REINFORCEMENT LEARNING FOR PORTFOLIO OPTIMIZATION

Aditi Gupta(2024701024)
Mohammad Wajahat(2024901002)
Ayan Datta (2021114017)

Portfolio Optimization

Process of selecting the optimal mix of financial assets that maximize returns while minimizing risk.

1. Single-Asset Portfolio Optimization: Trade one asset (e.g., AAPL)
2. Multi-Asset Portfolio Optimization: Allocate funds across assets
3. Risk-Awareness: Integrate Risk constraints.

Mean Variance optimization

Goal: Compared Mean-Variance Optimized Portfolio vs Equal-Weighted Portfolio using Dow 30 stocks.

Weights were optimized using Mean-Variance Optimization to maximize the Sharpe Ratio:

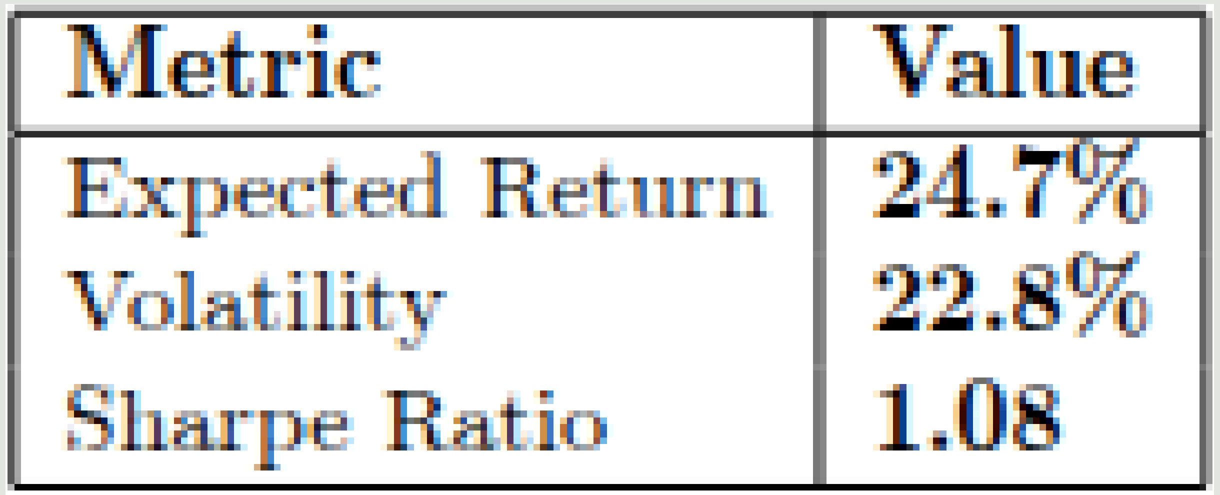
Sharpe Ratio = $E[Rp] - Rf/\sigma p$, where Rf is the risk-free rate

Optimized Portfolio Weights

Only three stocks were selected:

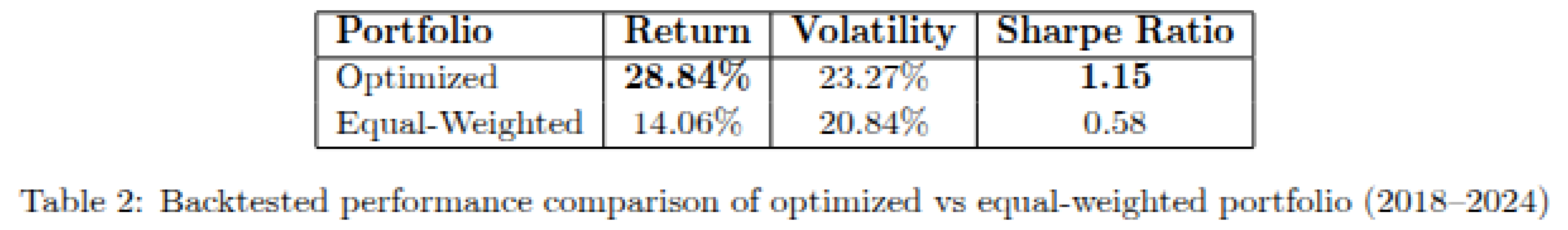
- AAPL: 27.60%
- MRK: 35.74%
- MSFT: 36.66%

All other 26 stocks received zero allocation, only these three had the best risk-adjusted performance.



Metric	Value
Expected Return	24.7%
Volatility	22.8%
Sharpe Ratio	1.08

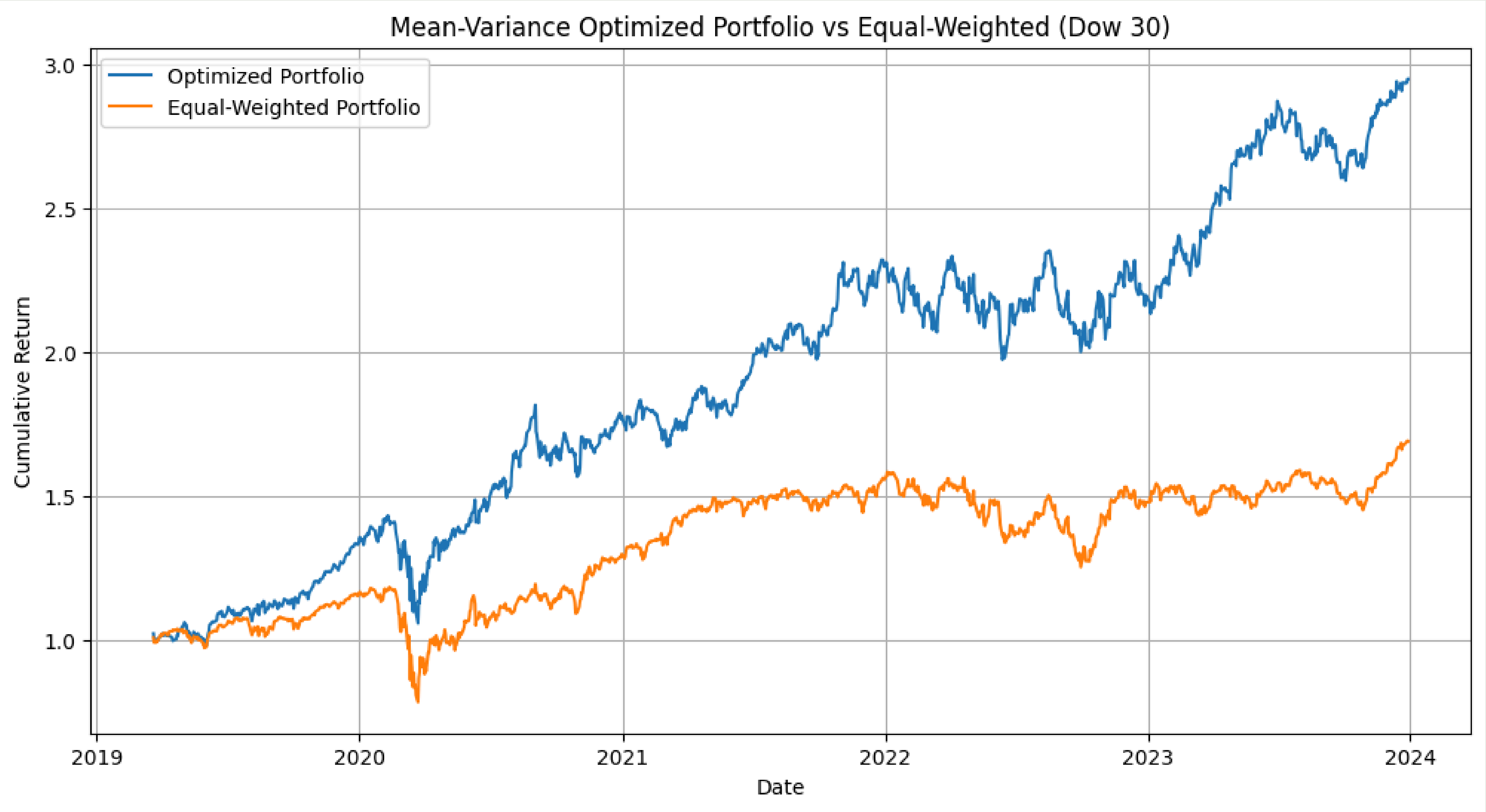
Backtested performance comparison of optimized vs equal-weighted portfolio(each stock equally weighted at ~3.33%) (2018–2024)



Portfolio	Return	Volatility	Sharpe Ratio
Optimized	**28.84%**	23.27%	**1.15**
Equal-Weighted	14.06%	20.84%	0.58

Table 2: Backtested performance comparison of optimized vs equal-weighted portfolio (2018–2024)

Doubled the return of the equal-weighted portfolio, with a significantly higher Sharpe Ratio, indicating superior risk-adjusted returns.



- Over-concentration: Only 3 out of 30 stocks were selected. This makes the portfolio vulnerable to stock-specific events.
- Look-ahead bias: Using historical data assumes future will resemble the past, which is not always true.

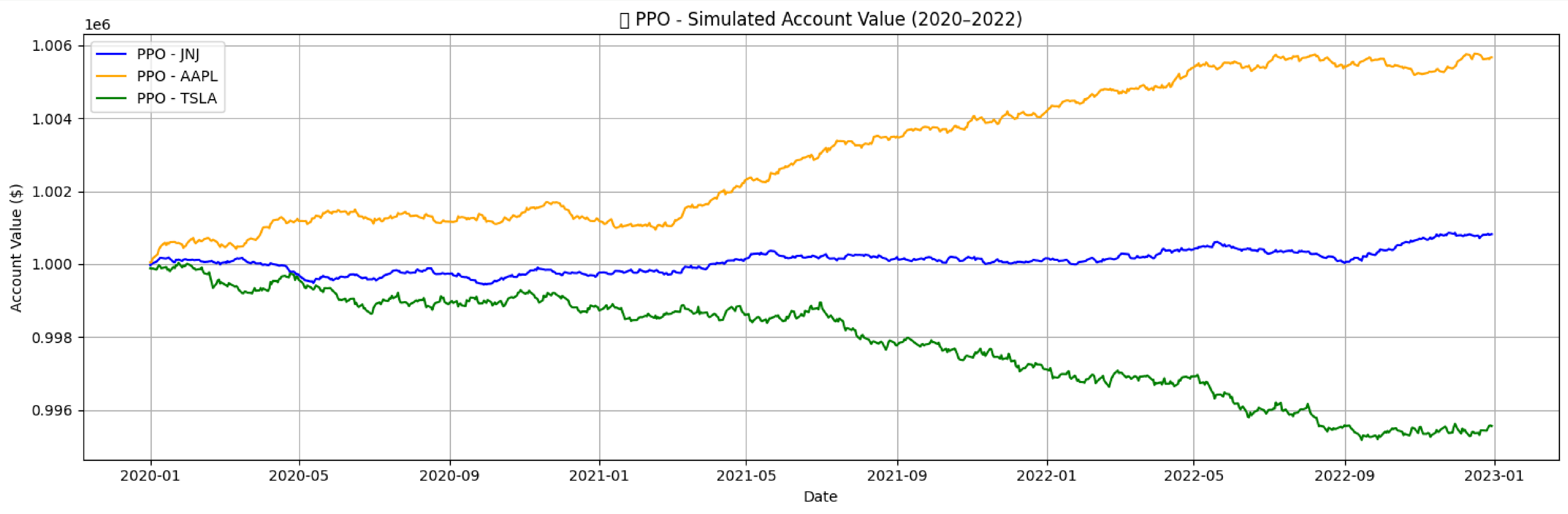
Single Asset Allocation

Goal: Training a Agent to trade over a single asset for maximum profit.

Data Used: DOW30 Constituents from yfinance(TSLA, AAPL, JNJ)

Agents Trained:
PPO (Proximal Policy Optimization)
A2C (Advantage Actor-Critic)

PPO



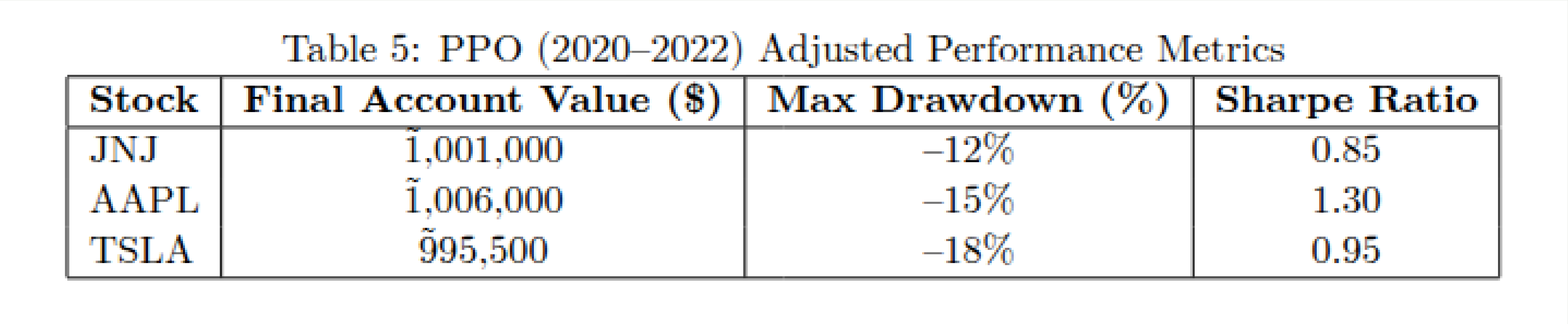
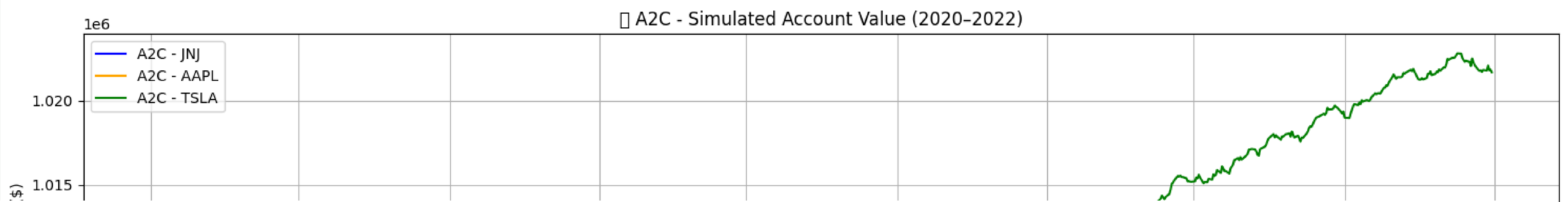


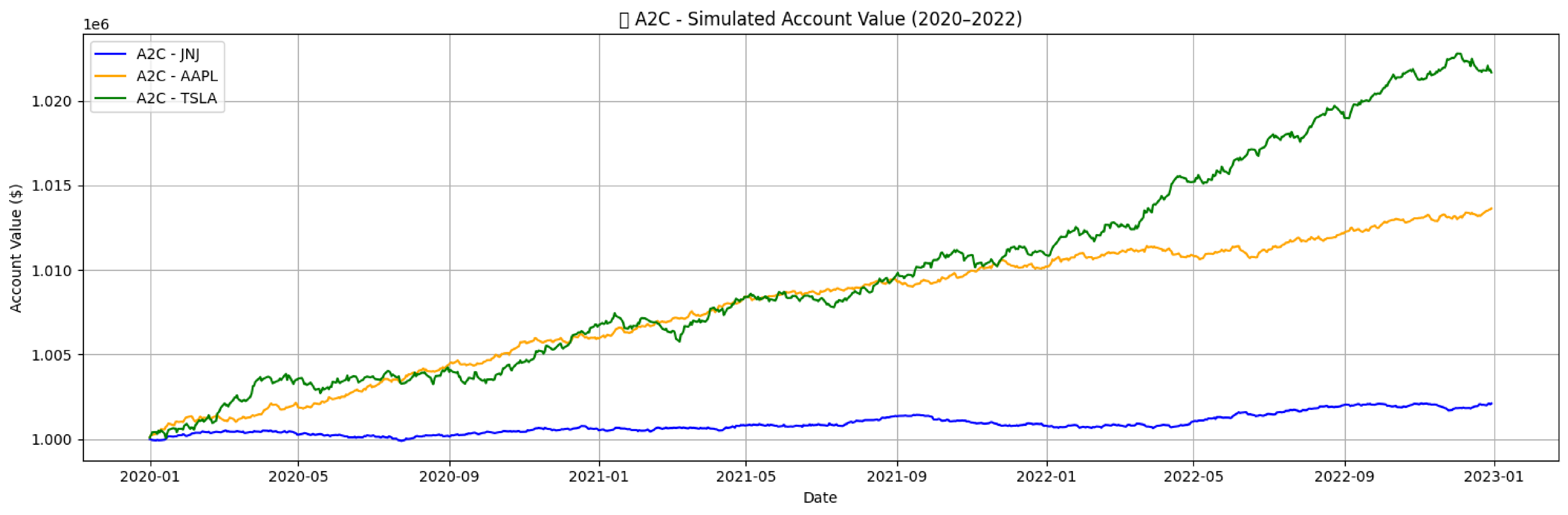
Table 5: PPO (2020–2022) Adjusted Performance Metrics

Stock	**Final Account Value ($)**	**Max Drawdown (%)**	**Sharpe Ratio**
JNJ	1̃,001,000	–12%	0.85
AAPL	1̃,006,000	–15%	1.30
TSLA	9̃95,500	–18%	0.95

PPO is more stable and better at long-term exploitation due to clipping.

A2C





AAPL	1̃,054,000	−14%	1.40
TSLA	1̃,002,000	−17%	0.80

A2C can be more reactive, which can lead to premature convergence (less exploration).
Capturing the volatility of JNJ and Apple more effectively.

Multi Asset Allocation

Goal: Allocate capital over multiple securities (Stocks, Bonds, Crypto)

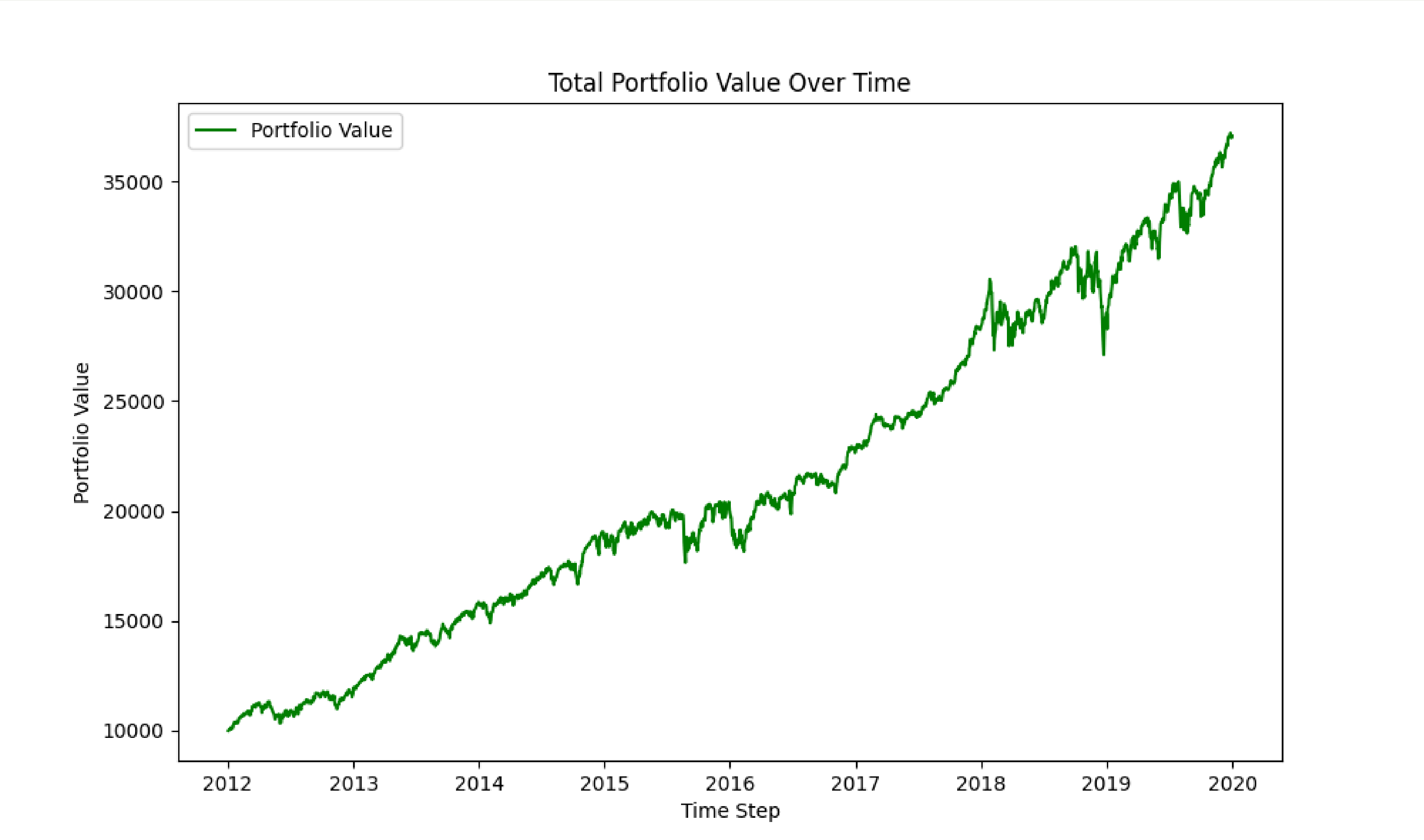
Data Used: DOW30 Constituents from yfinance

Agents Trained:

- DDPG (Deep Deterministic Policy Gradient)
- SAC (Soft Actor Critic)

Multi Asset DDPG

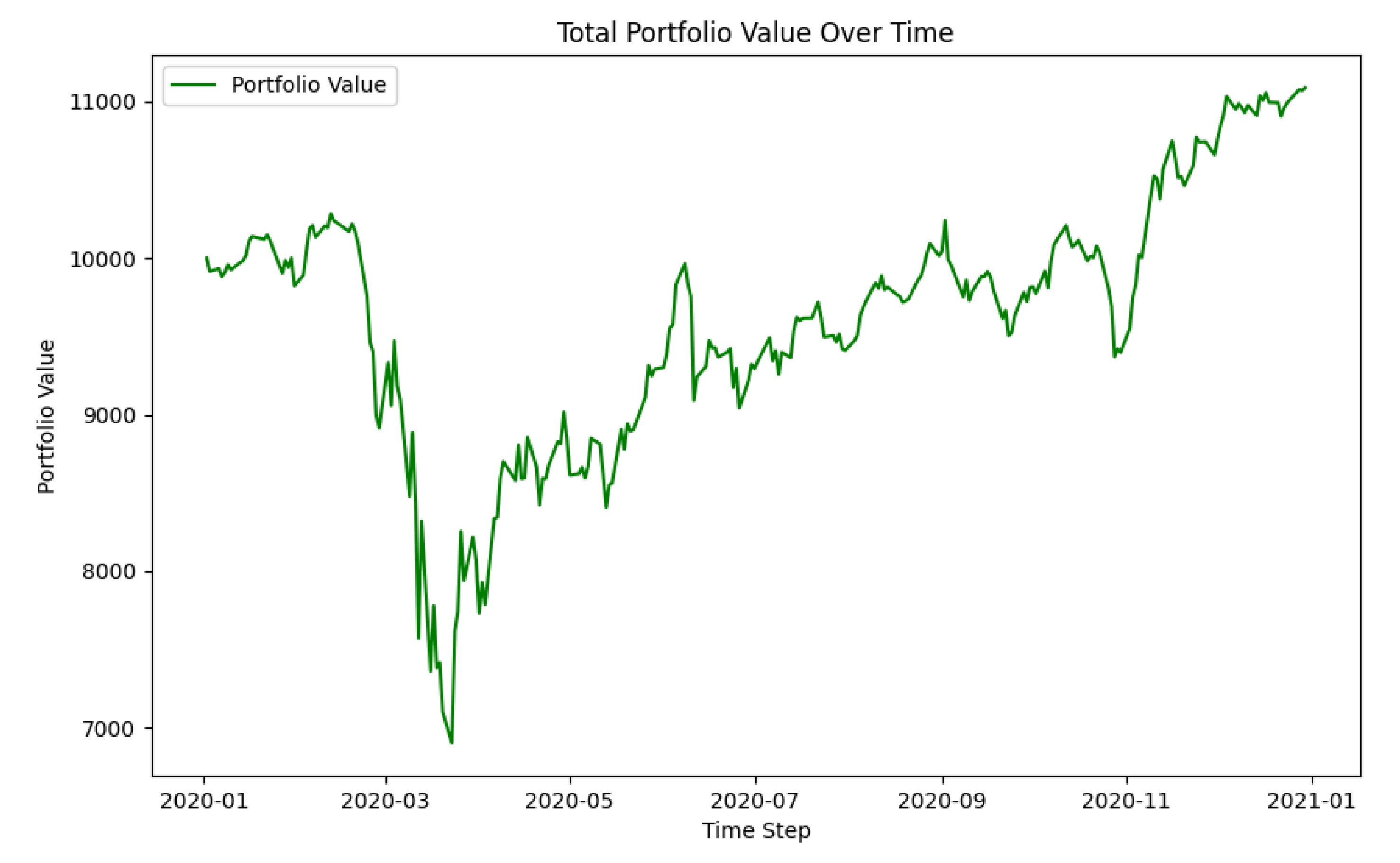
Trained over DOW 30 constituents from 2011-01-01 to 2019-12-31

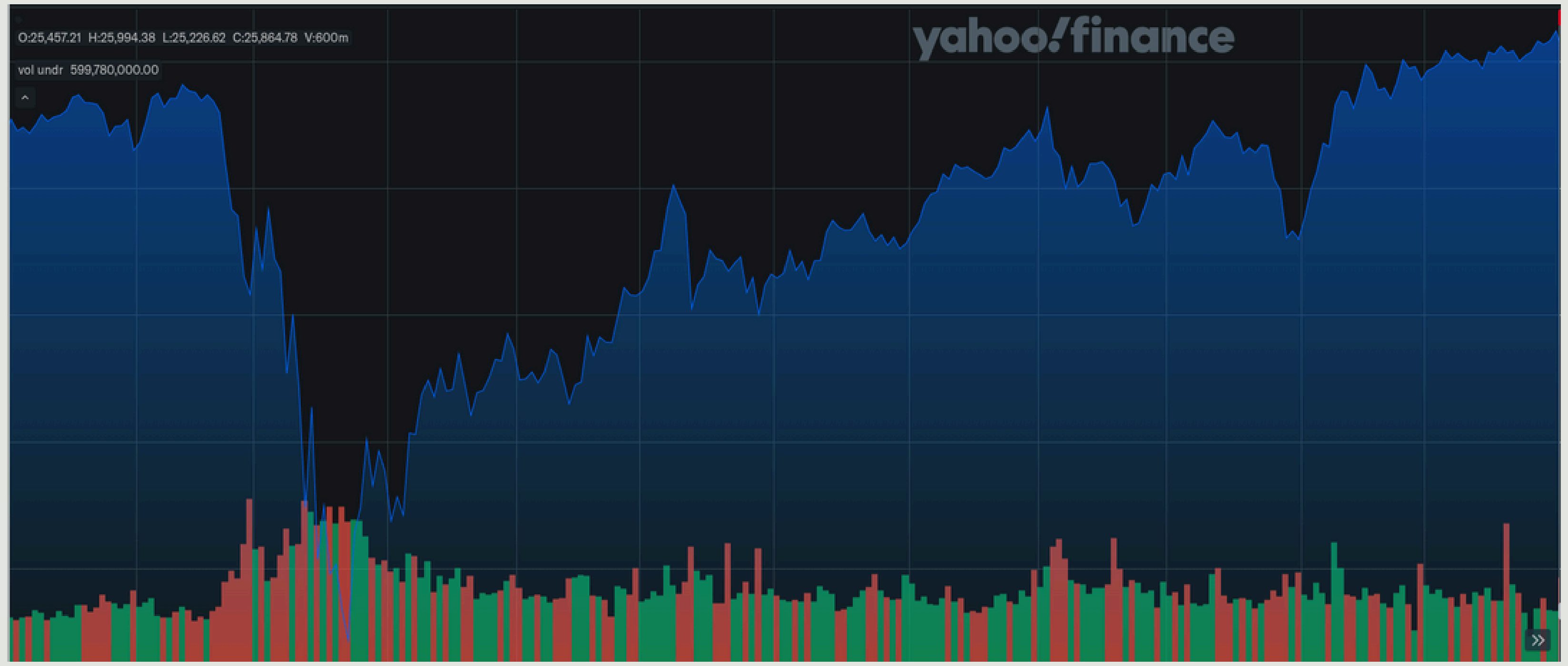


```
====================================
begin_total_asset:10000
end_total_asset:37086.09489681341
Max Drawdown: -15.39%
Sharpe Ratio:  1.3552901754452815
====================================
```

Multi Asset DDPG

Testing over DOW 30 constituents from 2020-01-01 to 2020-12-31

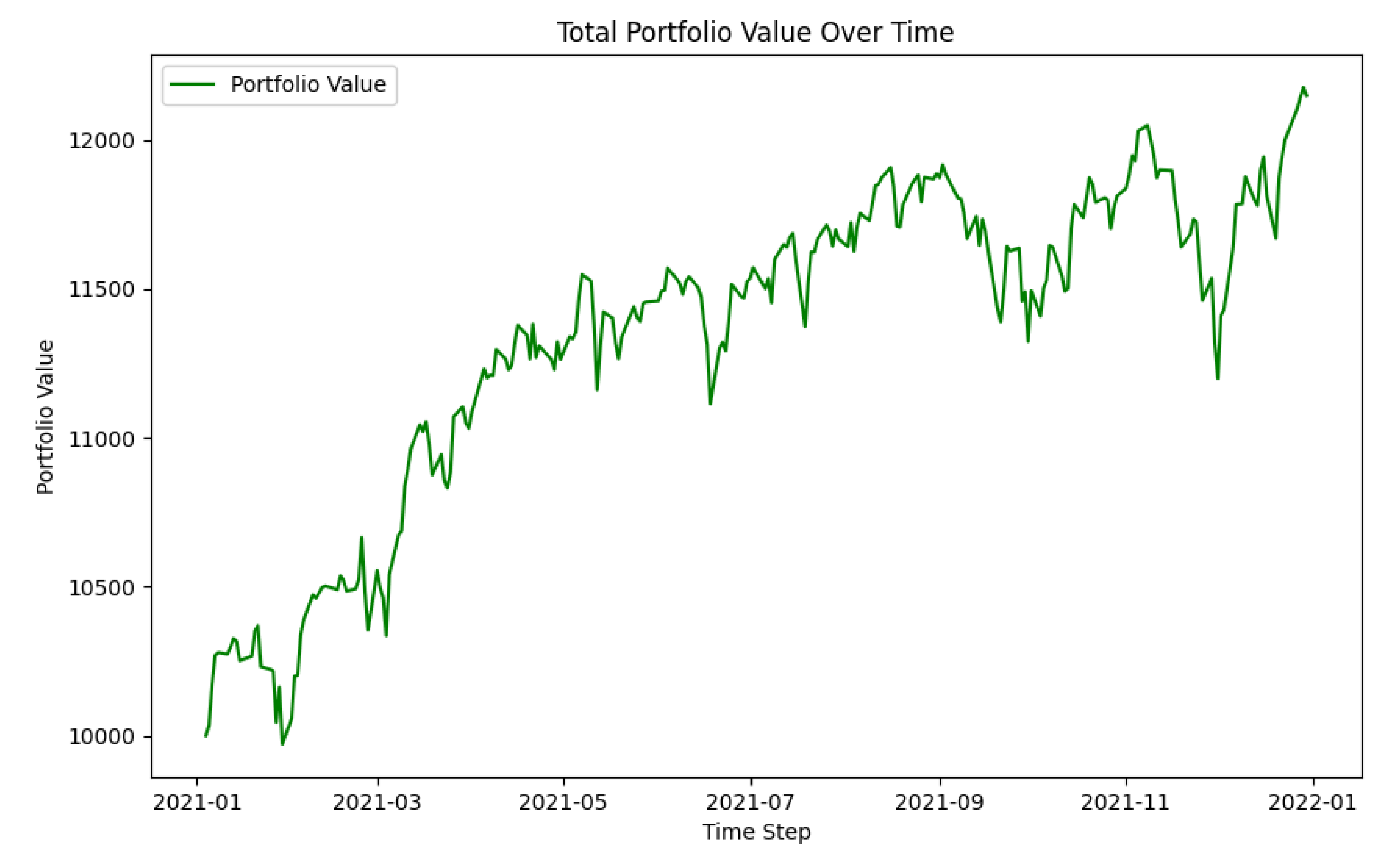


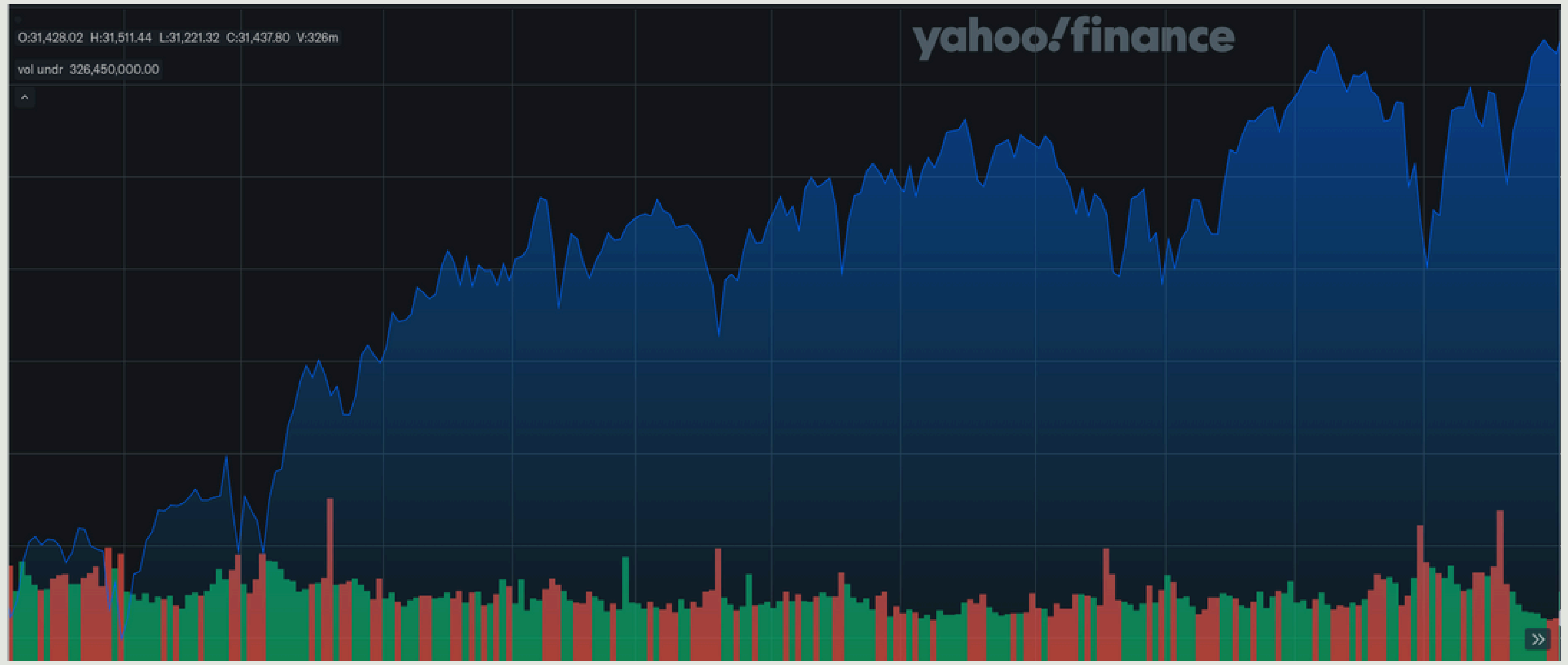


```
=================================
begin_total_asset:10000
end_total_asset:11089.507051905239
Max Drawdown: -32.89%
Sharpe Ratio:  0.4692018359035854
=================================
```

Multi Asset DDPG

Testing over DOW 30 constituents from 2021-01-01 to 2021-12-31

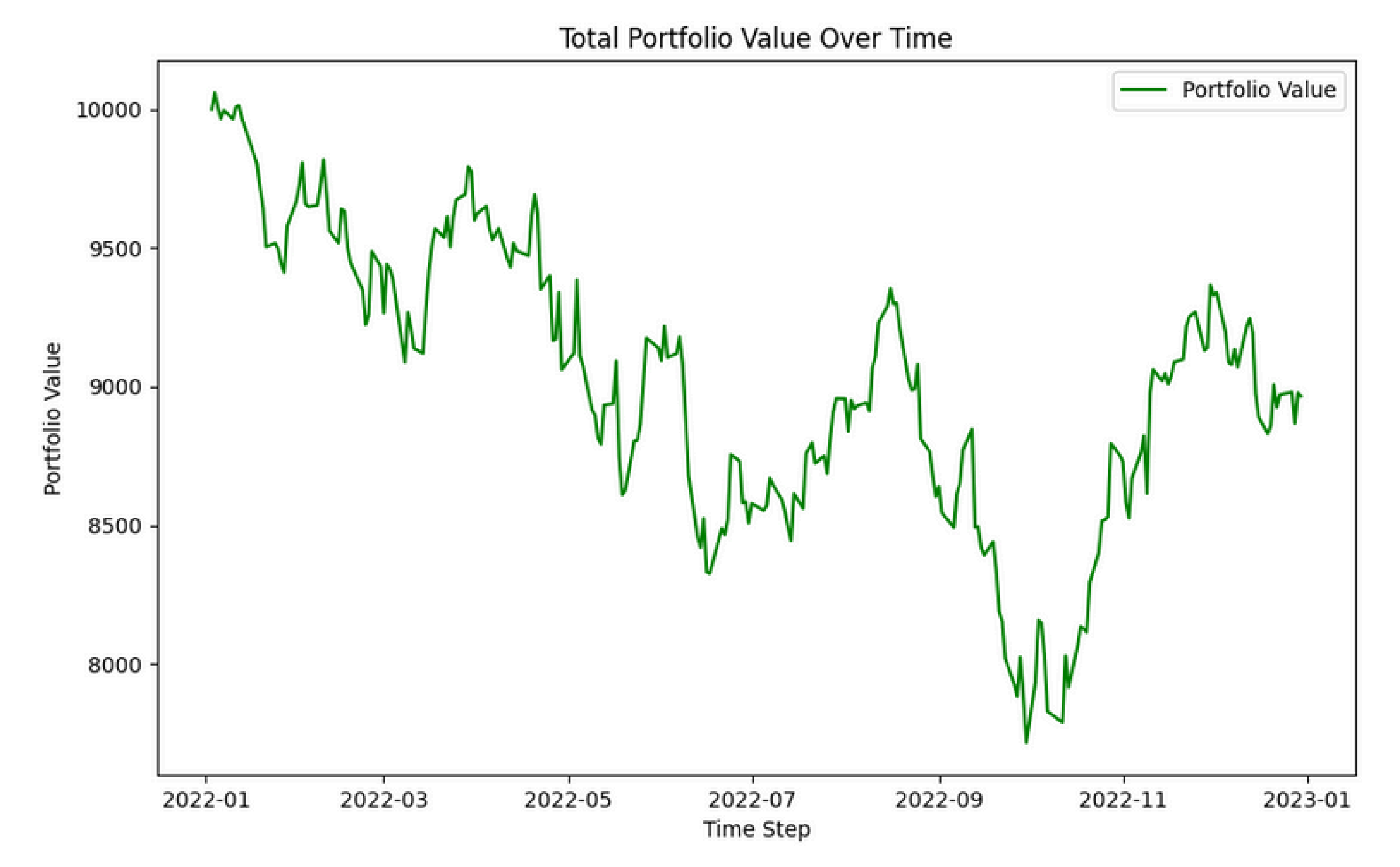


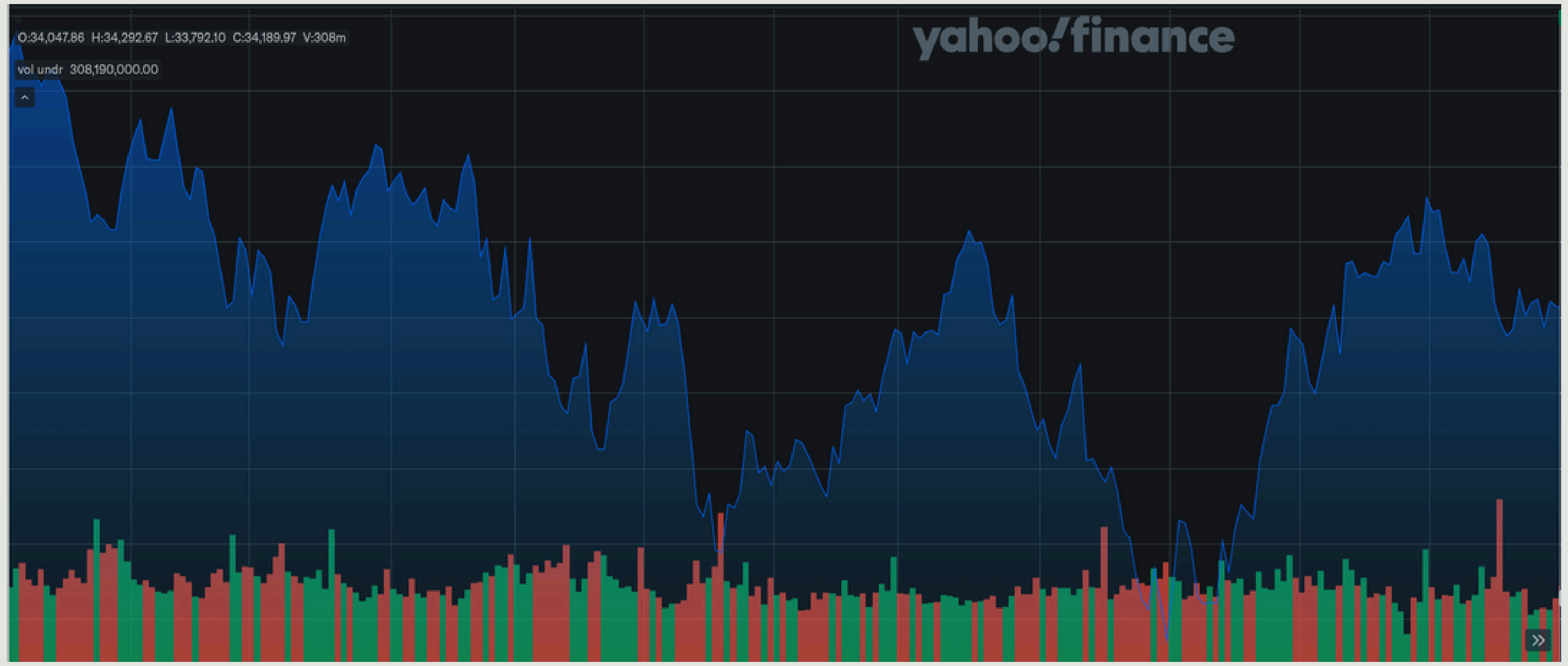


```
=================================
begin_total_asset:10000
end_total_asset:12148.115050715767
Max Drawdown: -7.05%
Sharpe Ratio:  1.660004123621777
=================================
```

Multi Asset DDPG

Testing over DOW 30 constituents from 2022-01-01 to 2022-12-31

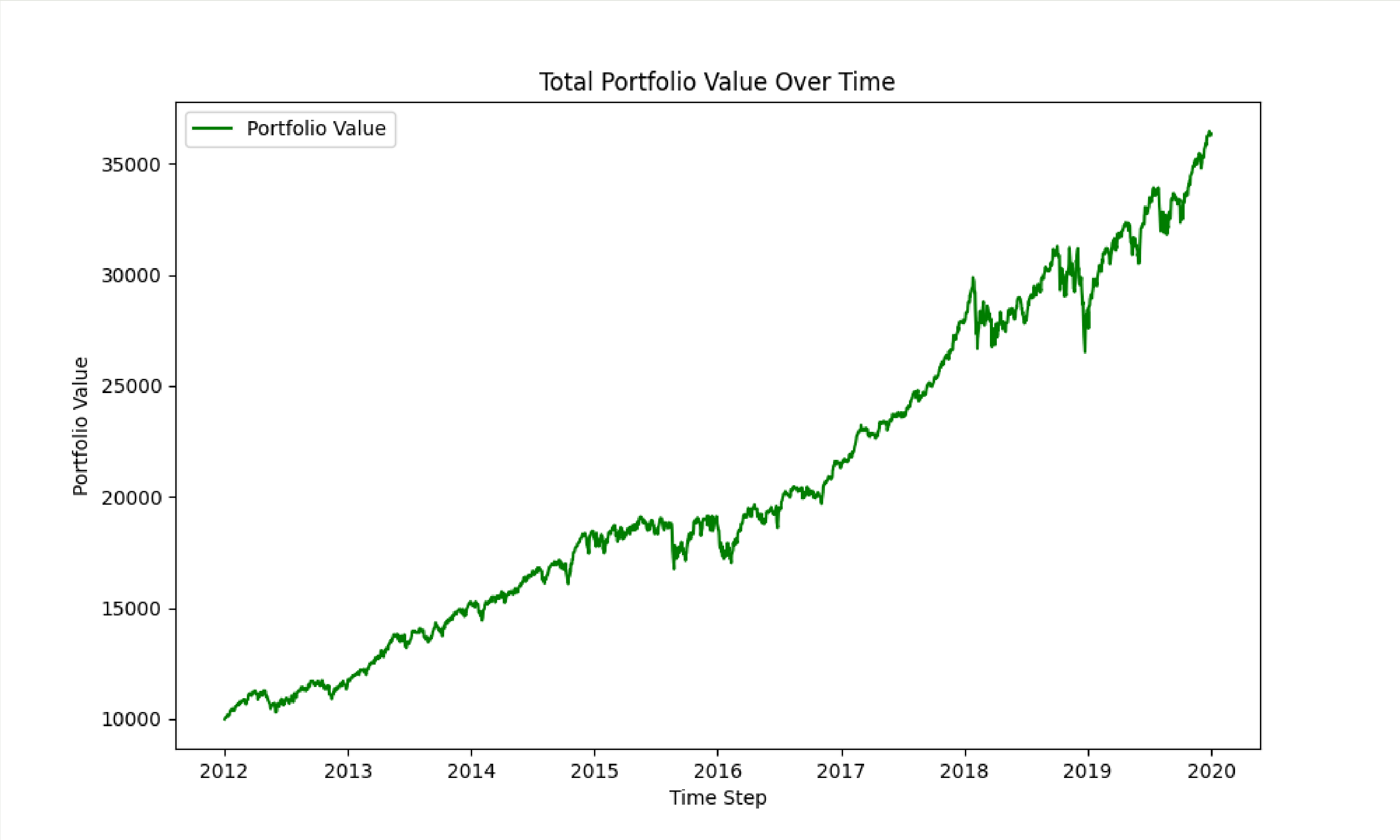




===================================
begin_total_asset:10000
end_total_asset:8965.736044101688
Max Drawdown: -23.28%
Sharpe Ratio: -0.4352328424637131
===================================

Multi Asset SAC

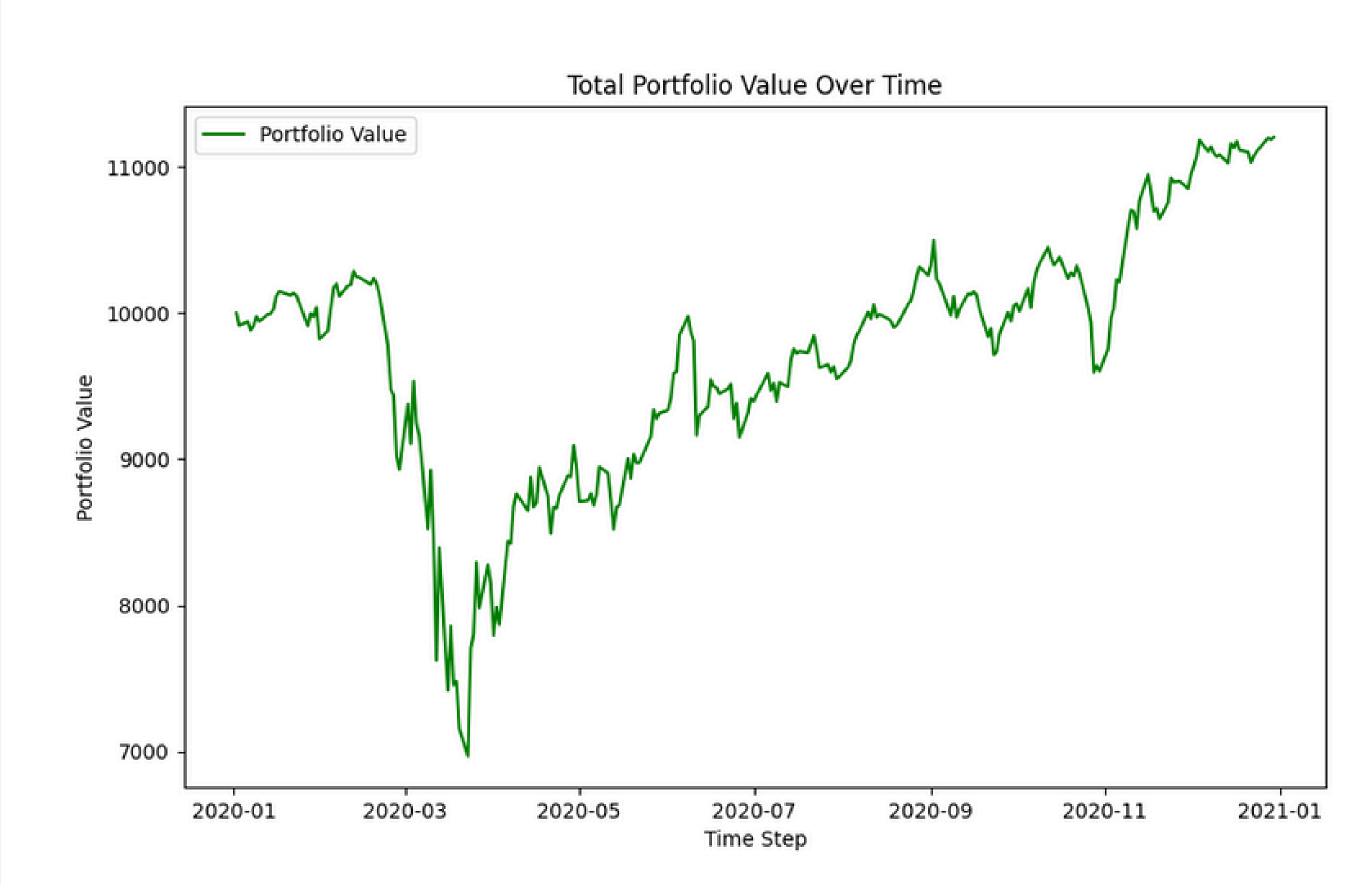
Trained over DOW 30 constituents from 2011-01-01 to 2019-12-31

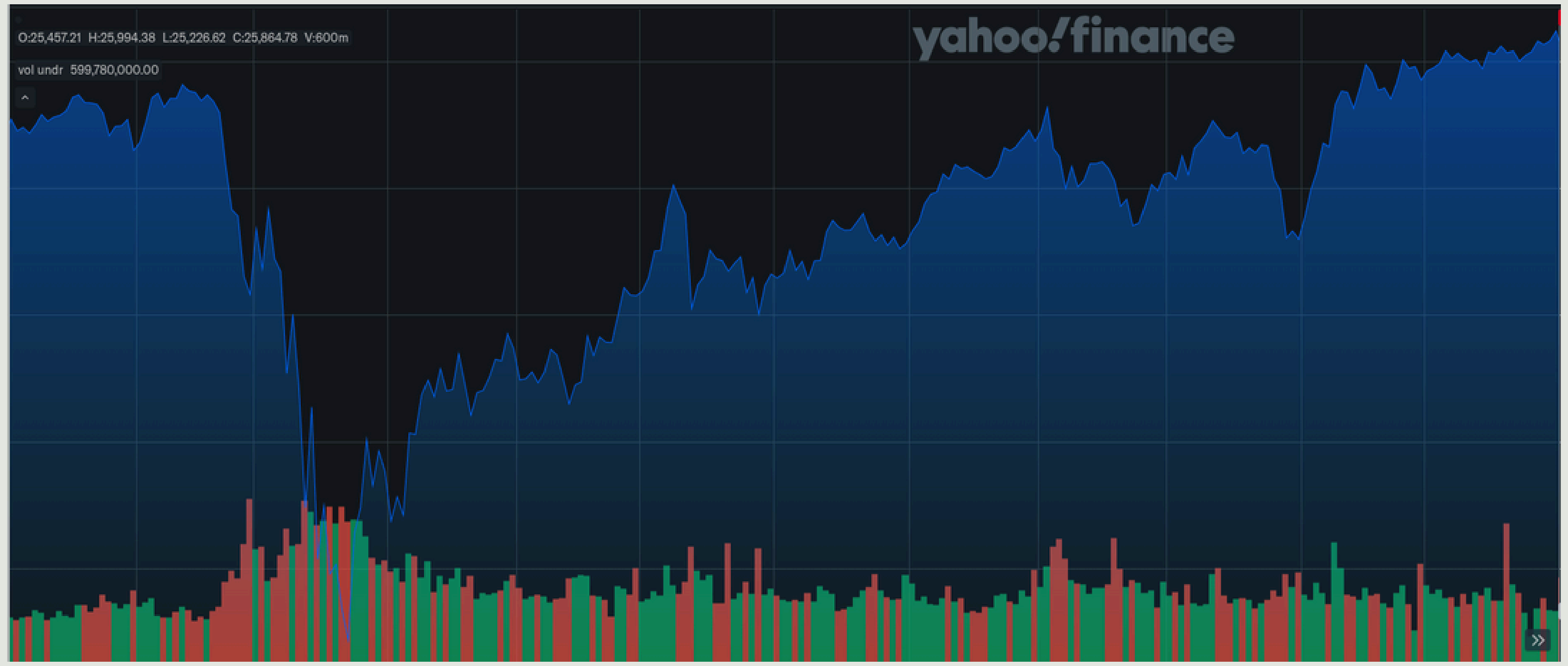


```
========================================
begin_total_asset:10000
end_total_asset:36356.92572878346
Max Drawdown: -15.23%
Sharpe Ratio:  1.3478822621384647
========================================
```

Multi Asset SAC

Testing over DOW 30 constituents from 2020-01-01 to 2020-12-31

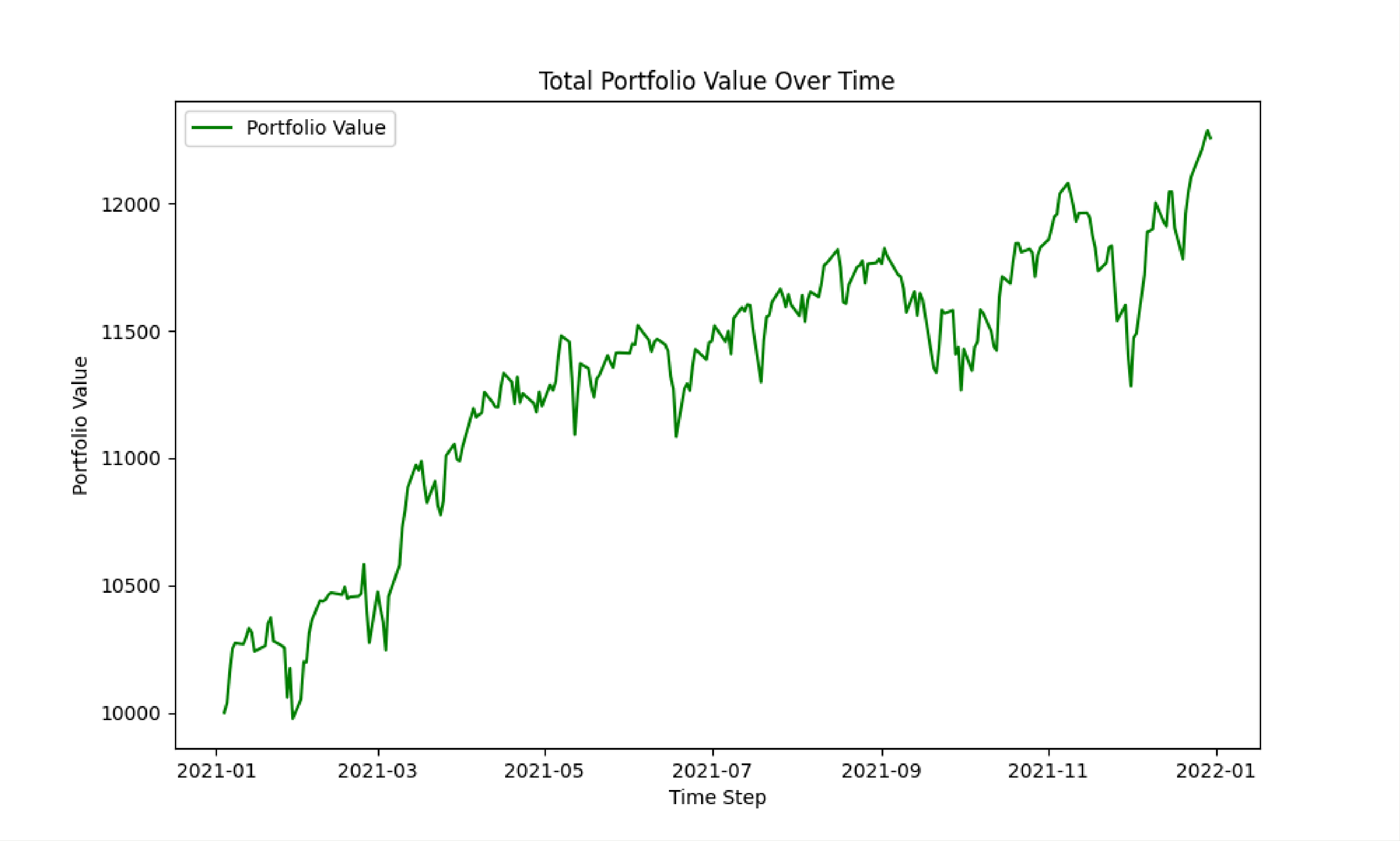


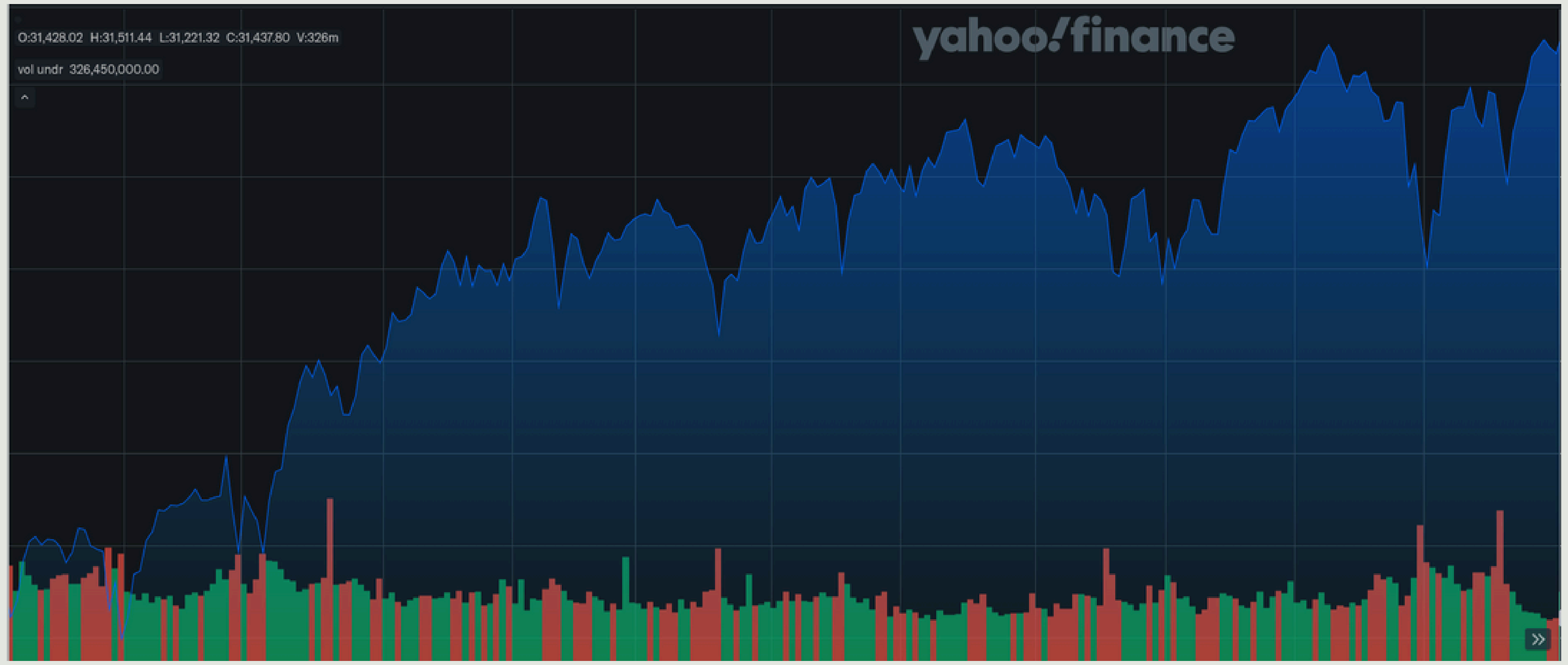


```
========================================
begin_total_asset:10000
end_total_asset:11203.525088493769
Max Drawdown: -32.28%
Sharpe Ratio:  0.4979580135476348
========================================
```

Multi Asset SAC

Testing over DOW 30 constituents from 2021-01-01 to 2021-12-31

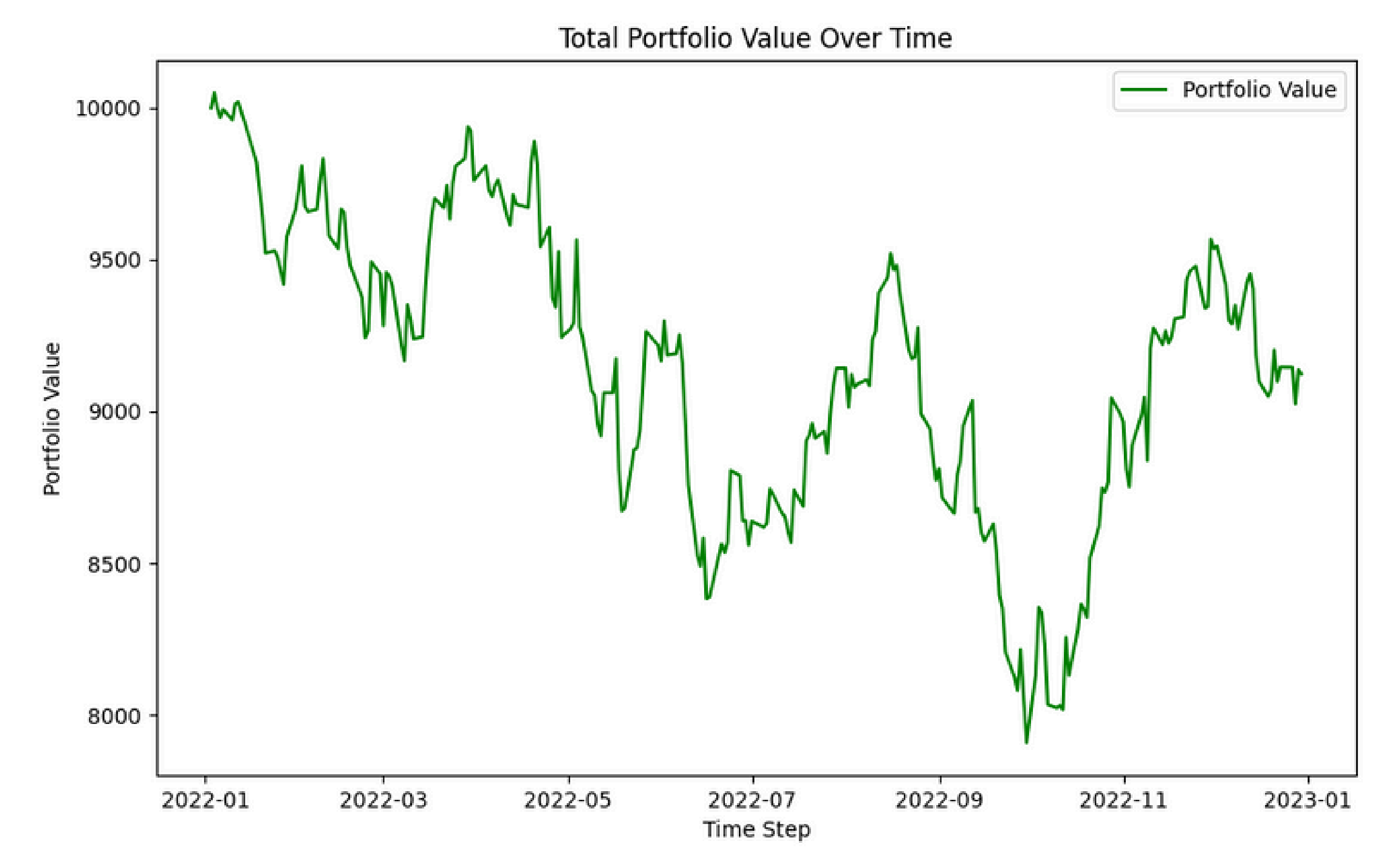


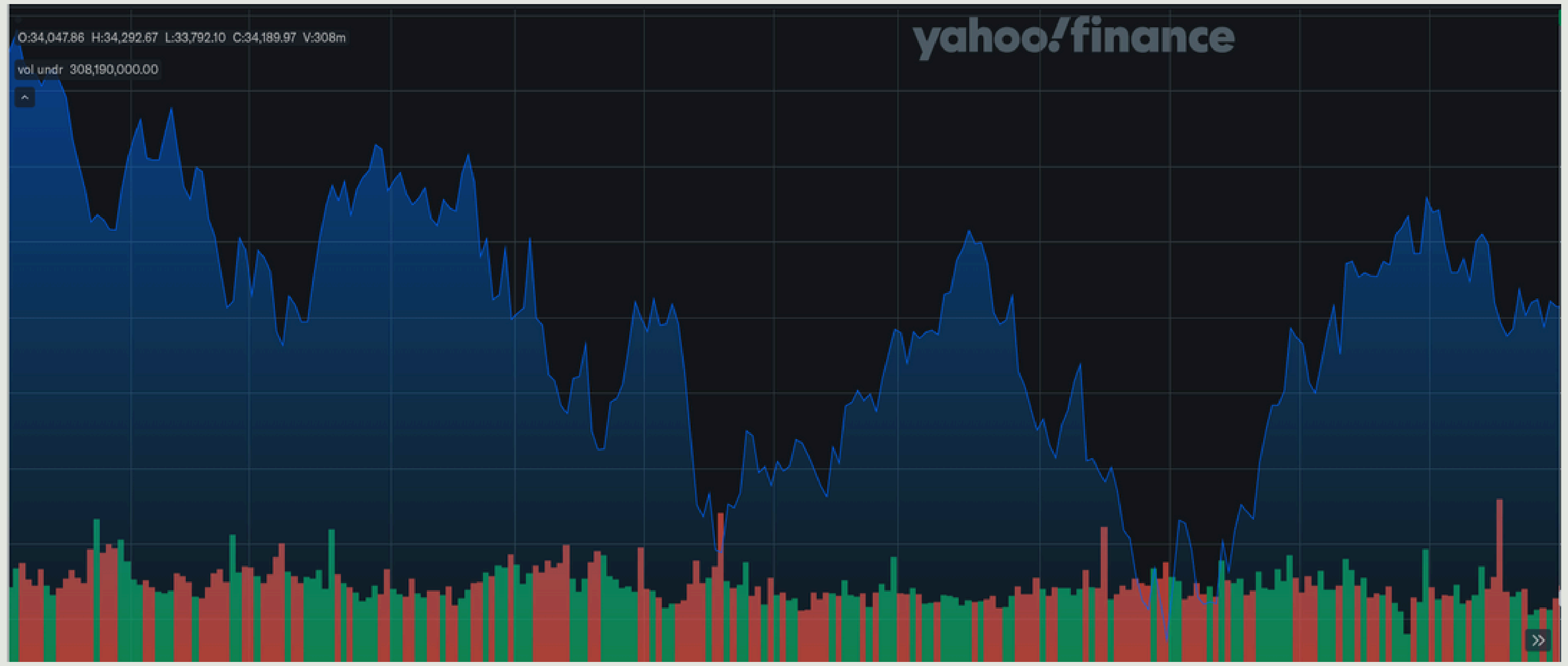


```
=================================
begin_total_asset:10000
end_total_asset:12258.057808166272
Max Drawdown: -6.60%
Sharpe Ratio:  1.7499127809902082
=================================
```

Multi Asset SAC

Testing over DOW 30 constituents from 2022-01-01 to 2022-12-31





=====================================
begin_total_asset:10000
end_total_asset:9123.364335576784
Max Drawdown: -21.31%
Sharpe Ratio: -0.3505502370952056
=====================================

Multi Asset Portfolio Comparision

Year	Model	Final Value	Sharpe Ratio	Max Drawdown
2020	DDPG	11089	0.46	-32.89%
	SAC	**11203**	**0.49**	**-32.28%**
2021	DDPG	12148	1.66	-7.05%
	SAC	**12258**	**1.74**	**-6.60%**
2022	DDPG	8965	-0.43	-23.28%
	SAC	**9123**	**-0.35**	**-21.31%**

Risk Aware Analysis

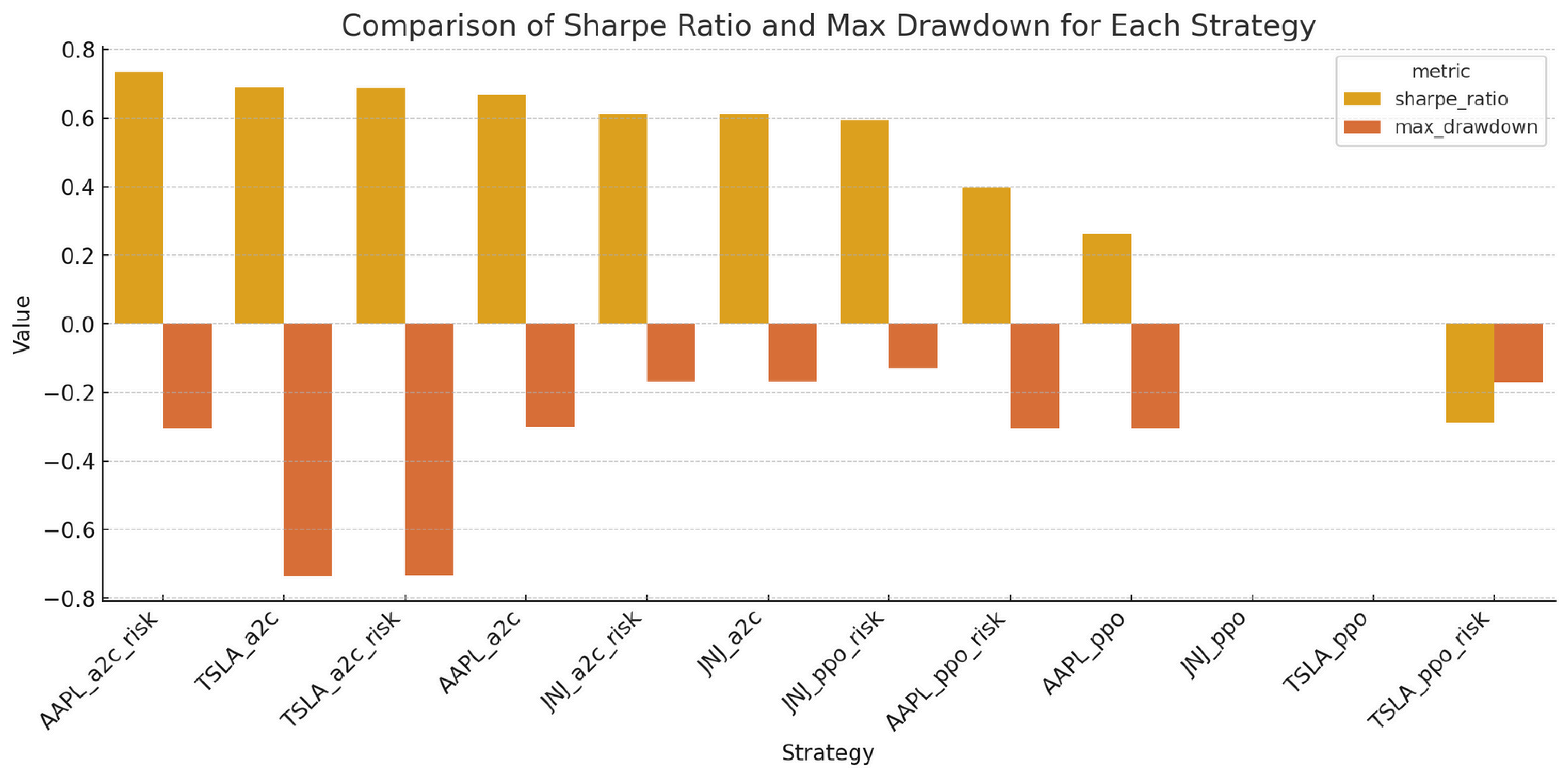
Normal Reward vs Modified Reward

Standard reward is typically the change in total asset value from one timestep to the next:
r(t)=portfolio value(t)−portfolio value(t−1)

Take the last 3 asset values and compute 2 returns.
Calculate the Sharpe ratio (risk-adjusted return)
Scale the reward (*100) to make the reward signal more significant.

Risk Aware Analysis

Single Asset - Visualizations



Risk Aware Analysis

CVaR Risk Aware Technique

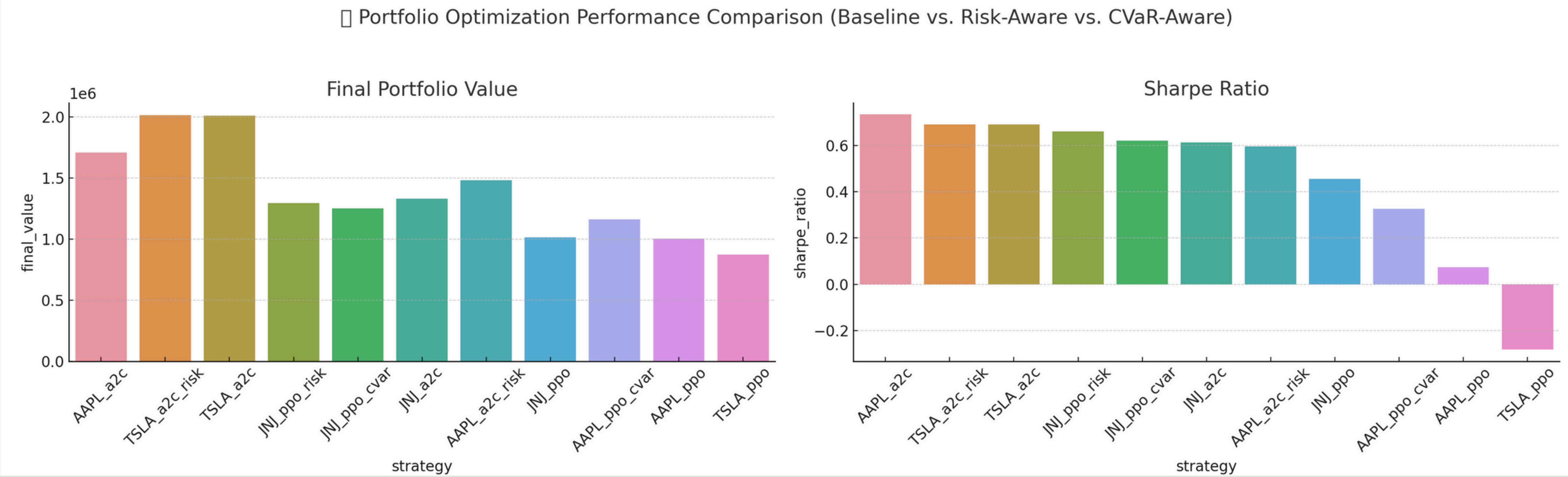
CVaR is a coherent risk measure that captures the expected loss in the worst α% of cases

ADVANTAGES :

1. Better Tail-Risk Handling
2. More Conservative Actions
3. Possibly Lower Cumulative Return

Risk Aware Analysis

Visualizations

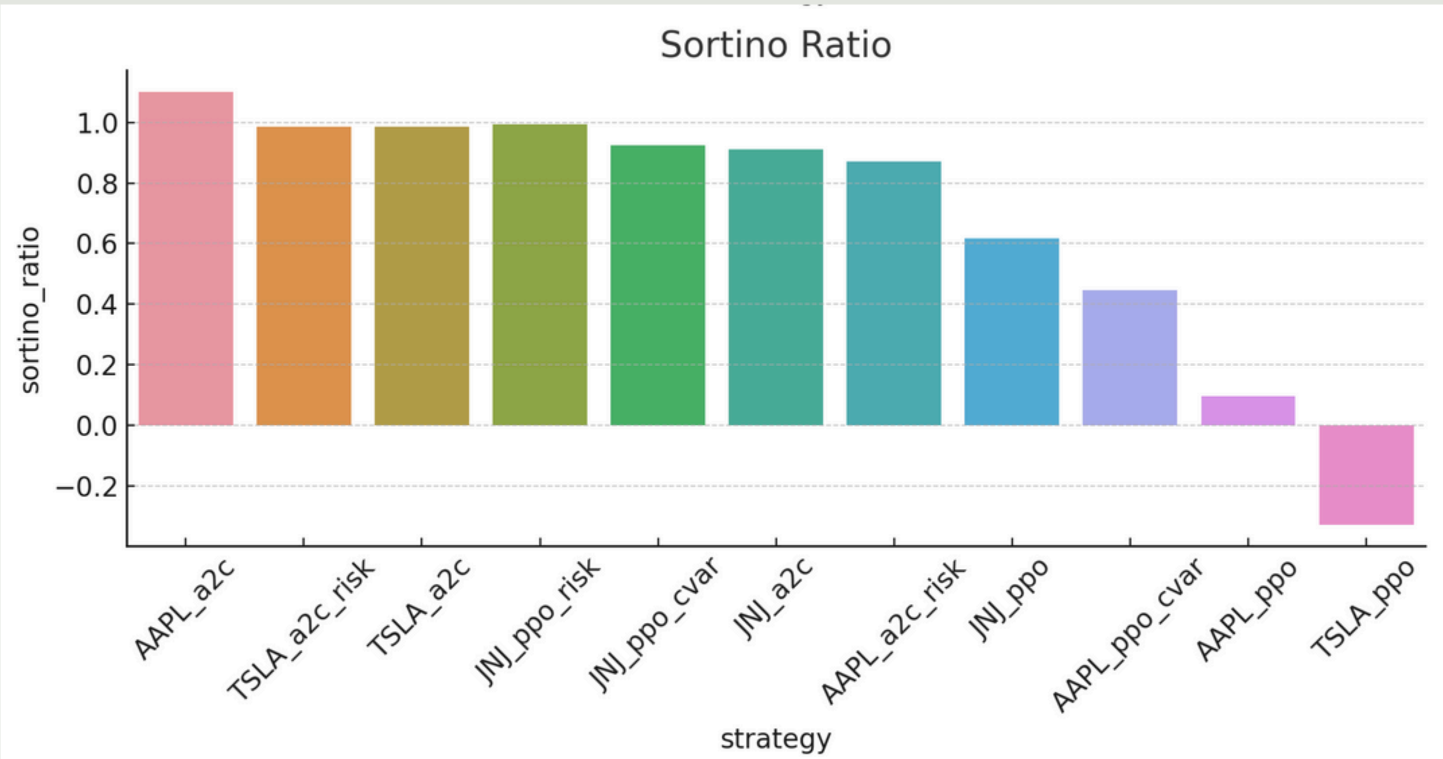

Portfolio Optimization Performance Comparison (Baseline vs. Risk-Aware vs. CVaR-Aware)

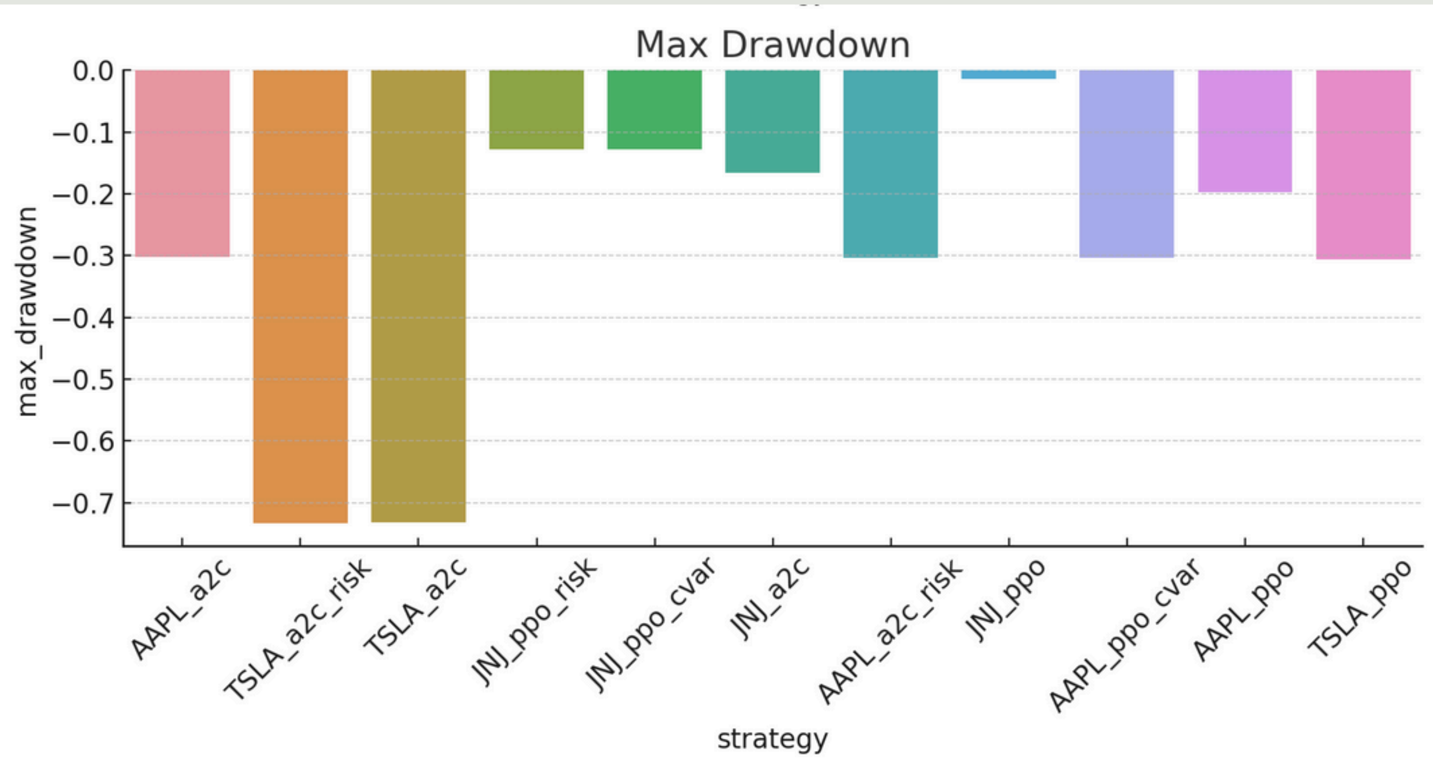
Risk-aware A2C strategies, particularly with TSLA, achieved high growth even while controlling volatility.

Baseline A2C with Apple (AAPL) yields the highest Sharpe, meaning strong return relative to volatility.

Risk Aware Analysis

Visualizations



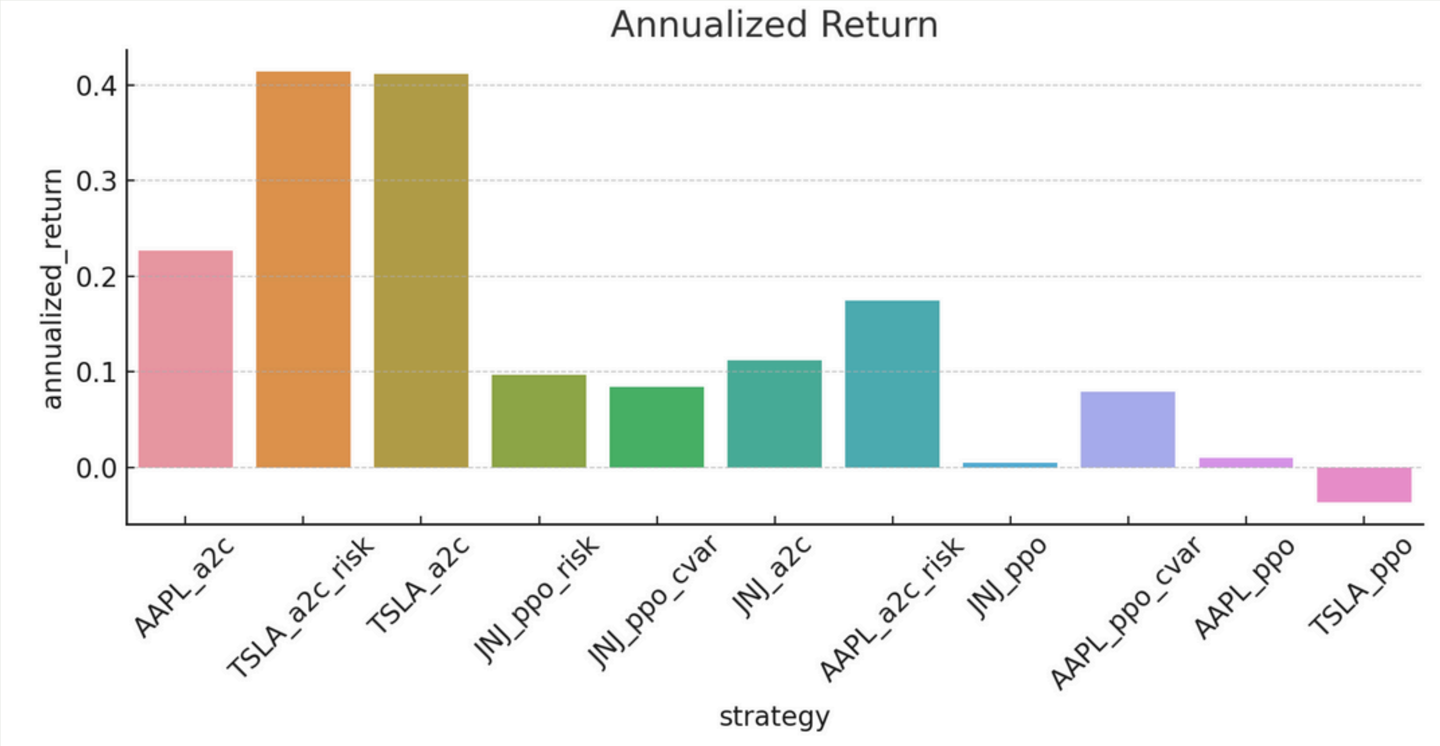


Risk-aware strategies handle downside risk well, especially for volatile stocks like TSLA.

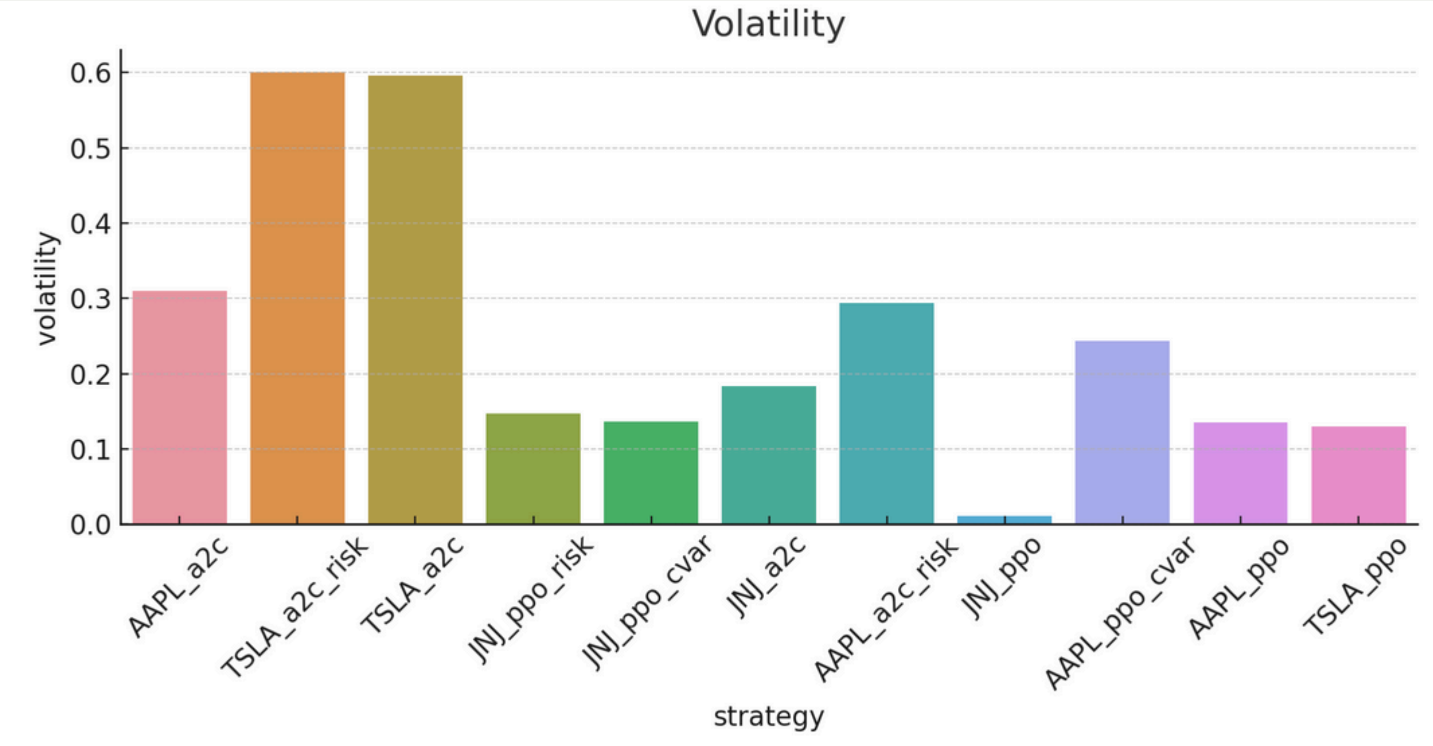
CVaR-aware methods are excellent at limiting large losses—ideal for conservative investors.

Risk Aware Analysis

Visualizations



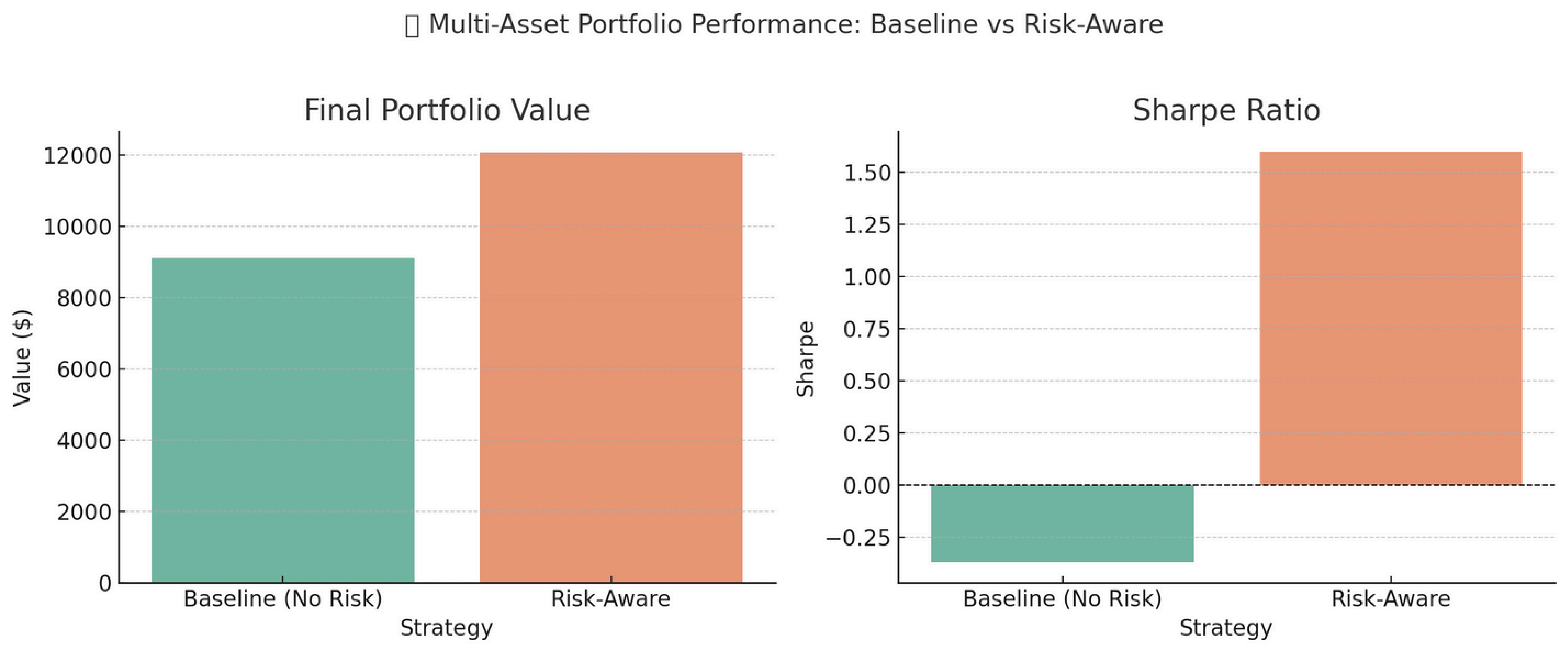
Higher return comes with higher risk; risk-aware A2C allows good trade-off.



CVaR-based models are clearly optimized for volatility control.

Risk Aware Analysis

Multi Asset - Visualizations



Baseline: Ends at $9,107, resulting in a net loss.
Risk-Aware: Grows to $12,071, marking a 20.7% return.
Risk-aware reward shaping helped the agent stabilize decisions and avoid capital erosion, even in a volatile multi-asset setting.

Baseline: Negative Sharpe Ratio (-0.37) indicates not only underperformance but also inefficient risk usage.
Risk-Aware: Sharpe Ratio of 1.60 indicates excellent return-to-volatility efficiency.

Single Asset Trading LLM Agent

Use gpt-4o-mini with access to python code interpreter loaded with a pandas dataframe of historical data

SYSTEM PROMPT:

```
You are a financial trading assistant operating on a daily time scale. Your primary objective is to analyze daily stock data and make informed trading decisions.

# Available Tools
Code Interpreter: Utilize this tool to execute Python code for data analysis. It's especially useful for processing historical stock data using libraries like pandas.

You can find the historical data in a variable `data`, pandas is already imported as `pd`, print the final answer you wanna see
data.head(2) gives:

{apple_data.head(2).to_markdown()}
Date is the index
# Daily Workflow
For each trading day:
- Data Analysis:
Review the provided date, open, and close prices for the day.
Employ the Code Interpreter to analyze historical data, identifying trends, patterns, or anomalies that could influence trading decisions.
- Decision Making:
Based on your analysis and news assessment, determine the number of shares to trade.
Your decision should be an integer within the range [-k, k], where:
Positive values indicate buying shares.
Negative values indicate selling shares.
Zero indicates no action.
- Output:
Clearly state your decision in the format: DECISION: [number]. For example, DECISION: 3 or DECISION: -2.
Provide a brief rationale for your decision, referencing specific data points or news items that influenced your choice.
Constraints & Guidelines
Trading Limits: Ensure that the number of shares traded does not exceed the absolute value of k in either direction.
Consistency: Maintain a consistent decision-making process, documenting your rationale for each action to facilitate future reviews and audits.
Tool Usage: Effectively utilize the available tools to support your analysis. If a tool fails or provides insufficient information, note this in your rationale.
Ethical Considerations: Avoid making decisions based on unverified or speculative information. Base your actions on concrete data analysis
Error Handling: If you encounter errors or inconsistencies in data or tool outputs, document these issues and proceed with caution, making decisions based on the most reliable information available.
```

Single Asset Trading LLM Agent

Use gpt-4o-mini with access to python code interpreter loaded with a pandas dataframe of historical data

Example Query:

User:
Date: 2025-05-05
Close: 199.34500122070312
High: 204.10000610351562
Low: 198.2100067138672
Open: 203.125
Volume: 47305475.0
Balance: 6508.538879394531
Shares Held: 24
Cost Basis: 206.42519067834925
Net Worth: 11292.818908691406

Generate the action to be taken using the analysis done using the tools provided in the format DECISION: [number]. For example, DECISION: 3 or DECISION: -2.

You can output the decisions from -24, ..., 32

-24 meaning sell all shares, and 32 meaning buy all shares possible

Always use the code tool before answering, use pandas to analyze the data variable already present, do not create a new one. Try to do some new analyses too.

Single Asset Trading LLM Agent

Example Output:

LLM Runs Python Code

```
# Analyzing the provided data for the date 2025-05-05

# Extracting the relevant data for analysis
current_close = 199.34500122070312
current_open = 203.125
current_high = 204.10000610351562
current_low = 198.2100067138672
shares_held = 24
cost_basis = 206.42519067834925
balance = 6508.538879394531

# Calculating the profit/loss per share
profit_loss_per_share = current_close - cost_basis

# Total profit/loss
total_profit_loss = profit_loss_per_share * shares_held

# Current net worth
net_worth = balance + (shares_held * current_close)

# Analyzing the price movement
price_change = current_close - current_open
price_change_percentage = (price_change / current_open) * 100

# Analyzing volume
average_volume = data['volume'].mean()  # Average volume over the historical data
current_volume = 47305475.0

# Determine if the current volume is above or below average
volume_analysis = "above average" if current_volume > average_volume else "below average"

# Preparing the analysis results
analysis_results = {
    'profit_loss_per_share': profit_loss_per_share,
    'total_profit_loss': total_profit_loss,
    'net_worth': net_worth,
    'price_change': price_change,
    'price_change_percentage': price_change_percentage,
    'volume_analysis': volume_analysis
}

analysis_results
```

Single Asset Trading LLM Agent

Tool Output:

```
 stdout:
{'profit_loss_per_share': -7.080189457646128, 'total_profit_loss': -169.92454698350707, 'net_worth': 11292.818908691406, 'price_change':
-3.779998779296875, 'price_change_percentage': -1.8609224759615386, 'volume_analysis': 'below average'}

    --
    stderr:
```

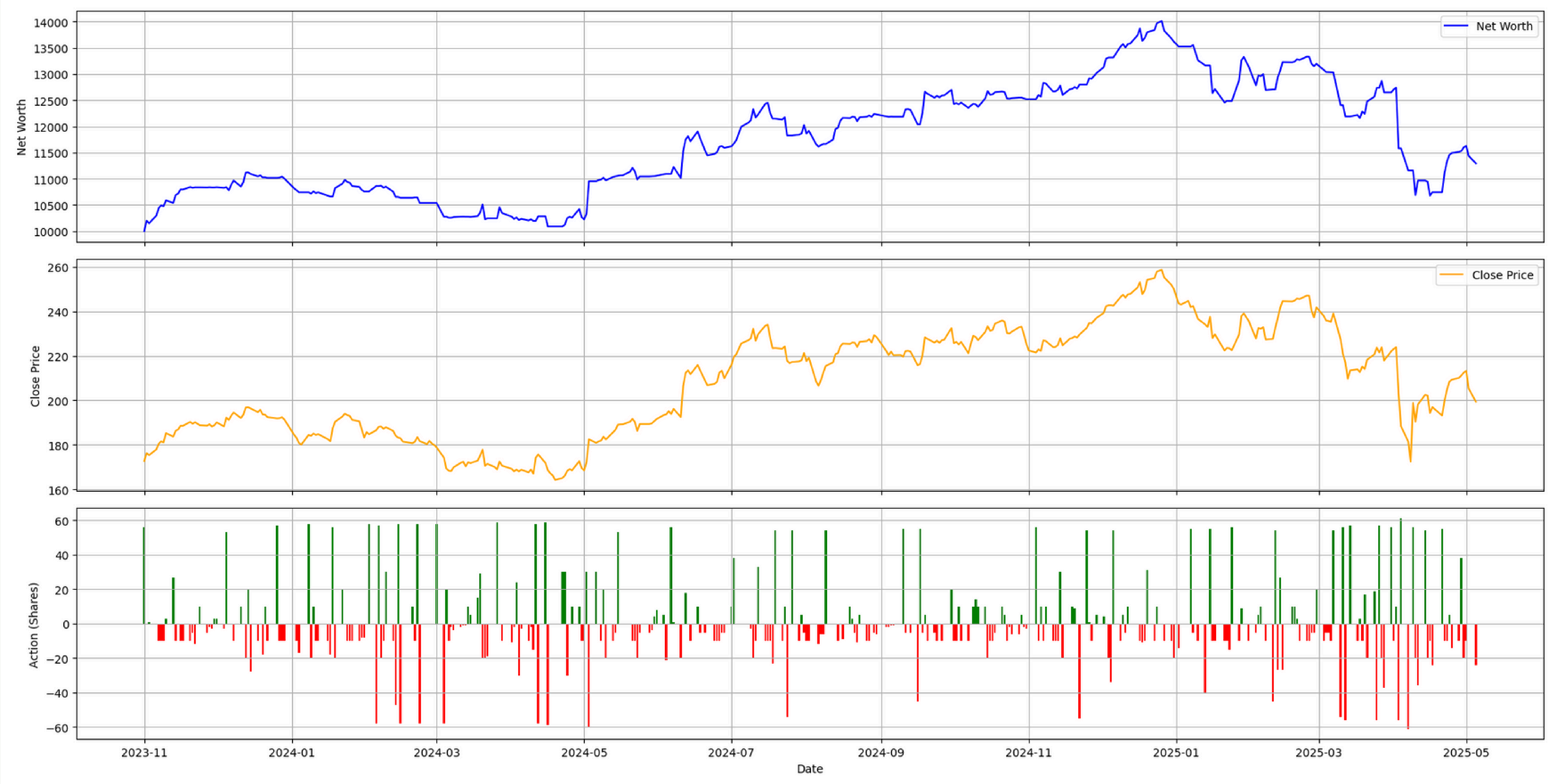
LLM OUTPUT:

```
Based on the analysis conducted for the date 2025-05-05, here are the key findings:

- **Profit/Loss per Share**: -$7.08
- **Total Profit/Loss**: -$169.92
- **Net Worth**: $11,292.82
- **Price Change**: -$3.78
- **Price Change Percentage**: -1.86%
- **Volume Analysis**: Below average

### Decision Rationale:
1. The current close price is significantly lower than the cost basis, resulting in a loss per share. This indicates that holding the shares is currently
unprofitable.
2. The price has decreased by 1.86% from the opening price, suggesting a bearish trend for the day.
3. The trading volume is below average, which may indicate a lack of strong market interest or momentum.

Given these factors, it would be prudent to minimize losses by selling shares. Since I hold 24 shares, I will decide to sell all of them.

### Final Decision:
DECISION: -24
```

Single Asset Trading LLM Agent

Experimental Setup:

- Testing from November 2023 to 5 May 2025 (GPT-4o-mini’s cutoff date is October 2023).
- Trading on a daily basis.
- AAPL stocks.
- Chat history is reset everyday.

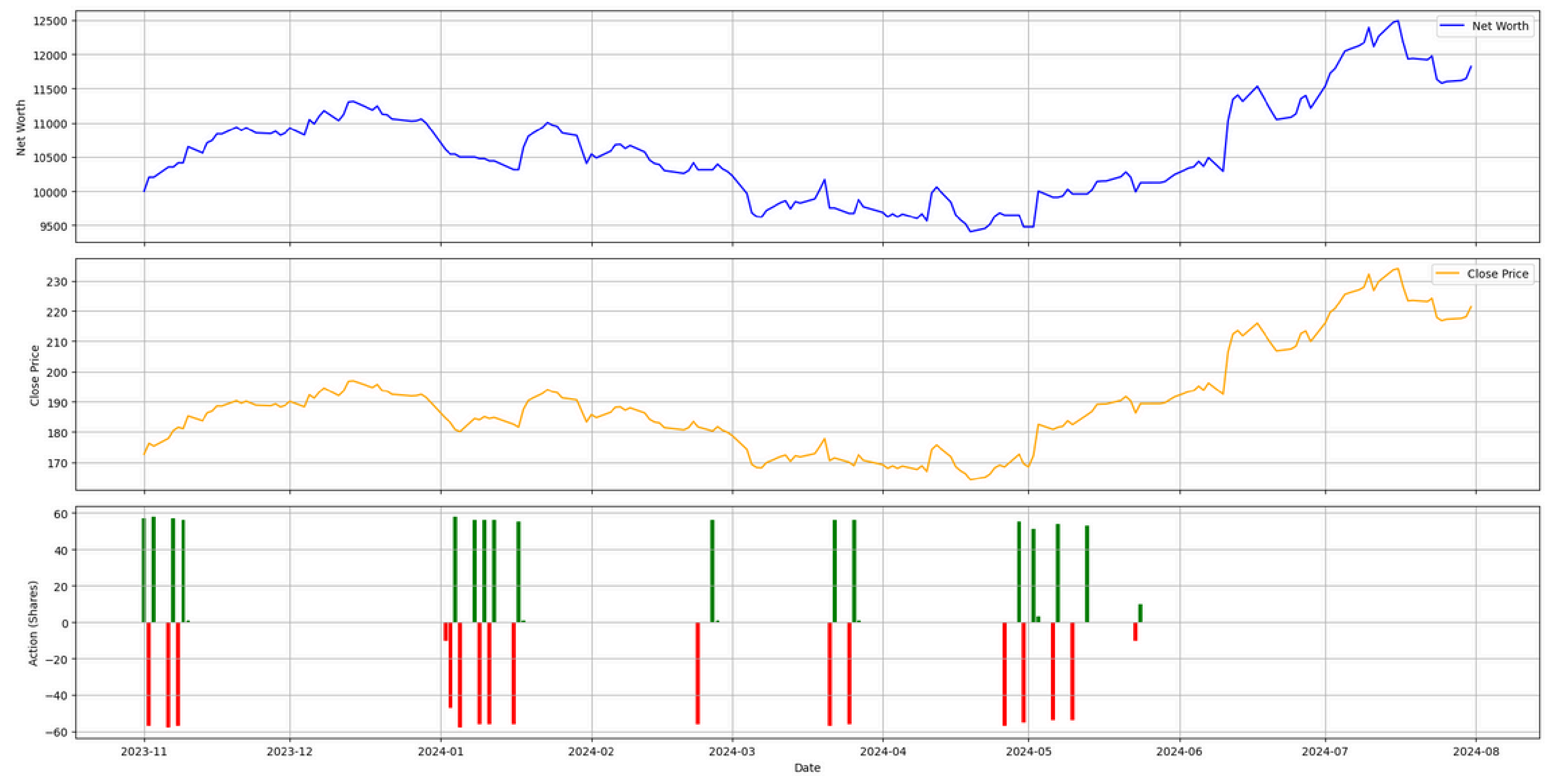


=== Performance Summary ===
Final Net Worth: 11292.82
Total Return (%): 12.93
Sharpe Ratio (LLM): 0.54
Sharpe Ratio (Stock): 0.47
Max Drawdown (%): -23.81
Buy Count: 112
Sell Count: 205

Single Asset Trading LLM Agent

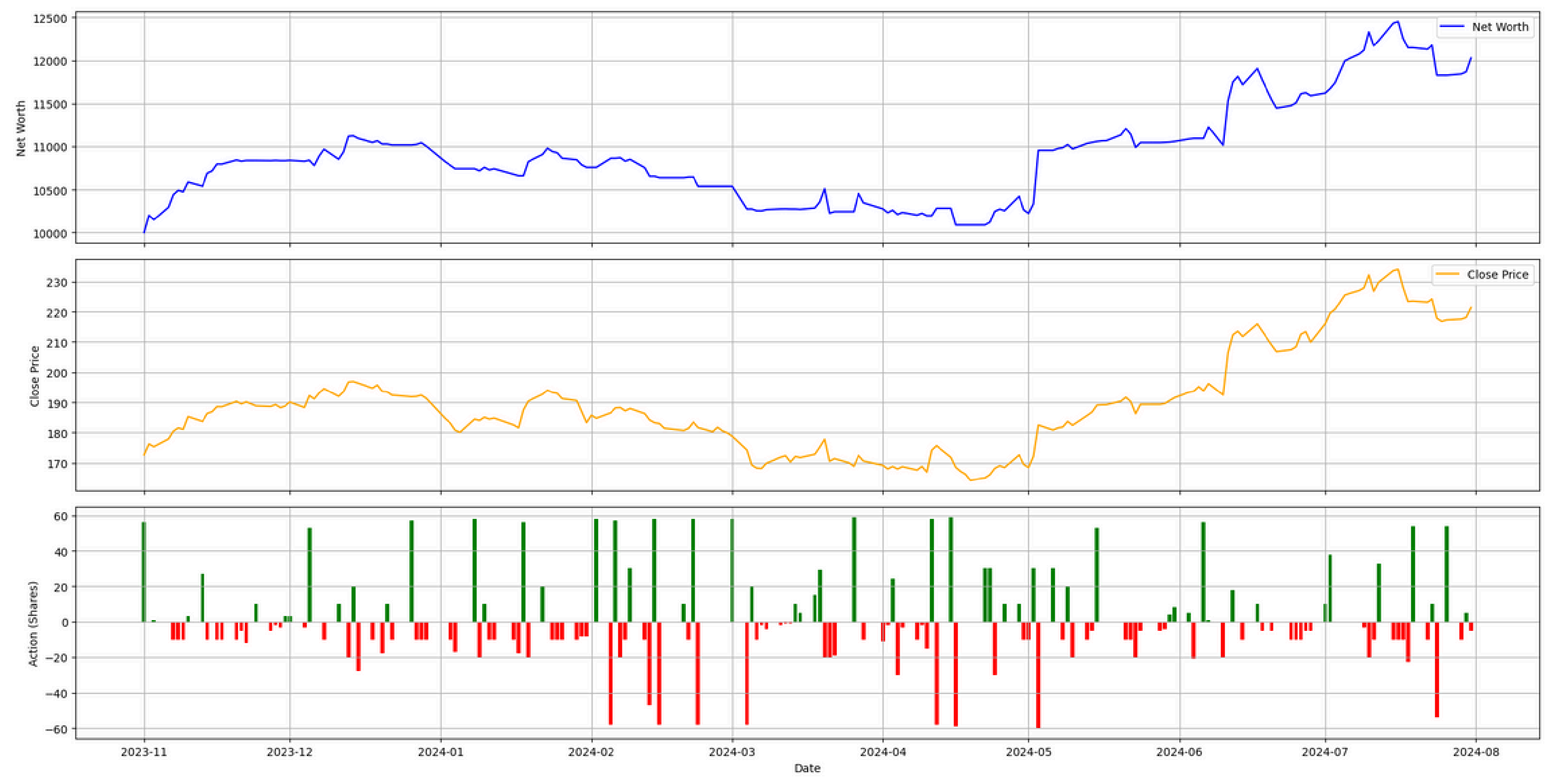
Comparing Persistent Chat History vs Non-Persistent Chat History

With Chat History



=== Performance Summary ===
Final Net Worth: 11819.16
Total Return (%): 18.19
Sharpe Ratio (LLM): 1.15
Sharpe Ratio (Stock): 1.59
Max Drawdown (%): -16.86
Buy Count: 22
Sell Count: 17

Without Chat History

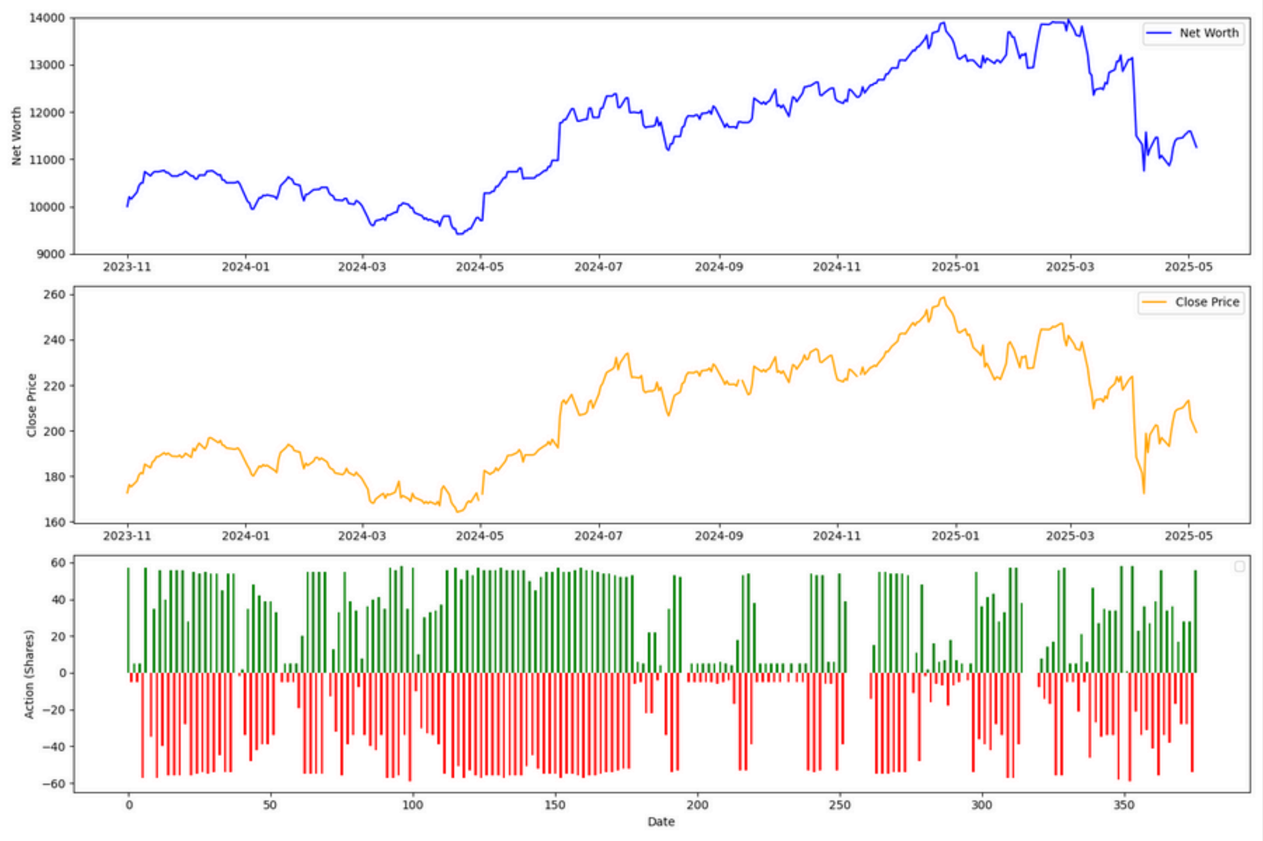


=== Performance Summary ===
Final Net Worth: 12026.31
Total Return (%): 20.26
Sharpe Ratio (LLM): 1.71
Sharpe Ratio (Stock): 1.59
Max Drawdown (%): -9.29
Buy Count: 54
Sell Count: 99

Single Asset Trading LLM Agent

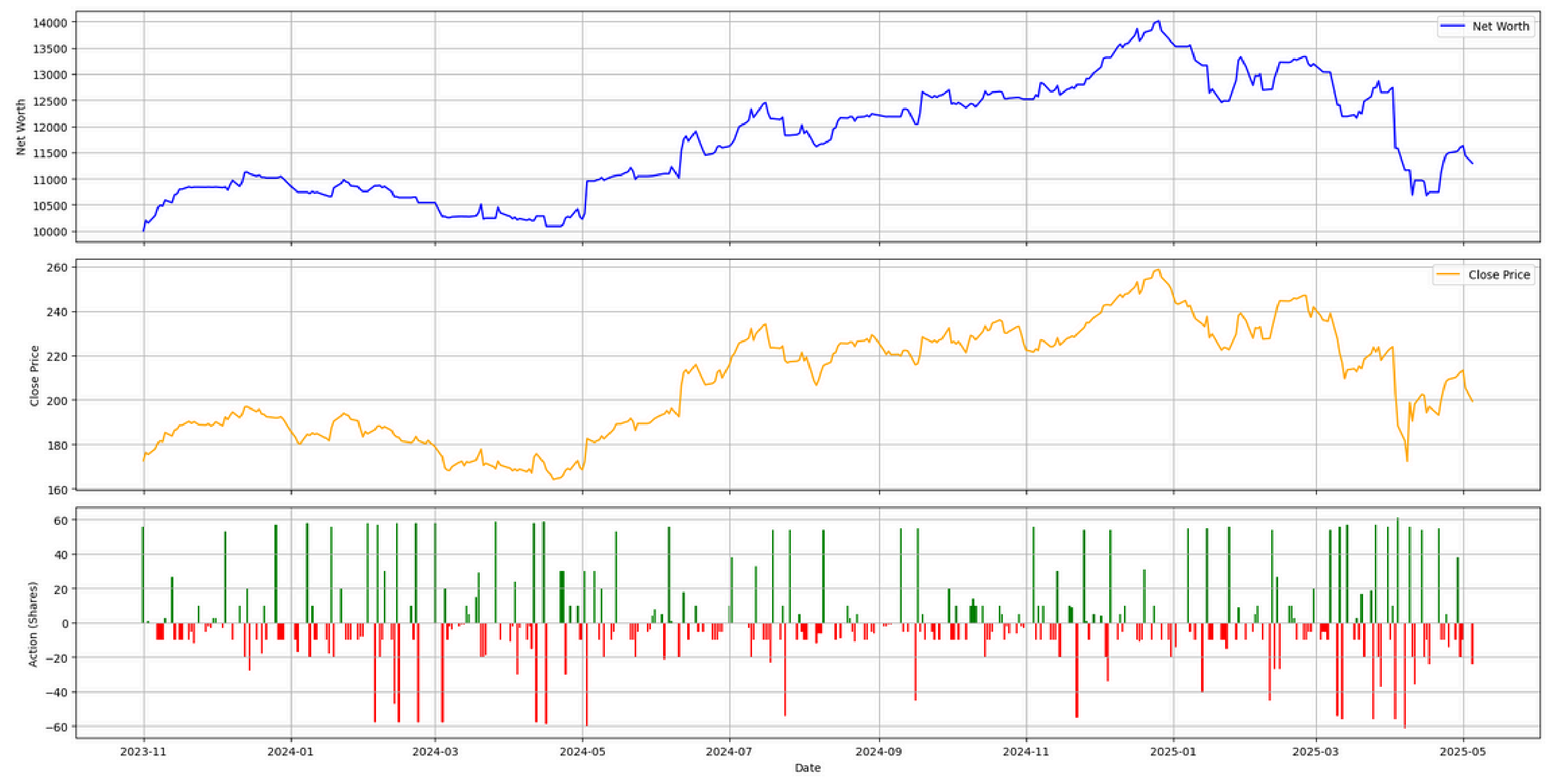
Comparing Persistent Chat History vs Non-Persistent Chat History

A2C
Trained from 2011 January to 2023 October



=== Performance Summary ===
Final Net Worth: 11254.48
Total Return (%): 12.54
Sharpe Ratio (LLM): 0.468
Sharpe Ratio (Stock): 0.47
Max Drawdown (%): -22.92

LLM Trader



=== Performance Summary ===
Final Net Worth: 11292.82
Total Return (%): 12.93
Sharpe Ratio (LLM): 0.54
Sharpe Ratio (Stock): 0.47
Max Drawdown (%): -23.81

Challenges

- YahooDownloader API Rate Limits
  - The free Yahoo Finance API used via YahooDownloader has limited request capacity
- Evolving FinRL Ecosystem
  - The FinRL library is actively developing, and some components are still experimental or under-documented
- Lack of Real-Time Market Sentiment/News Data
  - High-quality, real-time sentiment/news data is not readily available in free sources, limiting the realism of sentiment-driven portfolio adjustments.
- High Dimensionality in Multi-Asset Optimization
  - Optimizing a portfolio across multiple assets introduces a large and complex hyperparameter space, increasing computational cost and training difficulty.
- Extensive Use of LLMs Involves Cost
  - Leveraging Large Language Models (LLMs) for market interpretation and report generation incurs significant API or infrastructure costs, especially at scale.
- Lack of GRPO Support in Multi Turn Tool Calling setups for low resource devices

Multi Agent Trading

Goal: Meta-controller that dynamically combines the actions of different agents based on market volatility to produce superior portfolio returns with reduced risk.

Dataset: AAPL, MSFT, GOOG, AMZN, META, NVDA, TSLA, AMD

- Data Period: Jan 1, 2022 – Dec 31, 2023

Agents: PPO and DDPG

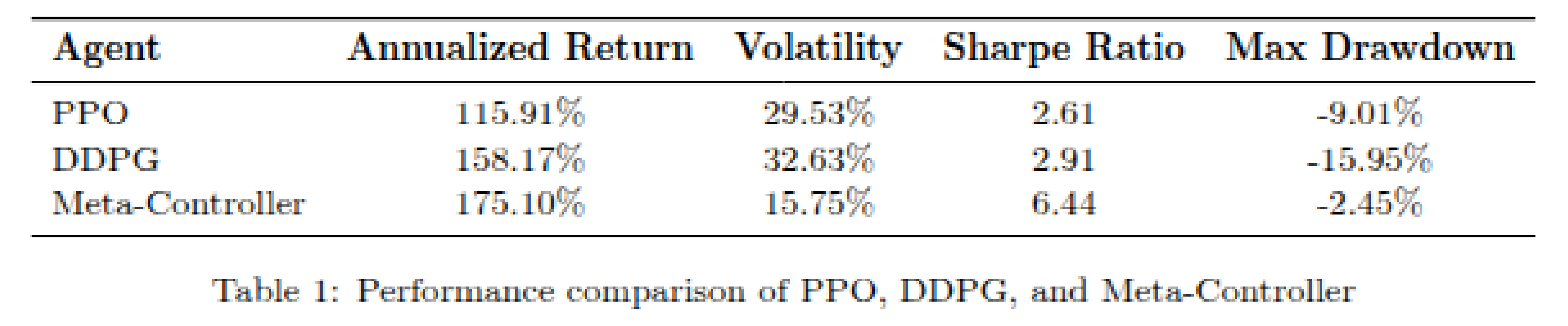
- Observation Window: 5 days
- Action: Portfolio weights (normalized)

The agent’s goal is to optimize portfolio value over time based on asset price returns.

Meta-Controller

Used Proximal Policy Optimization (PPO) for short-term volatility handling and Deep Deterministic Policy Gradient (DDPG) for long-term stable returns.
Model learns to dynamically adapt to market volatility by switching between exploration-oriented PPO and precision-focused DDPG.

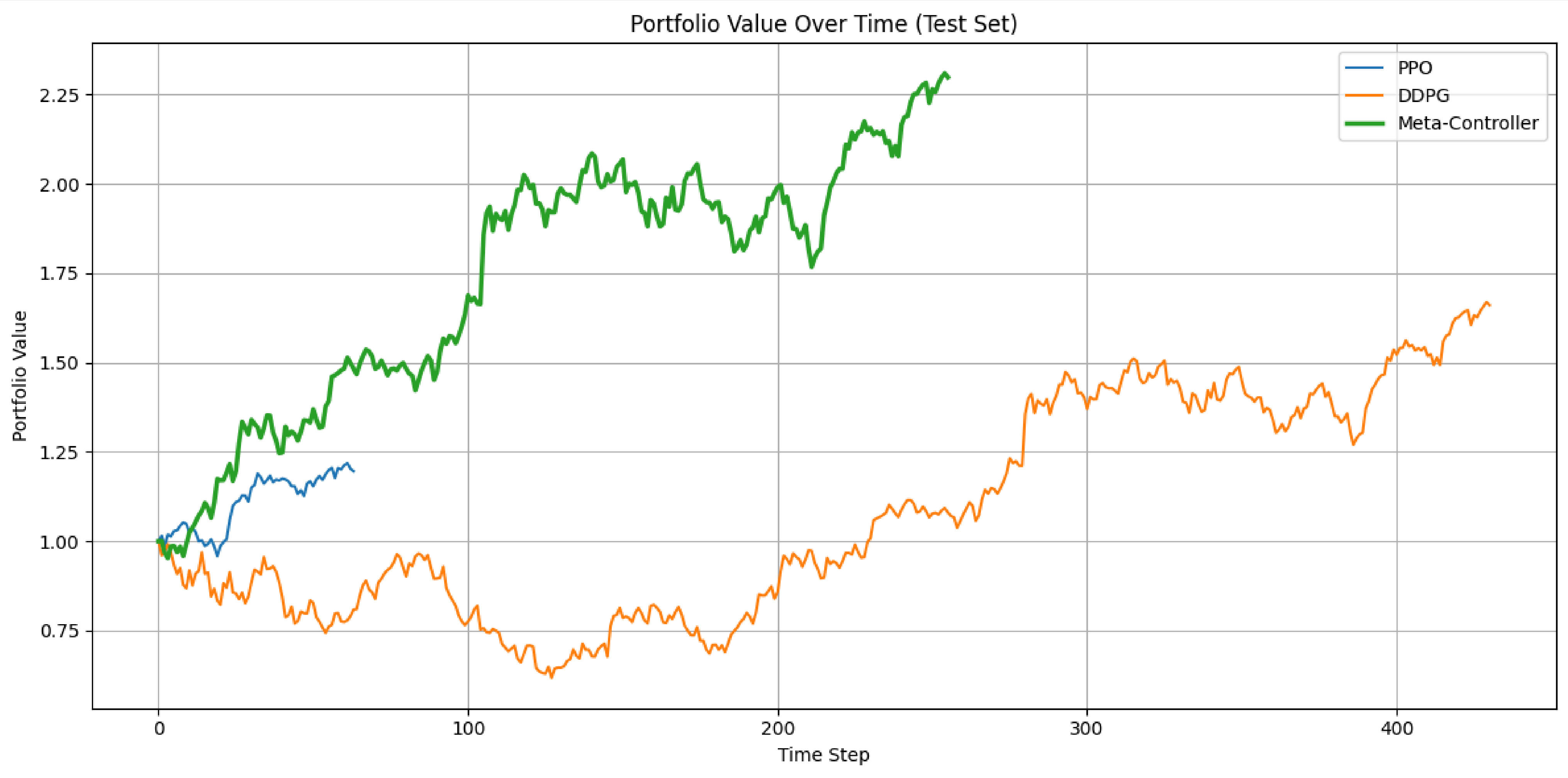
$$\text{Actionfinal} = \alpha \cdot \text{PPO} + (1 - \alpha) \cdot \text{DDPG}$$

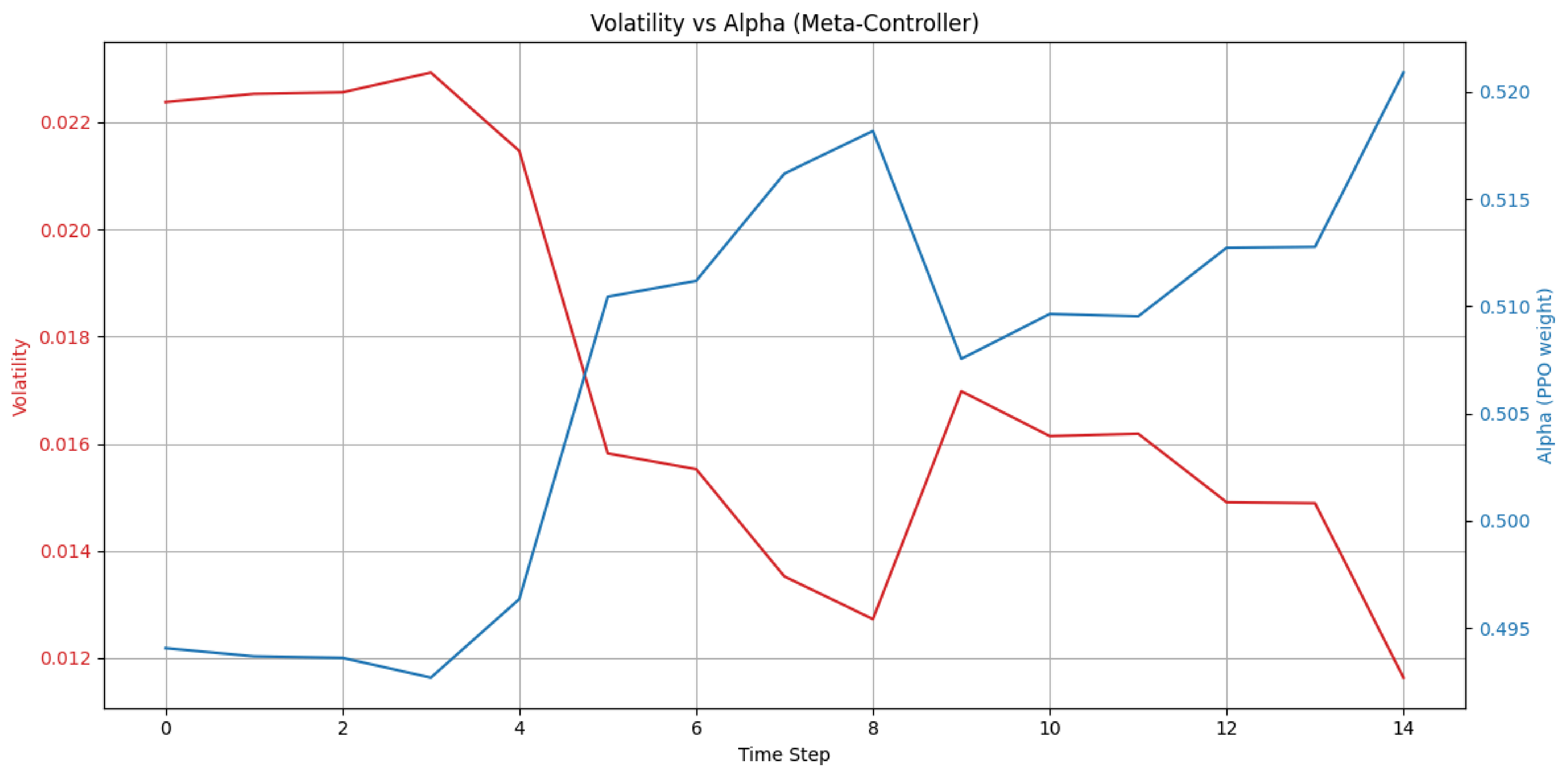


Agent	Annualized Return	Volatility	Sharpe Ratio	Max Drawdown
PPO	115.91%	29.53%	2.61	-9.01%
DDPG	158.17%	32.63%	2.91	-15.95%
Meta-Controller	175.10%	15.75%	6.44	-2.45%

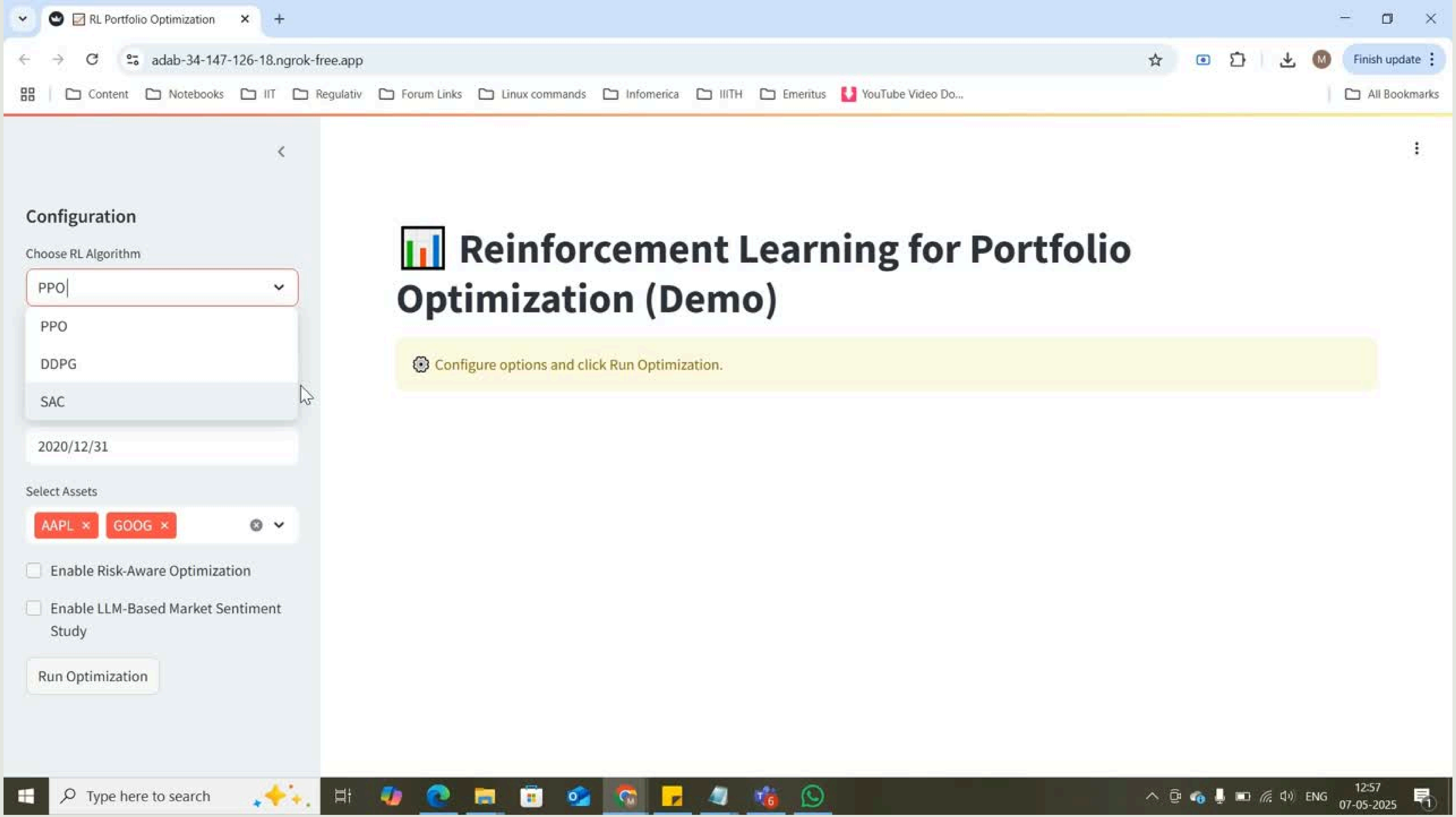
Table 1: Performance comparison of PPO, DDPG, and Meta-Controller

- Meta-controller outperforms both agents
- Better risk-adjusted return
- Reduced drawdown from adaptive switching


Portfolio Value Over Time (Test Set)
PPO
DDPG
Meta-Controller
Portfolio Value
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0
100
200
300
400
Time Step


Volatility vs Alpha (Meta-Controller)
Volatility
0.022
0.020
0.018
0.016
0.014
0.012
Alpha (PPO weight)
0.520
0.515
0.510
0.505
0.500
0.495
0
2
4
6
8
10
12
14
Time Step

Demo



📊 Reinforcement Learning for Portfolio Optimization (Demo)

Configure options and click Run Optimization.

Configuration

Choose RL Algorithm: PPO

- PPO
- DDPG
- SAC

2020/12/31

Select Assets: AAPL × GOOG ×

☐ Enable Risk-Aware Optimization

☐ Enable LLM-Based Market Sentiment Study