indicators, and the portfolio's current state, yielding a 392-dimensional observation vector.

Market Features (378 total)

Multi-horizon returns (50), Volatility (40), Momentum (40), Technical indicators- RSI, MACD, Bollinger Bands, and ATR (90), Cross-asset correlations (45), Statistical features (20), Regime indicators (3), Base prices/returns (30)

Portfolio Features (14 total)

Current weights for each ETF (SPY, QQQ, IWM, EFA, EEM, TLT, IEF, GLD, DBC, VNQ) + cash, portfolio value, recent return, and drawdown.

Environment Setup

We formulate portfolio management as a Markov Decision Process (MDP).

At each timestep *t*, the agent can take action by reallocating capital across 10 ETFs + cash.

- Transaction costs use rate c = 0.001
- Reward is normalized daily return minus cost

$$r_t = 100\left(\frac{V_{t+1}-V_t}{V_t}\right) - 100\frac{C_t}{V_t}.$$

Constraints:

- Portfolio weights must sum to 1
- Weight for a single holding cannot exceed 0.6

Objective: Identify a policy that maximizes expected discounted return.

Methods & Experiments

We train an ensemble of three Actor-Critic agents, using PPO clipping to prevent the model from overreacting to market noise. The final strategy averages their outputs to ensure stable, robust decision-making.

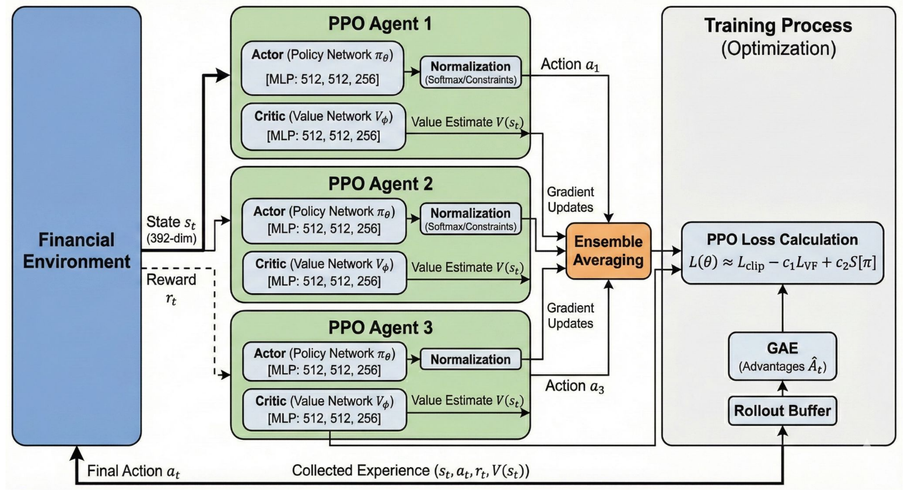

Figure 1: Overview of the Ensemble PPO Framework

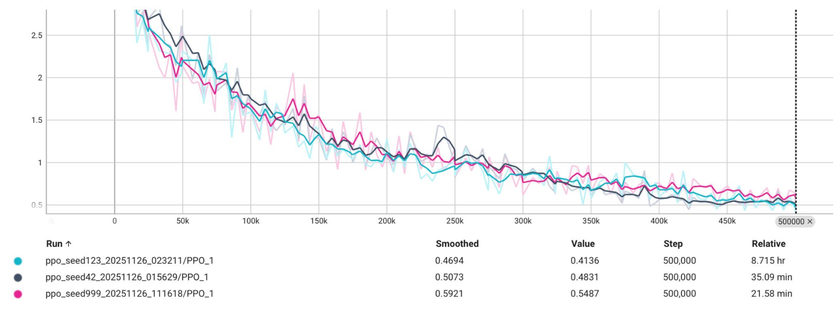

Figure 2: Value Loss Curves

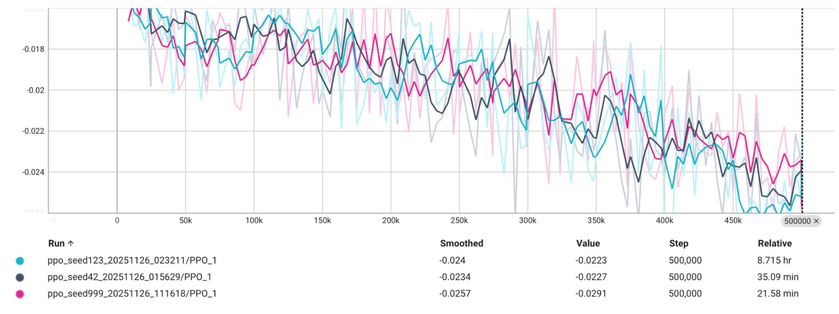

Figure 3: Policy Gradient Loss Curves

Results

Ensemble PPO achieved the strongest performance, with a **15.76% annualized return** and a **1.436 Sharpe ratio** on the 2022–2024 test set, outperforming all baseline strategies.

Figure 4: Normalized Portfolio Value

Strategy	Total Return	Ann. Return	Volatility	Sharpe	Sortino	Max DD	Calmar	Final Value
Ensemble PPO	**34.08%**	**15.76%**	10.97%	**1.436**	**2.222**	**-10.00%**	**1.575**	**$134,078**
60/40 Portfolio	28.48%	13.32%	10.66%	1.250	1.970	-12.60%	1.057	$128,356
Equal Weight	25.47%	11.99%	10.02%	1.196	1.845	-9.59%	1.250	$125,340
Risk Parity	22.93%	10.85%	9.35%	1.161	1.800	-8.83%	1.229	$122,809

Table 1: Summary of Metrics

Discussions & Future Research

Discussion: Our ensemble PPO model outperformed all baselines, achieving the highest returns and Sharpe ratio. Its stability comes from ensembling, transaction-cost penalties, and a broad feature set that supports adaptation to changing market conditions. The main limitation was slower response during sharp reversals, but overall the method captured allocation patterns that fixed strategies could not.

Future Research: Future work could incorporate sequence models such as LSTMs or Transformers to improve responsiveness to rapid market regime changes. Expanding the asset universe and exploring online or continual learning may further enhance generalization and real-time adaptability.

References:

[1] "Yahoo Finance," Yahoo! Finance, https://finance.yahoo.com/ (accessed Dec. 8, 2025).