DRL Portfolio Optimization

Andy Au

Aug 25, 2025

Summary of results

Portfolio	AnnExcessRet	Vol	Sharpe	MaxDD	CAGR
RL (PPO)	24.63%	12.31%	2.00	-6.35%	32.13%
Markowitz	31.68%	18.06%	1.75	-5.80%	40.56%
Naive	12.67%	15.44%	0.82	-12.05%	16.48%
SPY	1.49%	26.48%	0.06	-19.00%	0.78%

- PPO highest Sharpe Ratio (SR)
- Markowitz highest raw and CAGR but also higher vol.
- RL attains 22.3% lower return but also 31.8% lower vol vs Markowitz
- MC tail positioning (1,000,000 sims): Markowitz Sharpe 1.75 $\approx$ top 0.3%; RL Sharpe 2.00 $\approx$ top 0.01% of simulated paths.

Goal of Project

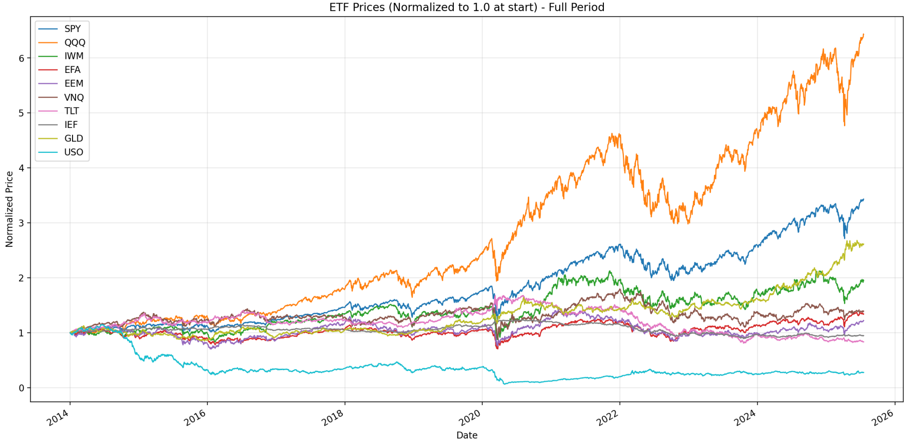
- Develop a PPO agent that learns allocation policy directly from engineered market features.
- PPO chosen for stability in continuous action spaces and sample efficiency
- Allocate capital daily across 10 liquid ETFs (SPY, QQQ, IWM, EFA, EEM, VNQ, TLT, IEF, GLD, USO).
- Benchmark vs:
  - Naive equal weight.
  - Markowitz mean–variance.
  - Monte Carlo random allocation envelope.

Data

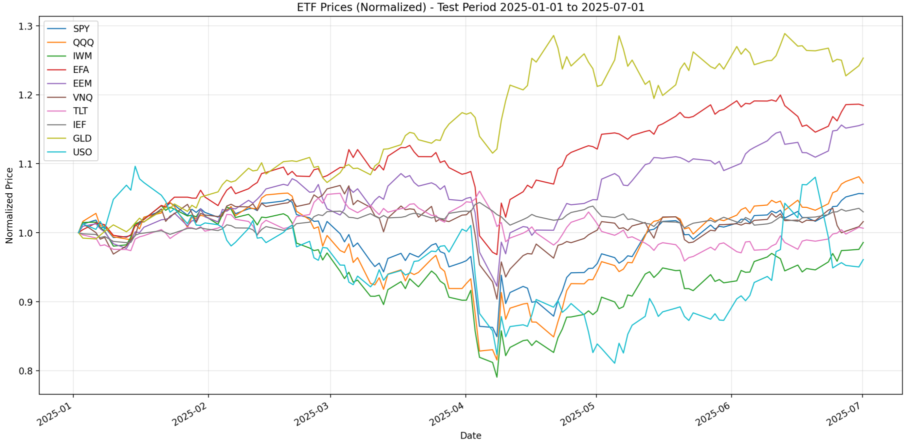
Test Period: 2025-01-01 to 2025-07-01 (6 months)

- Annual RFR = 0.04 for all Sharpe calculations
- Markowitz uses 2024-07-01 onward.
- RL uses 2019-01-01 onward.
- RL Feature stack (274 dimensions):
    - stacked normalized log-return lags (63d z-scores), lags 0–9
    - multi-horizon simple returns 1,5,21,63d
    - extra momentum returns 20d,60d (not in base set)
    - RSI(14)
    - realized volatility windows 5,21,63
    - downside semivol windows 21,63
    - cross-sectional percentile ranks (21d return, 21d vol)
    - rolling mean pairwise correlation (window 21) per asset
    - absolute daily returns
    - cyclical time (day_sin, day_cos, month_sin, month_cos)

Full ETF Price Visualization



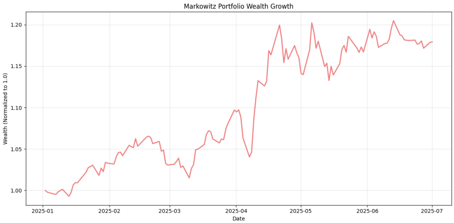
Test ETF Price Visualization

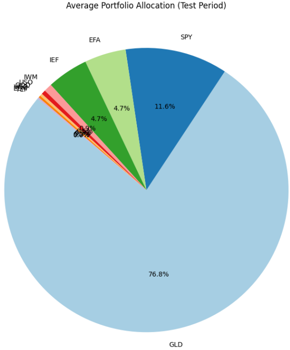


Markowitz Implementation

- Input: Historical return window, compute $\hat{\mu}$ and $\hat{\Sigma}$.
- Solve for max Sharpe with constraints (long only, sum to 1).
- Rebalance at fixed frequency (daily) with rolling lookback
- Output metrics; very standard, nothing fancy.

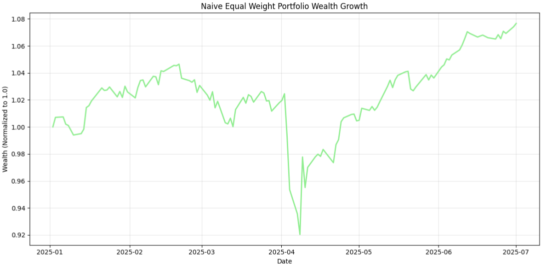
Markowitz Results





- Sharpe: 1.75
- Lookback: 6 months
- Rebalancing: Daily
- Annualized Excess Return: 31.68%
- Annualized Volatility: 18.06%
- Notable sensitivities: Covariance noise, regime shifts, low dimensionality.

Naive Implementation + Results



- Equal weight daily
- No transaction cost
- Sharpe: 0.82
- Annualized Excess Return: 12.67%
- Annualized Volatility: 15.44%
- Provides baseline risk-adjusted performance.

RL Implementation: Environment

- Observation: Feature vector + previous weights
- Action: Unconstrained logits $\rightarrow$ temperature + clipping $\rightarrow$ softmax weights.
- Constraints: Per asset caps (35% training; relaxed in refit to 80%).
- Turnover cost modeled linearly (daily rebalancing, configurable bps).

RL Reward Shaping

- Base: Excess portfolio return $r_p - r_f -$ turnover cost.
- Movement bonus (Encourages adaptive reallocations).
- Momentum term (Alignment with price trends).
- Variance penalty (Penalize high var over rolling window).
- Two sided HHI band:
  - Penalize over concentration (HHI too high).
  - Penalize uniform stagnation (HHI too low).
- Advantage tilt (Encourage assets with above avg returns).
- Optional L2 regularization (Penalize large action logits)
- Optional reward normalization (Scales rewards rolling).

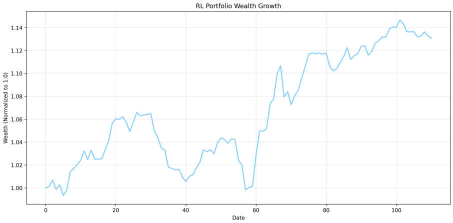
RL Process: Training

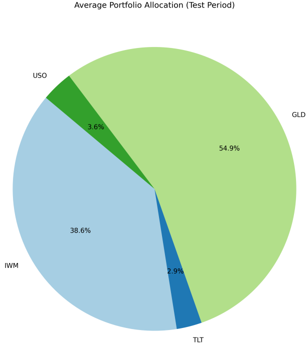
- Algorithm: PPO (SB3) with SDE, entropy / KL / logit-clip annealing.
- Validation: Multi-window Sharpe with soft worst-window penalty; early stopping on adjusted mean Sharpe.
- Saved checkpoints: Best model by validation; final model.
- Feature normalization frozen at end of training period.

RL Process: Monthly Refit

- Months: 2025-01-01 to 2025-07-01
- For each month:
    i. Freeze normalization up to prior day
    ii. Refit (fine tune) on recent 90-day slice
    iii. Evaluate within that month (no leak)
- Refit overrides: Lower turnover cost, higher max position size for adaptivity.
- Allows agent to learn recent market regimes; could potentially try denser refit windows but risks overfitting

RL Results: Wealth + Allocations



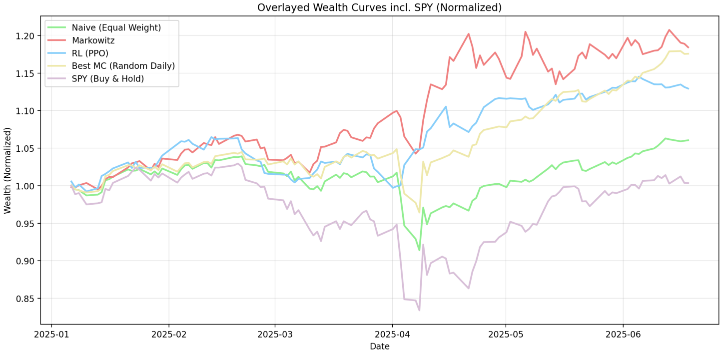


- Sharpe: 2.00
- Annualized Excess Return: 28.63%
- Annualized Volatility: 12.31%
- HHI (Concentration): 0.437
- Max Asset Weight: 54.0%

Monte Carlo process

- Simulate 1,000,000 random daily allocation paths over test window to build performance envelope
- For each path: sample daily weight vector from a Dirichlet(1) , apply daily rebalancing to the test-period prices, compute wealth series starting at 1.0
- Compute metrics, collect distributions etc
- Among 1M random daily rebalancing simulations, RL's 2.00 Sharpe $\approx$ 99.99th percentile; Markowitz's 1.75 $\approx$ 99.7th percentile"

Comparison (Normalized Wealth)



- RL dominates risk-adjusted path vs Naive and Markowitz.
- Monte Carlo “best Sharpe path” contextualizes chance extremes.

Quantitative Comparison

Portfolio	AnnExcessRet	Vol	Sharpe	MaxDD	CAGR
RL (PPO)	24.63%	12.31%	2.00	-6.35%	32.13%
Markowitz	31.68%	18.06%	1.75	-5.80%	40.56%
Naive	12.67%	15.44%	0.82	-12.05%	16.48%
SPY	1.49%	26.48%	0.06	-19.00%	0.78%

- PPO highest Sharpe Ratio (SR)
- Markowitz highest raw and CAGR but also higher vol.
- RL attains 22.3% lower return but also 31.8% lower vol vs Markowitz
- MC tail positioning (1,000,000 sims): Markowitz Sharpe 1.75 $\approx$ top 0.3%; RL Sharpe 2.00 $\approx$ top 0.01% of simulated paths.

Conclusion

- PPO with structured reward outperformed traditional baselines on 2025 test window.
- Monthly refit improved regime responsiveness without large overfit footprint.
- Diversification band + movement + advantage components yielded balanced exploration/adaptation.
- Future enhancements:
  - GRPO
  - Attention based policy (Didn't work well when I tried, maybe doing it wrong)
  - Ensembling
  - CVaR penalties
  - Bayesian / shrinkage layer for Markowitz baseline fairness.