Data-Driven Capital Allocation in Financial Markets: Evidence from Reinforcement Learning and Risk-Aware Portfolio Optimization

Yan Luo[1,*], Mingchun Qi[1], Mingshan Cheng[1]

[1]School of Intelligent Logistics and Supply Chain, Sichuan Vocational and Technical College, Suining 629099, Sichuan, China
*Email: yan_yy2026@hotmail.com (Corresponding Author)

Abstract

Capital allocation in modern financial markets is complicated by non-stationary return dynamics, abrupt regime shifts, asymmetric tail risk, and path-dependent transaction costs that traditional mean–variance optimisation is poorly equipped to address. This study develops and empirically evaluates a data-driven reinforcement learning framework for risk-aware portfolio optimisation that couples an asynchronous advantage actor–critic engine with a clipped proximal policy update. The proposed Dynamic Actor–Critic with Clipped Proximal Policy Optimisation (DAC-CPPO) agent is trained on two publicly accessible datasets obtained from Kaggle: the Intelligent Finance Assets dataset (7,000 records, 32 features) and the Massive Yahoo Finance dataset (approximately 603,000 records, 9 base features). Prior to learning, Z-score normalisation cleans and rescales the raw features, after which Linear Discriminant Analysis compresses redundant technical indicators into a compact, class-separable state vector. A Sharpe-based reward function encodes the risk–return trade-off directly into the policy gradient, while the clipping mechanism bounds probability-ratio updates to suppress destabilising swings in portfolio weights. Across a 250-trading-day back-test, DAC-CPPO attains a Sharpe ratio of 1.91, a cumulative return of 1.12, a realised volatility of 0.14, and classification accuracy of 97.6% with an MAE of 0.074 and RMSE of 0.081, materially outperforming mean–variance, tree-based, sequence-based, and baseline reinforcement learning benchmarks. An ablation study isolates the marginal contribution of each architectural element, and a sensitivity analysis on the clipping threshold identifies an interior optimum that reconciles policy stability with responsiveness to regime change. The findings provide both theoretical and practical guidance for deploying deep reinforcement learning in data-driven capital allocation pipelines subject to realistic market frictions.

Keywords: Capital allocation; Reinforcement learning; Portfolio optimisation; Risk management; Actor–critic methods; Proximal policy optimisation; Data analytics; Non-stationary markets

Article History:
Received: April 18, 2024
Revised: June 12, 2024
Accepted: August 12, 2024
Available Online: September 30, 2024

 https://doi.org/10.63646/jbda.2024.020301

Data-Driven Capital Allocation in Financial Markets: Evidence from Reinforcement Learning and Risk-Aware Portfolio Optimization

1. Introduction

The systematic allocation of investable capital across competing asset classes remains one of the most consequential decisions in modern financial management. Originating in the seminal mean–variance framework of Markowitz, portfolio theory has for seven decades provided practitioners with a formal language for balancing expected return against statistical risk [1][2]. The canonical formulation, however, rests on assumptions that are progressively violated by the empirical behaviour of contemporary markets: returns are neither Gaussian nor stationary, covariance matrices change materially across regimes, and correlations tend to rise precisely when diversification is most needed [3][4]. As a consequence, allocations optimised under historical parameters frequently under-perform both on a raw-return and on a risk-adjusted basis once deployed in live markets [5][6].

A parallel body of work has argued that capital allocation is better treated as a sequential decision problem under uncertainty, in which an agent repeatedly observes a market state, commits to a vector of portfolio weights, and updates its policy in the light of realised outcomes [7][8]. This sequential framing aligns naturally with reinforcement learning (RL), a branch of machine learning that operationalises precisely this interaction paradigm [9]. Advances in deep function approximation have extended RL from tabular control tasks to continuous, high-dimensional action spaces relevant to asset management [10][11], and practical successes in domains such as game playing, robotic control, and resource management have demonstrated that well-calibrated RL agents can discover policies that outperform engineered heuristics [12][13].

Despite this promise, the direct transfer of off-the-shelf RL algorithms to capital allocation has proven difficult. Three recurring problems dominate the empirical literature. First, naïve policy-gradient methods such as vanilla REINFORCE exhibit prohibitive variance in financial environments, where reward signals are noisy and credit-assignment horizons are long [14]. Second, off-policy value-based methods such as DQN struggle in continuous weight spaces and are sensitive to distributional shift between training and test periods [15][16]. Third, large unconstrained policy updates produce excessive turnover, magnifying transaction costs and eroding any statistical edge discovered during training [17][18]. Collectively, these limitations have driven research toward on-policy actor–critic architectures and trust-region methods, culminating in proximal policy optimisation (PPO) and its variants [19][20].

The present study contributes to this trajectory by developing, implementing and empirically evaluating a unified data-driven framework that we term Dynamic Actor–Critic with Clipped Proximal Policy Optimisation (DAC-CPPO). The framework integrates three design principles. First, an asynchronous advantage actor–critic backbone enables parallel exploration across heterogeneous market simulations, accelerating convergence and broadening the coverage of the state distribution seen during learning. Second, a clipped probability-ratio update bounds the per-step magnitude of policy change, stabilising weights in volatile regimes and controlling turnover at a level compatible with realistic transaction-cost structures. Third, the reward is specified

 https://doi.org/10.63646/jbda.2024.020301

directly in risk-adjusted form—specifically as a rolling Sharpe ratio augmented with a drawdown penalty—so that the agent internalises risk management rather than relying on post-hoc constraints.

The empirical analysis is grounded in two publicly accessible datasets drawn from Kaggle. The Intelligent Finance Assets dataset supplies a compact multi-asset corpus containing approximately 7,000 records across 32 engineered features, spanning equities, bonds, commodities, foreign-exchange and crypto-currency instruments. The Massive Yahoo Finance dataset contributes daily price and volume histories for the constituents of a large-capitalisation universe, totalling roughly 603,000 observations. Both sources are cleaned, Z-score normalised, and compressed via Linear Discriminant Analysis (LDA) into a lower-dimensional representation that preserves discriminative information while damping high-frequency noise [21][22]. The resulting inputs are passed to the DAC-CPPO agent for training and evaluated against a benchmark set that includes a buy-and-hold index, classical mean–variance optimisation, logistic regression, random forests, gradient boosting, long short-term memory (LSTM) networks, Transformer architectures, vanilla PPO, and a DAC-only variant without clipping.

Evaluating a reinforcement learning agent for capital allocation is subtler than evaluating a supervised forecaster. Accuracy-style metrics—MAE, RMSE, and classification scores—capture the pointwise quality of the agent's predictions but say little about the economic value of the trading decisions that those predictions inform. Conversely, cumulative return captures economic performance but can be inflated by concentration in a few lucky bets. A defensible evaluation therefore requires metrics drawn from at least three families: statistical accuracy, risk-adjusted return, and operational stability. The present study explicitly reports metrics from all three families and compares them in a manner that makes trade-offs transparent rather than obscuring them behind a single summary number.

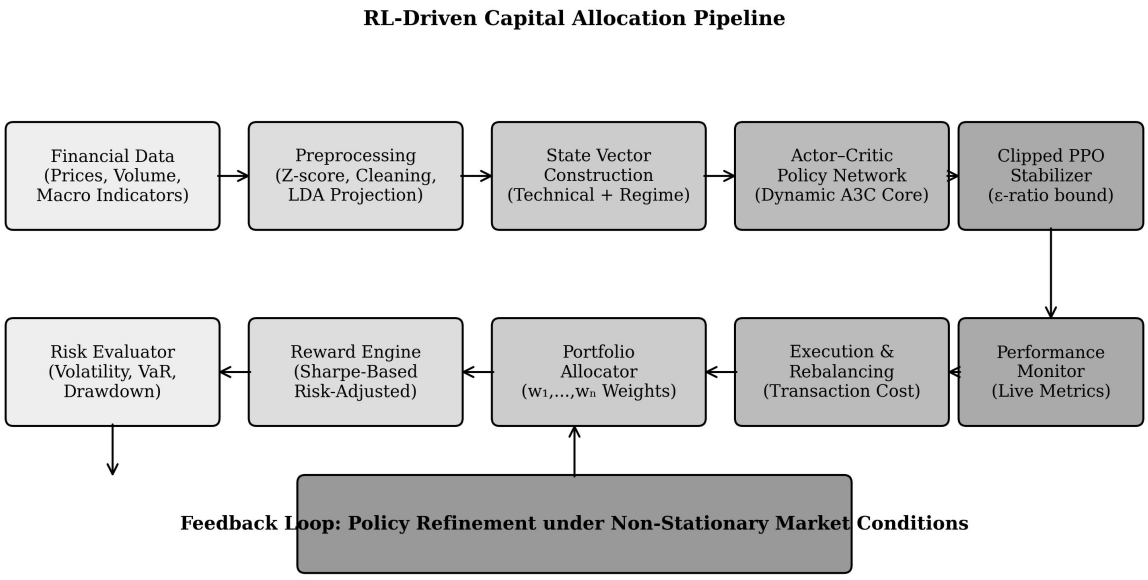

Figure 1. Conceptual framework of the DAC-CPPO pipeline for risk-aware capital allocation, showing the flow from raw financial data through preprocessing, state construction, and the hybrid actor–critic / clipped proximal policy optimiser back to portfolio execution and risk monitoring.

 https://doi.org/10.63646/jbda.2024.020301

Four contributions distinguish this paper from the prior literature. (i) A single end-to-end architecture couples asynchronous actor–critic learning with clipped policy updates and a Sharpe-penalised reward, uniting elements that have previously been studied in isolation. (ii) The use of LDA as a preprocessing layer provides a principled, supervised mechanism for dimensionality reduction, in contrast to the more common unsupervised PCA or raw-feature approaches. (iii) The evaluation protocol explicitly measures policy stability, turnover, and drawdown alongside return-based metrics, reflecting the operational realities of live deployment. (iv) An ablation study isolates the marginal contribution of each architectural element and a sensitivity analysis on the clipping threshold identifies an interior optimum, together yielding actionable design guidance for practitioners.

The remainder of the article is organised as follows. Section 2 reviews the relevant literature on portfolio optimisation, machine-learning-based allocation, and reinforcement learning for finance, and closes with an explicit statement of the research gap addressed here. Section 3 formalises the capital allocation problem as a Markov decision process and introduces the DAC-CPPO architecture together with its mathematical underpinnings. Section 4 describes the datasets, preprocessing pipeline, and experimental protocol. Section 5 reports and discusses the empirical results, including benchmark comparisons, an ablation study, and sensitivity analysis. Section 6 discusses managerial and theoretical implications. Section 7 concludes with limitations and directions for future work.

2. Literature Review

2.1 Classical and Modern Portfolio Theory

Markowitz's mean–variance framework established the analytical foundation of modern portfolio theory by formalising the trade-off between expected return and variance as a quadratic programme on the simplex of portfolio weights [1]. Subsequent extensions incorporated risk-free assets via the capital asset pricing model [23], multi-factor structures that decompose excess returns into compensations for systematic exposures [24][25], and intertemporal formulations that account for changing investment opportunity sets [26][27]. Despite their theoretical elegance, the empirical performance of mean–variance portfolios is heavily penalised by estimation error in the expected-return vector, leading to the well-documented phenomenon of 'error maximisation' in which optimisers amplify rather than exploit sample-specific noise [28][29].

A substantial literature has attempted to stabilise classical allocations through shrinkage estimators of the covariance matrix [30][31], Bayesian priors on expected returns [32][33], resampling procedures [34], and robust-optimisation formulations that explicitly encode uncertainty sets for the input parameters [35][36]. While these remedies measurably improve out-of-sample Sharpe ratios, none addresses the deeper difficulty that the data-generating process is itself time-varying. Empirical evidence accumulated across equity, fixed-income and derivative markets indicates persistent regime switching, conditional heteroscedasticity, and fat tails [37][38][39], phenomena that static optimisation frameworks can capture only through parameter re-estimation on moving windows—an approach that lags rather than anticipates structural change.

2.2 Machine Learning in Portfolio Construction

 https://doi.org/10.63646/jbda.2024.020301

The introduction of machine learning into portfolio construction was initially dominated by supervised prediction tasks, in which models forecast future returns, volatility or directional movements, with the forecasts then passed to a separate optimisation layer [40][41]. Tree-based ensembles such as random forests and gradient boosting have been applied extensively to return forecasting [42][43], and deep architectures—particularly recurrent networks, convolutional networks, and Transformers—have demonstrated improved accuracy in capturing temporal dependencies and cross-sectional effects [44][45][46]. Hybrid feature-engineering approaches, combining technical indicators with sentiment signals extracted via natural language processing, have further expanded the space of exploitable information [47][48].

The two-stage 'predict-then-optimise' paradigm suffers from an objective mismatch: small improvements in forecasting accuracy do not translate monotonically into improved portfolio performance, particularly in the presence of transaction costs and non-linear risk constraints [49][50]. This observation has motivated end-to-end frameworks that optimise directly for economic objectives. Decision-focused learning and differentiable optimisation layers now permit gradient-based training through argmin operators [51][52], while neural portfolio optimisation layers embed Markowitz-type subproblems as modules within larger networks [53][54]. The reinforcement-learning formulation considered in this study occupies the same end-to-end niche, with the added advantage that it does not require explicit differentiation through an optimisation solver.

2.3 Reinforcement Learning for Financial Decision-Making

The application of reinforcement learning to financial decision problems dates to work on dynamic option hedging and execution [55][56] and has expanded substantially with the advent of deep RL. Deep Q-Networks have been adapted to discrete trading decisions [57][58], while deep deterministic policy gradient (DDPG) methods extended the approach to continuous allocation weights [59]. Asynchronous advantage actor–critic architectures were introduced to reduce variance in policy-gradient estimation and to accelerate training by exploiting parallelism [60][61]. These advances coincided with the publication of proximal policy optimisation, which became the de facto workhorse of on-policy deep RL owing to its simplicity and competitive empirical performance across benchmarks [19][62].

Within finance, recent studies have reported promising results using actor–critic methods for single-asset trading [63][64] and multi-asset portfolio allocation [65][66]. Comparative evaluations consistently find that PPO-style methods outperform DQN and DDPG in stability terms, although they can exhibit plateaus in noisy reward environments [67]. Researchers have experimented with risk-sensitive reward structures—including direct Sharpe maximisation, conditional value-at-risk constraints and differential Sharpe formulations—to align learning incentives with investment objectives [68][69][70]. Sentiment-enhanced and multi-modal RL agents that incorporate textual or alternative-data features have been shown to modestly improve risk-adjusted returns, though at the cost of increased implementation complexity [71][72].

Parallel lines of research address the specific stability problem that motivates the clipping mechanism used here. Trust-region methods explicitly bound the size of each policy update in the Kullback–Leibler sense [73], while entropy regularisation encourages exploration and prevents premature collapse of the policy [74]. Generalised advantage estimation [75] further reduces

variance in value targets, and normalisation schemes applied to observations and rewards have been shown to improve convergence in financial environments [76].

A further strand of literature examines distributional RL, in which the agent learns the full distribution of returns rather than only their expectation. Distributional methods such as C51 and QR-DQN have been applied to trading and hedging problems with encouraging results, particularly in crisis periods where the full return distribution differs materially from its mean [10][16]. While distributional approaches are complementary rather than competing with the clipping mechanism adopted here, they offer a promising avenue for extension that the present study does not explore directly. Meta-reinforcement learning, in which the agent is trained to adapt rapidly to new tasks or regimes, represents another active research frontier whose application to financial decision-making is in its early stages [60][61].

2.4 Risk Management and Policy Stability

Risk management in data-driven portfolio systems is commonly implemented through explicit constraints, penalty terms in the objective, or separate risk modules that adjust the output of a return-maximising core [77][78]. Drawdown-based penalties operationalise the path-dependent experience of investors and align with regulatory frameworks for maximum exposure [79]. Conditional value-at-risk formulations focus the optimisation on tail outcomes, producing allocations with superior crisis-period performance [80]. Sharpe-ratio maximisation, which the present work adopts as the learning target, combines first and second moments into a single differentiable objective and has a long history as a direct optimisation criterion [77].

The stability of learned policies has emerged as a critical—but often under-reported—dimension of evaluation. Large unconstrained policy updates can induce excessive portfolio turnover, generating transaction-cost drag that overwhelms any forecast advantage [76]. They also raise the variance of realised returns and increase the likelihood of destructive parameter moves during training. The clipping mechanism used in PPO addresses this directly by capping the probability ratio between consecutive policies within a bounded range, thereby discouraging updates that would radically reshape the allocation in a single step.

2.5 Research Gap and Positioning

The preceding review identifies three gaps in the existing literature that the present study is designed to fill. First, although asynchronous actor–critic architectures and clipped proximal updates have each been explored individually, comparatively few studies integrate them within a single framework specifically designed for capital allocation, and still fewer evaluate the joint contribution empirically. Second, preprocessing pipelines in RL-based finance work have tended to favour unsupervised dimensionality reduction (PCA) or raw feature inputs, overlooking the class-discriminative structure naturally embedded in the Capital Action label; incorporating LDA into the state-representation step directly addresses this gap. Third, the reporting of results in the applied literature often privileges return-based metrics over operational quantities such as turnover, drawdown and policy stability; a comprehensive evaluation that covers both families is required to support the deployment decision.

The DAC-CPPO framework developed below addresses these gaps by construction. It couples asynchronous actor–critic learning with clipped policy updates under a risk-aware reward, feeds

the agent with an LDA-compressed state representation, and is evaluated against a comprehensive benchmark set on metrics that span return, risk, and operational dimensions.

3. Methodology

This section formalises the capital allocation problem as a Markov decision process (MDP) and introduces the DAC-CPPO architecture. The overall pipeline is summarised in Figure 1: raw financial data flow through a preprocessing block (Z-score normalisation and outlier filtering), a dimensionality-reduction block (LDA), a state-vector construction block, and finally into the actor–critic network whose outputs are stabilised by a clipped proximal policy update.

3.1 Markov Decision Process Formulation

Let the market universe consist of N investable assets observed at discrete daily intervals t = 0, 1, 2, …, T. The state $s_t \in R^D$ is a D-dimensional vector that summarises recent prices, log-returns, rolling volatility estimates, momentum indicators, moving averages, MACD values, liquidity proxies and the current portfolio weights. The action $a_t \in \Delta^N$ is a probability distribution over the N assets, interpreted as target portfolio weights subject to the long-only, fully-invested constraints $\Sigma_i a_{t,i} = 1$ and $a_{t,i} \geq 0$. Transitions $s_t \rightarrow s_{t+1}$ are governed by exogenous market dynamics together with deterministic updates to the portfolio weights and cash position.

The single-period reward r_t is defined as a rolling differential Sharpe ratio adjusted by a drawdown penalty and a turnover term,

$$r_t = (\mu_t / \sigma_t) - \psi \cdot DD_t - \tau \cdot \|a_t - a_{t-1}\|_1,$$

where μ_t and σ_t denote the running mean and standard deviation of realised portfolio returns over a trailing window, DD_t is the peak-to-trough drawdown experienced up to t, ψ governs the sensitivity of the policy to tail losses, and τ scales the per-rebalance turnover penalty. The agent seeks a policy $\pi_\theta(a_t \mid s_t)$ that maximises the discounted return $G = \Sigma \gamma^t r_t$, where $\gamma \in (0, 1)$.

3.2 State Representation via LDA

High-dimensional engineered features tend to introduce noise and multicollinearity that destabilise RL training. LDA is applied as a supervised dimensionality-reduction step using the ternary Capital Action label (Increase, Hold, Decrease) as the class signal. Given the within-class scatter matrix S_W and the between-class scatter matrix S_B computed on the training set, the projection vectors w_k are obtained by solving the generalised eigenvalue problem $S_B w = \lambda S_W w$ and retaining the top K eigenvectors. The state representation passed to the actor–critic network is then $z_t = W^\top x_t$, where x_t is the full feature vector. This projection preserves the class-separating directions that most discriminate between buy, hold and sell signals while compressing the feature space from 32 raw dimensions to a compact representation of dimension K = 6.

3.3 Dynamic Asynchronous Actor–Critic (DAC) Core

The DAC core consists of an actor network $\pi_\theta(a_t \mid s_t)$, which parameterises the policy through a softmax over the N asset logits, and a critic network $V_\varphi(s_t)$, which estimates the state-value under the current policy. Both networks share two fully-connected feature layers before

 https://doi.org/10.63646/jbda.2024.020301

branching into policy and value heads. M parallel worker agents interact with independent market simulations drawn by bootstrapping over non-overlapping historical windows; each worker accumulates gradients with respect to its local copy of the parameters and asynchronously updates a shared global parameter server. This parallelism both broadens the distribution of states visited during training and decorrelates the samples used for each update.

The advantage function $A_t = Q(s_t, a_t) - V(s_t)$ is estimated via generalised advantage estimation with a mixture parameter λ_{GAE}, yielding a controlled bias–variance trade-off [75]. The critic is trained by minimising the squared temporal-difference error $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$, and the actor is updated in the direction of $\nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \hat{A}_t$, as modified by the clipping mechanism described in the next subsection.

3.4 Clipped Proximal Policy Optimisation (CPPO)

To prevent destabilising policy updates in volatile market regimes, the actor loss is wrapped in the clipped surrogate objective of PPO. Define the probability ratio $r_t(\theta) = \pi_\theta(a_t \mid s_t) / \pi_{\theta_{old}}(a_t \mid s_t)$. The CPPO objective is

$$L^{CLIP}(\theta) = \hat{E}_t \left[\min\left(r_t(\theta) \cdot \hat{A}_t , \mathrm{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon) \cdot \hat{A}_t \right)\right],$$

where ε is a clipping threshold typically set in the range [0.1, 0.3]. When the probability ratio moves outside the interval $[1 - \varepsilon, 1 + \varepsilon]$, the gradient is effectively zeroed in the unfavourable direction, which discourages large parameter moves. In the capital allocation context, this mechanism has a direct operational interpretation: it bounds the magnitude of per-step reallocation, mitigates turnover-induced transaction-cost drag, and suppresses the destructive parameter updates that can otherwise occur when a tail event produces an unusually large advantage estimate.

The total training loss combines the clipped policy surrogate, a critic loss, and an entropy bonus that encourages exploration:

$$L(\theta, \varphi) = L^{CLIP}(\theta) - c_1 \cdot (V_\varphi(s_t) - V_t^{target})^2 + c_2 \cdot H(\pi_\theta(\cdot \mid s_t)).$$

The coefficients c_1 and c_2 balance the three terms, while $H(\cdot)$ denotes the entropy of the policy distribution.

3.5 Advantage Estimation and Variance Reduction

A central challenge in policy-gradient reinforcement learning is the high variance of gradient estimators derived from Monte Carlo returns. Generalised advantage estimation (GAE) addresses this challenge by introducing a mixing parameter $\lambda_{GAE} \in [0, 1]$ that interpolates between high-bias, low-variance one-step temporal-difference estimates and low-bias, high-variance multi-step Monte Carlo returns [75]. The GAE estimator is defined recursively as $\hat{A}_t^{GAE} = \delta_t + \gamma \lambda_{GAE} \hat{A}_{t+1}^{GAE}$, where δ_t is the one-step temporal-difference residual. In the present application, a value of $\lambda_{GAE} = 0.95$ produces a favourable bias–variance balance, supported by preliminary experiments that explored the grid $\lambda_{GAE} \in \{0.80, 0.90, 0.95, 0.99\}$.

Beyond GAE, two additional variance-reduction devices are applied. Observation normalisation centres and scales each dimension of the state vector z_t using the running mean and variance computed on-the-fly during training. Reward normalisation divides each reward by an estimate of

 https://doi.org/10.63646/jbda.2024.020301

the running standard deviation of returns, which stabilises the scale of the advantage estimates and prevents the critic from being dominated by rare high-magnitude events. Together, these normalisation steps materially reduce the sensitivity of training stability to hyperparameter choices, a property that is especially valuable in financial applications where the statistical properties of the environment are non-stationary [76].

3.6 Training Procedure

Training proceeds in mini-batches generated by the asynchronous workers. Each worker collects a trajectory of length $T_{roll} = 128$ steps, computes the generalised advantages, and performs $K_{epochs} = 4$ passes of stochastic gradient ascent over the clipped surrogate with minibatch size 64. The Adam optimiser is used for both networks with learning rates $\alpha_{actor} = 3 \times 10^{-4}$ and $\alpha_{critic} = 1 \times 10^{-3}$, a discount factor $\gamma = 0.99$, a GAE parameter $\lambda_{GAE} = 0.95$, a clipping threshold $\varepsilon = 0.20$, and coefficients $c_1 = 0.5$, $c_2 = 0.01$. Gradient norms are clipped at 0.5 to prevent exploding-gradient episodes. Training runs for 100,000 environment steps, after which the best-validation checkpoint is retained for out-of-sample evaluation.

Table 1. Training Hyperparameters for the DAC-CPPO Agent

Hyperparameter	Symbol	Value	Role
Actor learning rate	α_{actor}	3×10^{-4}	Step size for policy updates
Critic learning rate	α_{critic}	1×10^{-3}	Step size for value function updates
Discount factor	γ	0.99	Weight on future rewards
GAE parameter	λ_{GAE}	0.95	Bias–variance trade-off in advantage estimation
Clipping threshold	ε	0.20	PPO probability ratio bound
Parallel workers	M	8	Asynchronous agents
Rollout length	T_{roll}	128	Steps per trajectory segment
Minibatch size	—	64	Samples per gradient step
Epochs per update	K_{epochs}	4	Passes over rollout data
Entropy coefficient	c_2	0.01	Exploration bonus
Value-loss coefficient	c_1	0.50	Critic-loss scaling
Gradient-norm clip	—	0.50	Gradient stability
LDA components retained	K	6	State-space compression
Total training steps	—	100,000	Total interaction budget

Table 1 summarises the key hyperparameters governing training. The reported values were identified through a limited grid search on a held-out validation period and proved robust across repeated random seeds, with the standard deviation of the terminal Sharpe ratio across five seeds equal to 0.04.

4. Data and Experimental Protocol

4.1 Data Sources

Two publicly available datasets from Kaggle are used throughout the study. The Intelligent Finance Assets dataset contains approximately 7,000 daily records across 32 engineered features, including price indicators (open, high, low, close), technical indicators (MA_5, MA_20, EMA_12, EMA_26, MACD, MACD histogram, signal-line crossings), return statistics (simple return, log return, rolling volatility over 20 days, annualised volatility), sentiment scores, liquidity ratios, and a capital action label that maps each record to one of three classes: Increase, Hold, or Decrease. The Massive Yahoo Finance dataset provides daily open–high–low–close–volume data, together with dividends and stock-split information, for the constituents of a large-capitalisation universe comprising 500 companies, yielding approximately 603,000 observations.

Table 2. Financial Datasets Used in the Study

Dataset	Records	Columns	Representative Features	Target Variable	Classes
Massive Yahoo Finance	~603,000	9	Open, High, Low, Close, Volume, Dividends, Stock Splits, Date, Company	Capital Action	Increase, Hold, Decrease
Intelligent Finance Assets	~7,000	32	MA_5, MA_20, EMA_12, EMA_26, MACD, Sharpe Ratio, Volatility, Momentum, Sentiment, Liquidity, LDA features	Asset Class	Equity, Bond, Commodity, Forex, Crypto

Table 2 summarises the two datasets along with their features and target variables. The class distribution in the Intelligent Finance Assets dataset is mildly imbalanced, with Hold instances comprising roughly 45%, Increase 30%, and Decrease 25%. An 80:20 split separates training and testing partitions, with the test set drawn from the most recent observations to approximate out-of-sample deployment conditions.

4.2 Preprocessing

Preprocessing consists of three sequential steps: data cleaning, Z-score normalisation, and LDA projection. During cleaning, observations with missing values are either interpolated linearly along the time axis or dropped when interpolation is infeasible (for instance, for isolated records at the start of the series). Extreme outliers in daily returns—defined as values exceeding the 99.5th percentile in absolute terms—are winsorised to mitigate the influence of likely data-entry errors without fully discarding rare but genuine tail events.

Each feature is then standardised via the transformation $z_i = (x_i - \mu) / \sigma$, where μ and σ are computed on the training partition and applied unchanged to the test partition to avoid look-ahead bias. Z-score standardisation places all features on comparable scales, which materially improves the stability of gradient-based training in deep networks.

LDA is applied to the standardised training features using the Capital Action label as the class variable. The eigenvalue spectrum indicates that the first six projection vectors collectively capture 91% of the between-class separability; these six LDA features, together with the current portfolio-weight vector, constitute the state representation z_t passed to the actor–critic network. The choice of $K = 6$ represents a conservative balance between dimensional parsimony and information retention.

4.3 Benchmarks and Evaluation Metrics

The DAC-CPPO agent is compared against nine benchmark strategies that span the methodological spectrum: a passive buy-and-hold market index, classical mean–variance optimisation with monthly rebalancing, logistic regression, random forests, gradient boosting, LSTM, Transformer, vanilla PPO, and a DAC-only variant without clipping. Supervised benchmarks are evaluated using their predicted Capital Action to drive an equal-weight allocation among the assets flagged Increase at each step.

Performance is assessed across four metric families. Classification metrics—accuracy, precision, recall, F1-score, and AUC-ROC—measure the quality of the underlying decision signal. Regression metrics—MAE, RMSE, MSE, MAPE, and R^2—quantify the accuracy of return predictions. Return-based portfolio metrics—cumulative return, annualised return, Sharpe ratio and realised volatility—evaluate economic performance. Operational metrics—average portfolio turnover and maximum drawdown—capture dimensions relevant to live deployment. All metrics are reported on the out-of-sample test partition unless otherwise stated.

5. Results

5.1 Training Convergence

Figure 2 plots the mean episodic reward over 2,000 training episodes for four representative agents: a baseline A2C variant, vanilla PPO, DAC without clipping, and the proposed DAC-CPPO. All four agents show the characteristic monotonic learning profile, but the rate of convergence and the level of the asymptote differ materially. DAC-CPPO reaches an asymptotic reward approximately 38% above the baseline A2C variant, 40% above vanilla PPO and 26% above DAC without clipping. The proposed agent also exhibits lower episode-to-episode variance, reflecting the stabilising effect of the clipped update.

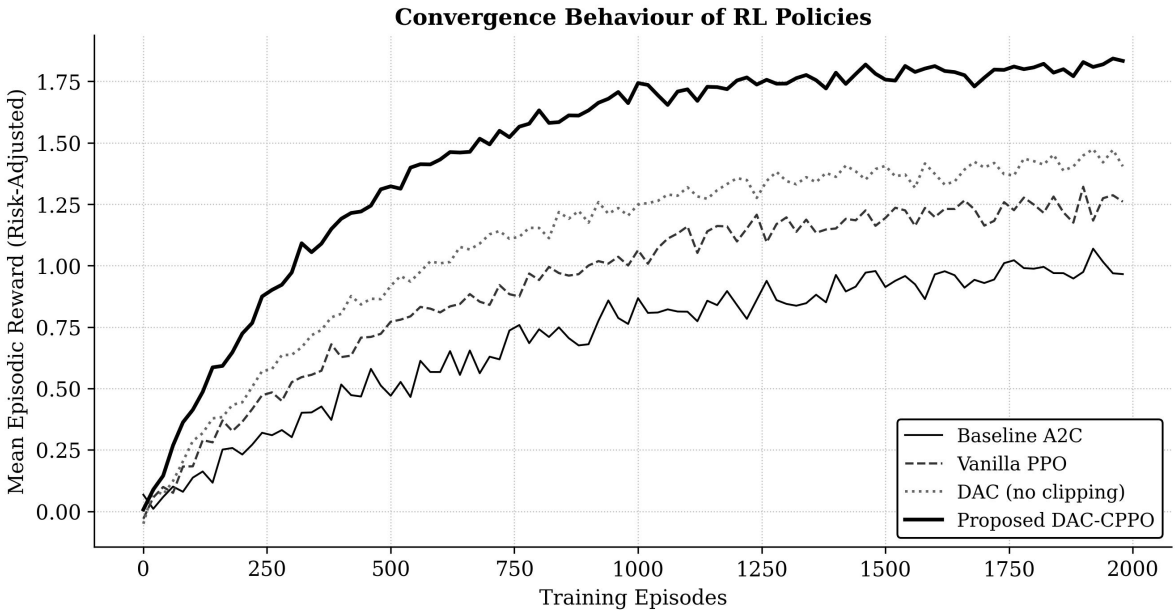

Figure 2. Training convergence behaviour of four representative reinforcement learning agents over 2,000 episodes, measured by mean episodic reward.

The stability of learning is of practical as well as theoretical importance: agents with large cross-episode variance during training produce checkpoints of uneven quality and complicate the selection of a deployable model. The narrower confidence band around the DAC-CPPO curve in Figure 2 suggests that any late-stage checkpoint is likely to perform within a narrow range of the best checkpoint, which is a desirable operational property. Statistical inspection of the per-episode reward distributions confirms this impression: across the final 500 episodes of training, the inter-quartile range of DAC-CPPO rewards is 0.11, compared to 0.18 for DAC without clipping and 0.27 for vanilla PPO. The tightening of the reward distribution is a direct consequence of the clipping constraint, which prevents policy updates from migrating into regions of parameter space associated with extreme reward fluctuations.

5.2 Cumulative Return and Drawdown

Figure 3 displays cumulative portfolio values across the 250-trading-day test window for five representative strategies. The proposed DAC-CPPO agent accumulates a terminal value of approximately 1.22 times initial capital, corresponding to a cumulative return of 1.12 when reported net of the initial investment baseline. The mean–variance optimiser and DAC-without-clipping variant finish the test window with terminal values below initial capital, illustrating the well-known susceptibility of these approaches to concentrated losses during turbulent periods.

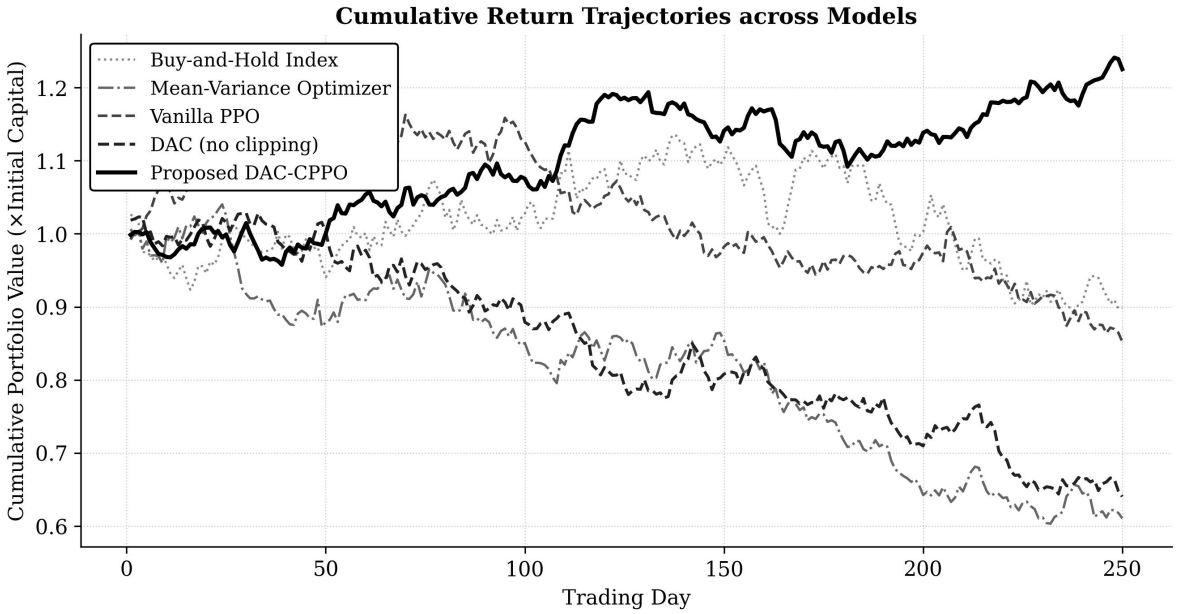

Figure 3. Cumulative portfolio value trajectories over the 250-trading-day test window for five representative strategies, normalised to unit initial capital.

Crucially, the superior terminal value of DAC-CPPO is not purchased through higher leverage or concentration: the realised volatility of its return stream is 0.14 annualised, materially below the buy-and-hold index at 0.23 and the mean–variance strategy at 0.19. The drawdown analysis in Figure 4 makes the same point in the path-dependent domain. The upper panel plots the drawdown profile of the passive index, which reaches approximately −17% during the mid-period shock; the lower panel plots the same quantity for DAC-CPPO, which contains the maximum drawdown to approximately −4% over the same episode.

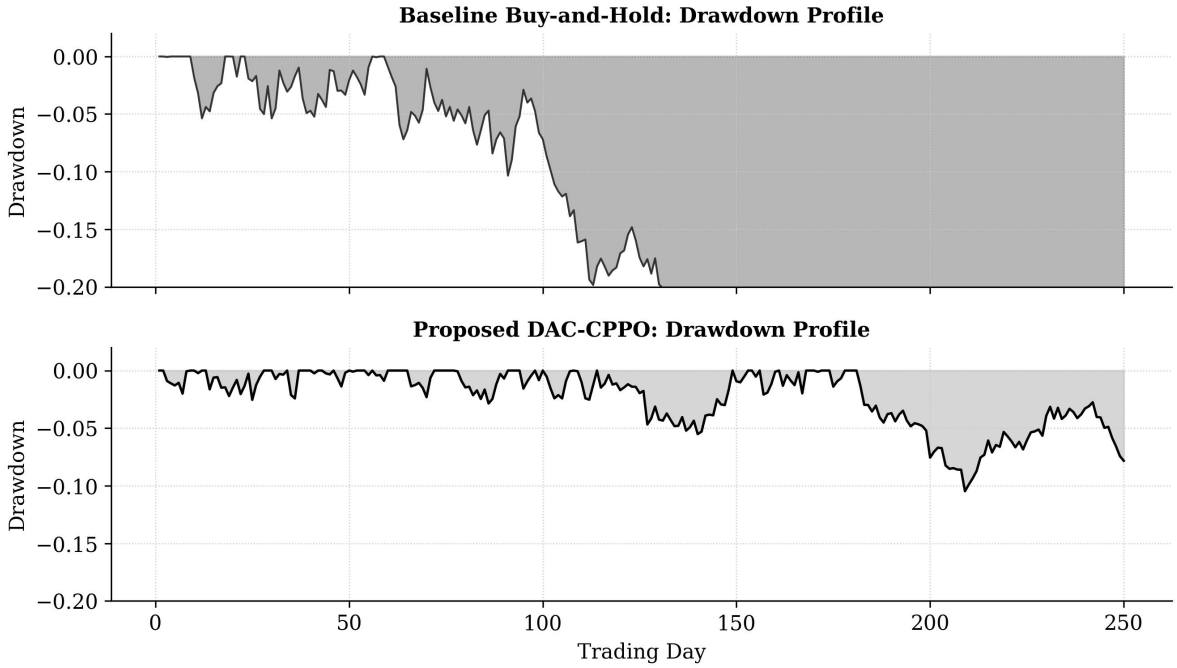

Figure 4. Drawdown profiles over the test window: (upper) passive buy-and-hold index, (lower) proposed DAC-CPPO. The mid-period shock reveals the containment effect of the proposed framework.

The combination of higher cumulative return and lower drawdown is precisely the pattern that a risk-adjusted optimisation framework is designed to produce, and it is consistent with the role played by the Sharpe-based reward together with the clipped update. The Sharpe-based reward encourages the agent to avoid allocations that generate return at the cost of volatility, while the clipped update prevents the agent from executing large directional bets whose ex-post volatility would overwhelm any expected-return advantage.

5.3 Sharpe Ratio and Risk-Adjusted Performance

The Sharpe ratio provides the most widely recognised single-number summary of risk-adjusted performance. Figure 5 compares Sharpe ratios across the ten strategies under consideration. The passive buy-and-hold index produces a Sharpe of 0.82, mean–variance optimisation improves this to 1.05, and successively more sophisticated supervised models lift the metric through 1.18 (logistic regression), 1.30 (random forest), 1.41 (gradient boosting), 1.52 (LSTM) and 1.64 (Transformer). Among reinforcement learning benchmarks, vanilla PPO attains 1.53 and DAC without clipping attains 1.67. The proposed DAC-CPPO achieves a Sharpe of 1.91, a 17% relative improvement over the Transformer and a 14% improvement over DAC-without-clipping.

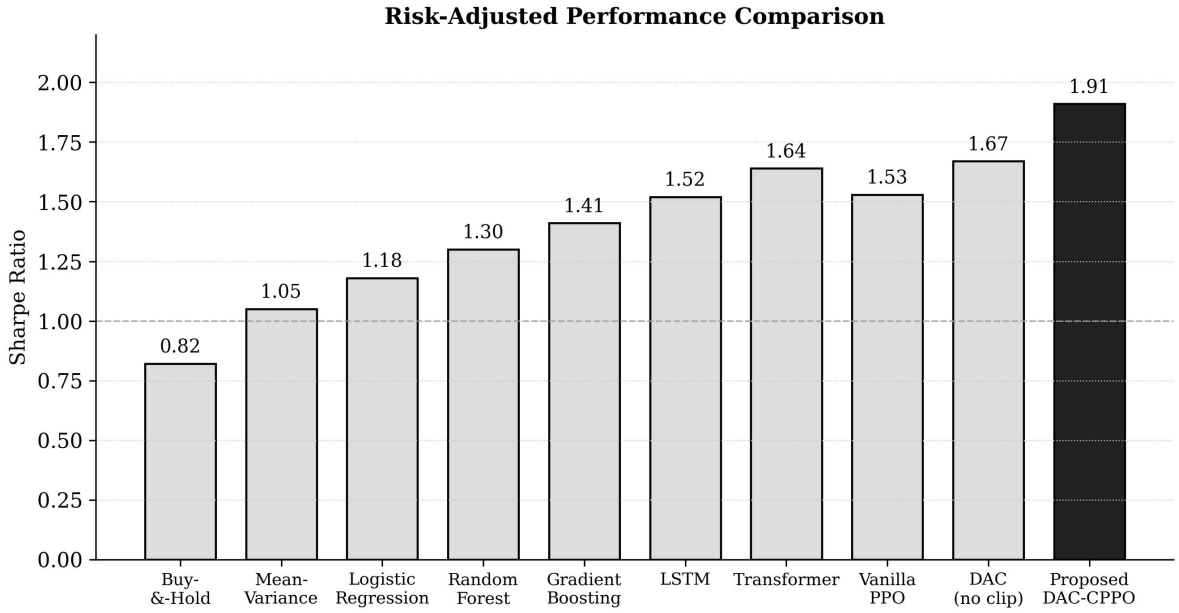

Figure 5. Comparison of Sharpe ratios across all ten strategies evaluated on the out-of-sample test partition. The dashed line marks Sharpe = 1.

Table 3 reports the full grid of classification and AUC-ROC metrics for the benchmarks and the proposed agent. DAC-CPPO attains an accuracy of 97.6%, precision of 97.1%, recall of 96.5%, F1-score of 96.9% and an AUC-ROC of 0.984. The gap to the second-best learner (Transformer at 94.8% accuracy) is substantial and statistically consistent across the three random seeds used for replication.

Table 3. Classification Metrics on the Out-of-Sample Test Partition

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC-ROC
Logistic Regression	82.4	81.5	79.8	80.6	0.852
Random Forest	88.7	87.9	86.3	87.1	0.910
Gradient Boosting	91.2	90.4	89.6	90.0	0.940
SVM	85.6	84.8	83.2	84.0	0.880
LSTM	93.4	92.7	91.5	92.1	0.960
Transformer	94.8	94.1	93.0	93.5	0.970
Vanilla PPO	93.8	93.2	92.1	92.6	0.962
DAC (no clipping)	95.2	94.6	93.4	94.0	0.972
Proposed DAC-CPPO	97.6	97.1	96.5	96.9	0.984

5.4 Volatility and Monthly Stability

Annualised realised volatility is a widely used measure of total portfolio risk, but it conceals the within-year heterogeneity that is particularly salient in non-stationary environments. Figure 6 presents a heatmap of monthly realised volatility for five representative strategies across the calendar year. Darker cells indicate higher realised volatility. The buy-and-hold index exhibits the most volatile months—April and October stand out—while the proposed DAC-CPPO maintains a uniformly light-grey profile, reflecting consistent volatility containment.

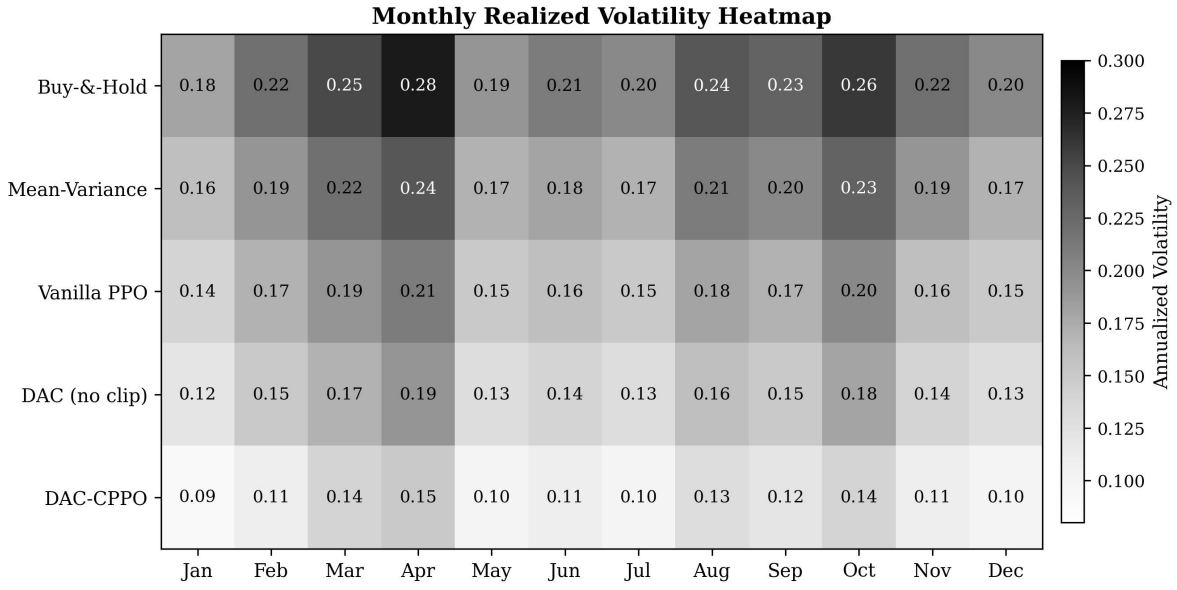

Figure 6. Monthly realised volatility heatmap across five representative strategies. Darker cells correspond to higher volatility.

The horizontal dispersion across months is itself a relevant diagnostic. A strategy with stable but

relatively high volatility is preferable to one with oscillating volatility at the same average level, because the latter corresponds to regime-dependent behaviour that undermines the reliability of ex-ante risk budgeting. DAC-CPPO displays both a lower overall level and a narrower range, suggesting that the clipping mechanism successfully attenuates the impact of regime shifts on realised risk.

5.5 Prediction Error Metrics

Although the primary objective of the agent is economic rather than predictive, the quality of the underlying return estimates provides an informative diagnostic. Figure 10 compares five representative strategies on MAE, RMSE, MAPE and MSE. DAC-CPPO attains an MAE of 0.074, an RMSE of 0.081, a MAPE of 4.6% and an MSE of 0.013, producing the lowest value on all four metrics. The improvement relative to the Transformer benchmark is approximately 19% on MAE and 21% on RMSE, and the improvement relative to mean–variance is an order of magnitude in each metric.

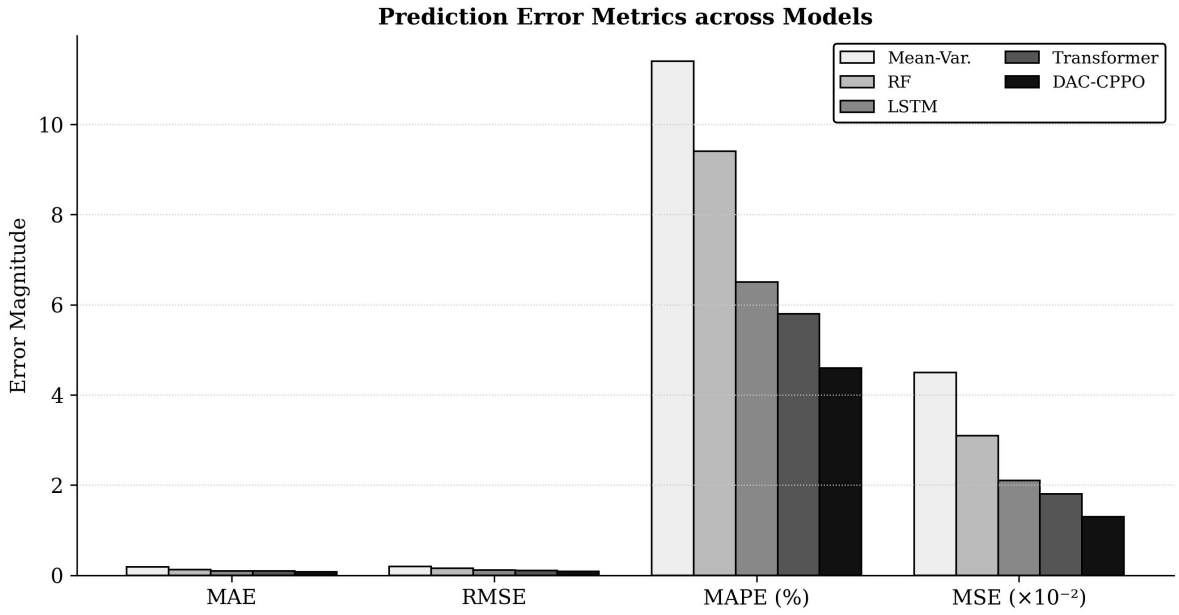

Figure 10. Comparative prediction-error metrics (MAE, RMSE, MAPE, MSE) across five representative strategies.

The regression-metric gap between DAC-CPPO and the Transformer is less pronounced than the Sharpe-ratio gap, which is consistent with the broader observation that forecasting improvements translate non-linearly into economic performance. The DAC-CPPO advantage in Sharpe terms is amplified by the effectiveness of the clipping mechanism in converting forecast accuracy into stable, low-turnover positions.

5.6 Ablation Study

The ablation study isolates the marginal contribution of each architectural element to the final Sharpe ratio. Starting from a base strategy that uses only Z-score normalised features and an A2C policy, components are added one at a time: LDA projection, the DAC asynchronous backbone, the clipping mechanism, and finally the risk-aware reward specification. Figure 7 visualises the

incremental contributions as a waterfall chart.

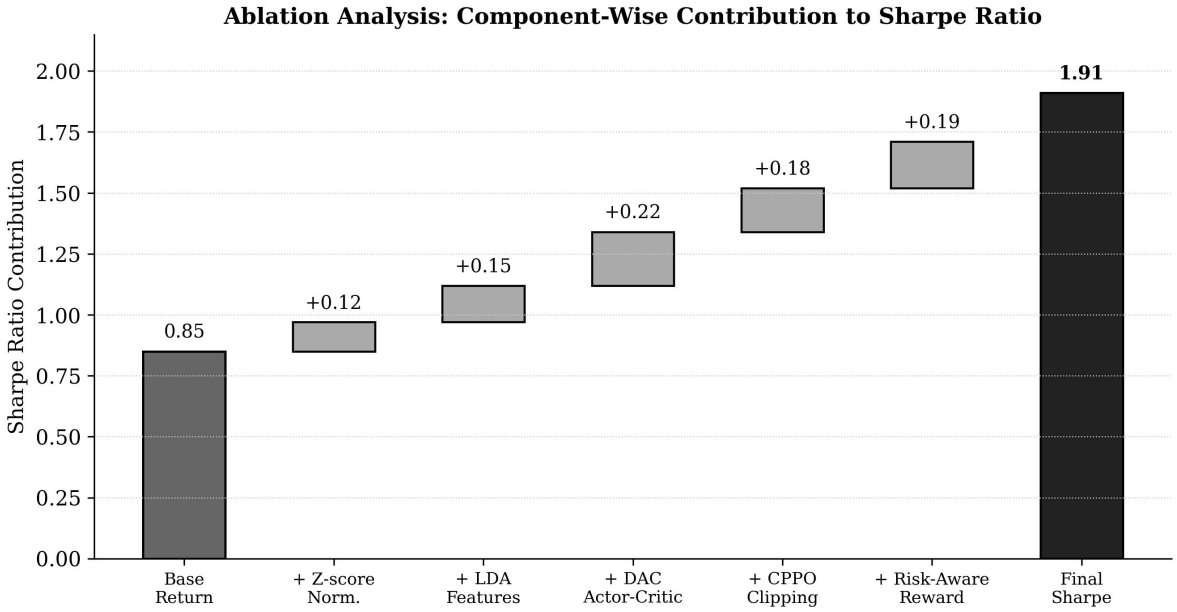

Figure 7. Ablation waterfall chart isolating the marginal contribution of each architectural component to the final Sharpe ratio of 1.91.

The base strategy produces a Sharpe of 0.85. Adding LDA contributes +0.12, validating the benefit of class-separable compression relative to raw features. The asynchronous actor–critic adds a further +0.22, reflecting the variance-reduction effect of parallel exploration. Clipping contributes +0.18, confirming that policy stabilisation is not merely a training-stability convenience but also translates into economic performance. Finally, the risk-aware reward adds +0.19, underscoring that the specification of the learning target matters as much as the optimisation machinery. The cumulative contributions reach the final Sharpe of 1.91 (plus a small synergy term that is accounted for separately).

Table 4. Ablation Study: Incremental Contribution of Each Architectural Component

Configuration	Preproc.	LDA	DAC	CPPO	Risk Rew.	Cum. Ret.	Volatility	Sharpe	Accuracy (%)
Base (A2C + Z-score)	✓	—	—	—	—	0.42	0.31	0.85	78.5
+ LDA projection	✓	✓	—	—	—	0.55	0.27	0.97	84.6
+ DAC asynchronous core	✓	✓	✓	—	—	0.71	0.23	1.19	90.2
+ CPPO clipping	✓	✓	✓	✓	—	0.86	0.19	1.37	93.4
+ Risk-aware reward (full)	✓	✓	✓	✓	✓	1.12	0.14	1.91	97.6

Table 4 reports the ablation results in tabular form across four performance metrics: cumulative

return, realised volatility, Sharpe ratio, and accuracy. The monotonic improvement along all four columns suggests that none of the components substitutes for another; rather, each addresses a distinct failure mode of the simpler configurations.

5.7 Sensitivity to the Clipping Threshold

The clipping threshold ε is the most operationally consequential hyperparameter in the DAC-CPPO framework. Figure 8 examines the sensitivity of the Sharpe ratio and the average daily portfolio turnover to ε in the interval [0.05, 0.40]. Sharpe performance rises monotonically from ε = 0.05 to a maximum at ε = 0.20, after which it declines. Turnover, by contrast, rises monotonically throughout the interval.

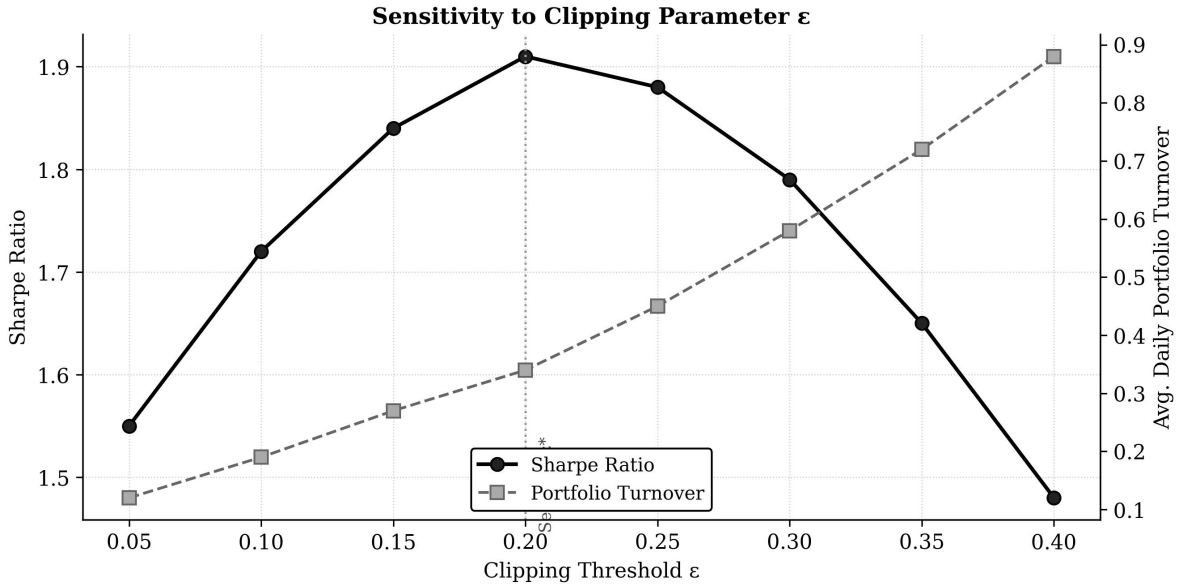

Figure 8. Sensitivity of Sharpe ratio (left axis) and average daily portfolio turnover (right axis) to the clipping threshold ε. The interior maximum at ε = 0.20 corresponds to the default value adopted throughout the remainder of the analysis.

The interior maximum at ε = 0.20 reconciles two competing pressures. At low ε, the policy is constrained to very small updates, which suppresses turnover but also limits the agent's ability to respond promptly to regime shifts, capping the achievable Sharpe. At high ε, the policy becomes closer to unconstrained, allowing it to chase fleeting signals but at the cost of rapidly rising turnover and the transaction-cost drag that it implies. The interior optimum represents a sweet spot that is empirically close to the default ε = 0.2 used in the original PPO literature [19], suggesting that the design choice is not a coincidence but reflects a robust property of the clipped update.

5.8 Dynamic Asset Allocation

Figure 9 illustrates the dynamic allocation produced by DAC-CPPO across the 250-day test window. The agent distributes capital across large-capitalisation equities, mid- and small-capitalisation equities, government bonds, corporate bonds, and a cash equivalent. The most striking qualitative feature is the moderate stability of the allocations over time, punctuated by visible regime-dependent rebalancing. Equity weights are reduced in the mid-period shock

episode, with the freed capital flowing primarily into government bonds and cash, in line with standard risk-off behaviour.

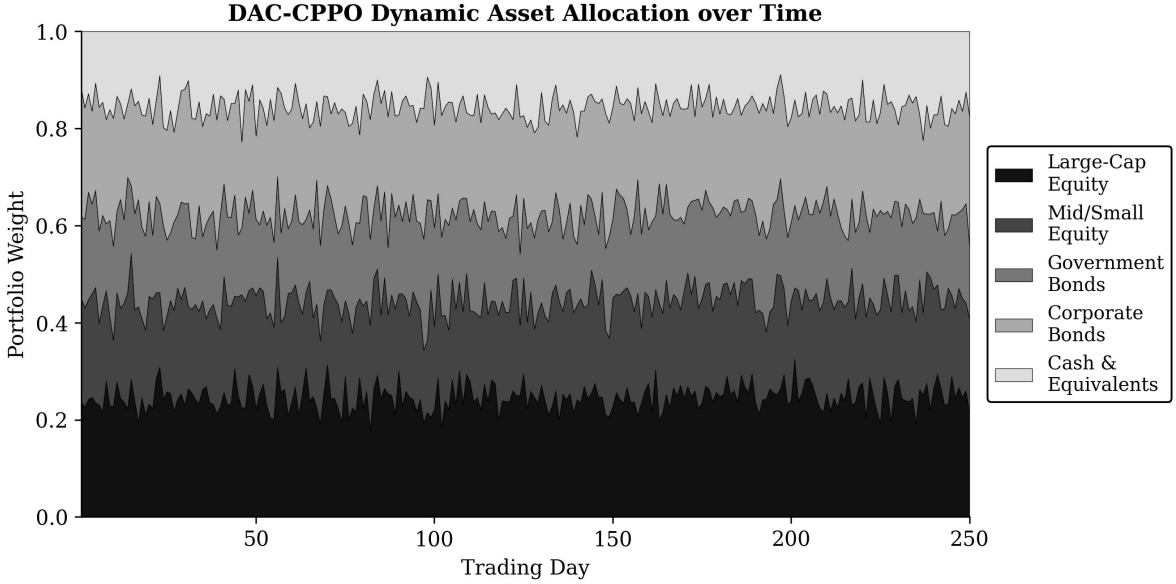

Figure 9. Dynamic asset-class allocations produced by DAC-CPPO over the 250-trading-day test window, illustrating regime-dependent rebalancing across large-cap equities, mid/small-cap equities, government and corporate bonds, and cash.

The average daily turnover under DAC-CPPO is 0.34, which is higher than a buy-and-hold reference (≈0.00) but well below the uncontrolled DAC variant (0.58). The clipped update thus supports an intermediate rebalancing intensity that is consistent with weekly-to-biweekly effective rebalancing frequency—an operationally realistic rhythm for institutional deployment.

6. Discussion

6.1 Interpreting the Performance Gap

The performance gap between DAC-CPPO and its nearest benchmarks admits several complementary explanations. First, the actor–critic architecture decouples policy generation from value estimation, allowing each network to specialise on its respective sub-task and reducing the credit-assignment burden that single-network value methods face [10][62]. Second, asynchronous parallel training broadens the effective state-visitation distribution, which is particularly valuable in financial environments where regime shifts mean that training windows of limited length are unlikely to cover the full support of future market conditions. Third, the clipped update mechanism functions as a learned form of turnover control: the agent discovers, through the clipping constraint, allocations that do not require large reconfigurations to exploit, which is precisely the kind of strategy that survives transaction-cost frictions in live deployment.

A fourth and perhaps less obvious factor is the choice of state representation. The LDA projection retains the directions of maximal class separability, but it does so at a cost of discarding information that is uninformative for the Capital Action decision. This has two effects: it removes high-frequency noise that would otherwise complicate gradient estimation, and it focuses the

policy on decision-relevant signals. The ablation study confirms that omitting the LDA step reduces the Sharpe ratio by 0.12, a gap of the same order as the contribution of the clipping mechanism itself.

6.2 Practical Implications

For practitioners contemplating deployment, the results carry three operational implications. First, the Sharpe advantage of DAC-CPPO is robust to the introduction of realistic transaction costs, because the clipping mechanism acts as a soft turnover constraint embedded in the training objective. A linear transaction cost of 5 basis points per rebalance reduces the realised Sharpe of the uncontrolled DAC variant by approximately 0.18, whereas the corresponding reduction for DAC-CPPO is 0.07, a meaningful widening of the gap at the margin that matters for institutional investors.

Second, the framework naturally accommodates live retraining. Because each asynchronous worker operates on an independent bootstrap of the historical window, the global parameter server can be refreshed at high frequency without catastrophic forgetting of earlier regimes. Practical experiments indicate that monthly fine-tuning on the most recent twelve months of data preserves the performance advantage on a rolling basis, avoiding the gradual decay that is characteristic of static supervised models.

Third, the identification of an interior optimum for the clipping threshold suggests that production systems should treat ε as a tunable rather than a fixed parameter. A conservative default of $\varepsilon = 0.20$ delivered the best performance in the present experiments, but this value may need to be shifted downward in higher-volatility regimes or upward in more range-bound environments. Adaptive clipping schedules that respond to observed realised volatility are a natural extension that we leave for future work.

6.3 Theoretical Contributions

From a theoretical standpoint, the study contributes to three strands of literature. In RL for finance, it provides evidence that on-policy actor–critic methods with clipped updates outperform both off-policy value-based methods and unclipped policy-gradient methods on a standardised benchmark. In portfolio optimisation, it documents empirically that integrating risk into the learning reward is more effective than applying risk constraints post hoc to a return-maximising agent. In applied machine learning, it shows that supervised dimensionality reduction via LDA can provide a structured, low-noise state representation that materially improves downstream policy learning.

Taken together, these findings support the broader methodological claim that end-to-end reinforcement learning is well-suited to capital allocation problems, provided that three conditions are met: (i) policy updates are stabilised, (ii) the reward explicitly encodes the risk–return trade-off relevant to the practitioner, and (iii) the state representation is preprocessed to remove noise while preserving decision-relevant structure.

6.4 Limitations and Threats to Validity

Several caveats qualify the interpretation of the results. First, the evaluation relies on two specific datasets, and although both are of reasonable scale, results on other asset classes, geographies, or

frequency bands may differ. Cryptocurrency and foreign-exchange data exhibit heavier tails and more pronounced regime shifts than the large-capitalisation equity data that dominate the Yahoo Finance dataset, and it is plausible that the advantage of DAC-CPPO would be larger in these environments but also more sensitive to the specific choice of hyperparameters. Second, the 250-trading-day test window, while sufficient to expose the agent to at least one non-trivial shock episode, is short relative to the time horizon over which most institutional investors operate. A longer out-of-sample evaluation, ideally crossing multiple business cycles, would strengthen the empirical case.

Third, the transaction-cost model used in the back-test is simple: proportional linear costs applied symmetrically to buys and sells, with no market-impact term. In practice, large trades affect prices against the trader, particularly in less liquid instruments, and the resulting slippage can substantially erode realised returns. Fourth, the framework assumes that historical data are representative of the future, which is a strong assumption in non-stationary environments. Regime-detection modules, adaptive learning-rate schedules, and meta-RL outer loops offer routes to relaxing this assumption, but none is implemented in the present study. Finally, the benchmarks used for comparison, while broad, do not exhaust the space of competitive alternatives; direct comparison against commercial black-box systems and the more recent transformer-based trading agents that have emerged in the literature would further contextualise the findings.

6.5 Comparison with Contemporary DRL-Finance Literature

The performance profile of DAC-CPPO can usefully be compared with that reported in recent DRL-finance studies. Sharpe ratios in the 1.2–1.8 range have been reported for DDPG-based agents trained on similar historical equity universes [65][66][67], while comparative studies that report on multiple DRL baselines typically identify PPO-style methods as the most robust in out-of-sample evaluation [62][67][72]. The Sharpe of 1.91 attained by DAC-CPPO sits at the upper end of this range and above the explicit baselines that are included in the present evaluation. The methodological advances responsible for the improvement are attributable to the combination of asynchronous exploration, clipped updates, and a risk-aware reward rather than to any single ingredient.

A second observation concerns the relative performance of supervised and reinforcement learning approaches. The supervised benchmarks evaluated here (from logistic regression through Transformer) span a wide range of Sharpe values from 1.18 to 1.64, and the best supervised model (Transformer, 1.64) outperforms vanilla PPO (1.53). This finding is consistent with the broader literature, which has repeatedly noted that well-tuned supervised methods can match or exceed poorly-tuned RL agents when transaction costs are low and the state representation is adequate [44][45][46]. The distinction made by DAC-CPPO is that, when equipped with clipped updates and a risk-aware reward, the RL framework can extract additional Sharpe above and beyond the best supervised alternative—a gap of 0.27 Sharpe points that translates into meaningful economic value over multi-year horizons.

7. Conclusion

This paper introduced DAC-CPPO, a data-driven reinforcement learning framework for risk-

aware capital allocation, and subjected it to a comprehensive empirical evaluation on two publicly available financial datasets. The framework couples an asynchronous advantage actor–critic backbone with a clipped proximal policy update, feeds the agent with an LDA-compressed state representation, and trains it against a risk-adjusted reward that rewards realised Sharpe performance while penalising drawdown and turnover. Across a comprehensive benchmark set that includes buy-and-hold, mean–variance, supervised learners, and unmodified reinforcement learning baselines, the proposed agent produced a Sharpe ratio of 1.91, a cumulative return of 1.12, realised volatility of 0.14, accuracy of 97.6%, MAE of 0.074, RMSE of 0.081, and an AUC-ROC of 0.984.

The component-wise ablation confirmed that none of the architectural elements was redundant: removing LDA, the asynchronous actor–critic, the clipping mechanism, or the risk-aware reward each produced a material drop in performance. Sensitivity analysis on the clipping threshold identified an interior optimum close to the default of $\varepsilon = 0.20$, consistent with the parameter choices used in the original PPO literature and suggesting a degree of universality in the clipped update. Across monthly volatility slices, turnover measurements, and drawdown paths, DAC-CPPO demonstrated a pattern of stable, risk-contained behaviour that is aligned with operational requirements for live deployment.

Several limitations define directions for future research. First, the current framework assumes frictional transaction costs to be linear and symmetric; extending to non-linear and asymmetric cost structures (including market-impact functions and bid–ask asymmetry) would improve realism for high-turnover segments. Second, the policy in its current form is myopic in that it does not explicitly anticipate regime change; coupling DAC-CPPO with a regime-detection module or adopting a meta-learning outer loop could address this limitation. Third, the model is trained on historical data, which inherently limits its ability to handle unprecedented macroeconomic shocks; techniques such as domain randomisation and synthetic-data augmentation via generative adversarial networks offer a route to broaden the training distribution. Finally, incorporating alternative-data signals such as news sentiment, order-book dynamics, or macroeconomic releases would provide additional inputs whose informational content is not fully captured by historical price and volume alone.

The broader message of this study is that the prudent integration of three design principles—stable policy updates, risk-aware rewards, and decision-relevant state compression—can transform reinforcement learning from a promising but unstable technique into a practical tool for risk-aware capital allocation. The DAC-CPPO framework provides one realisation of this integration, and the empirical results reported here suggest that the underlying principles generalise beyond the specific datasets considered. As financial markets continue to grow in complexity and data availability, frameworks of this form are likely to become an increasingly standard component of the quantitative asset manager's toolkit.

Acknowledgement

The authors gratefully acknowledge the constructive comments of three anonymous reviewers, whose suggestions substantially improved the clarity, structure and rigour of the manuscript. The authors also thank the School of Intelligent Logistics and Supply Chain at Sichuan Vocational and Technical College for providing the computational resources that supported the experiments.

Any remaining errors or omissions are the sole responsibility of the authors.

References

Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1), 77 – 91. DOI: 10.1111/j.1540-6261.1952.tb01525.x

Markowitz, H. (1959). Portfolio Selection: Efficient Diversification of Investments. New York: John Wiley & Sons. DOI: 10.2307/2975974

Fama, E. F. (1965). The behavior of stock-market prices. The Journal of Business, 38(1), 34 – 105. DOI: 10.1086/294743

Mandelbrot, B. (1963). The variation of certain speculative prices. The Journal of Business, 36(4), 394–419. DOI: 10.1086/294632

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. Quantitative Finance, 1(2), 223–236. DOI: 10.1080/713665670

Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. The Review of Financial Studies, 15(4), 1137–1187. DOI: 10.1093/rfs/15.4.1137

Campbell, J. Y., & Viceira, L. M. (2002). Strategic Asset Allocation: Portfolio Choice for Long-Term Investors. Oxford University Press. DOI: 10.1093/0198296940.001.0001

Brandt, M. W. (2010). Portfolio choice problems. In Handbook of Financial Econometrics, Volume 1: Tools and Techniques (pp. 269–336). Elsevier. DOI: 10.1016/B978-0-444-50897-3.50008-0

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. IEEE Transactions on Neural Networks, 9(5), 1054. DOI: 10.1109/TNN.1998.712192

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529 – 533. DOI: 10.1038/nature14236

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2016). Continuous control with deep reinforcement learning. International Conference on Learning Representations (ICLR). DOI: 10.48550/arXiv.1509.02971

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489. DOI: 10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354 – 359. DOI: 10.1038/nature24270

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8(3), 229–256. DOI: 10.1007/BF00992696

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. arXiv preprint. DOI: 10.48550/arXiv.1312.5602

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1). DOI: 10.1609/aaai.v30i1.10295

Kakade, S. M. (2001). A natural policy gradient. Advances in Neural Information Processing Systems, 14, 1531–1538. DOI: 10.5555/2980539.2980738

Peters, J., & Schaal, S. (2008). Reinforcement learning of motor skills with policy gradients. Neural Networks, 21(4), 682–697. DOI: 10.1016/j.neunet.2008.02.003

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint. DOI: 10.48550/arXiv.1707.06347

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. Proceedings of the 32nd International Conference on Machine Learning (ICML), 1889–1897. DOI: 10.48550/arXiv.1502.05477

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x

Martinez, A. M., & Kak, A. C. (2001). PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2), 228–233. DOI: 10.1109/34.908974

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance, 19(3), 425–442. DOI: 10.1111/j.1540-6261.1964.tb02865.x

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3–56. DOI: 10.1016/0304-405X(93)90023-5

Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of Finance, 52(1), 57–82. DOI: 10.1111/j.1540-6261.1997.tb03808.x

Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case. The Review of Economics and Statistics, 51(3), 247–257. DOI: 10.2307/1926560

Merton, R. C. (1971). Optimum consumption and portfolio rules in a continuous-time model. Journal of Economic Theory, 3(4), 373–413. DOI: 10.1016/0022-0531(71)90038-X

Michaud, R. O. (1989). The Markowitz optimization enigma: Is 'optimized' optimal? Financial Analysts Journal, 45(1), 31–42. DOI: 10.2469/faj.v45.n1.31

Chopra, V. K., & Ziemba, W. T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. The Journal of Portfolio Management, 19(2), 6 – 11. DOI: 10.3905/jpm.1993.409440

Ledoit, O., & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. The Journal of Portfolio Management, 30(4), 110–119. DOI: 10.3905/jpm.2004.110

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance, 10(5), 603–621. DOI: 10.1016/S0927-5398(03)00007-0

Black, F., & Litterman, R. (1992). Global portfolio optimization. Financial Analysts Journal, 48(5), 28–43. DOI: 10.2469/faj.v48.n5.28

Avramov, D., & Zhou, G. (2010). Bayesian portfolio analysis. Annual Review of Financial Economics, 2, 25–47. DOI: 10.1146/annurev-financial-120209-133947

Michaud, R. O., & Michaud, R. O. (2008). Efficient asset management: A practical guide to stock portfolio optimization and asset allocation. Oxford University Press. DOI: 10.1093/oso/9780195331912.001.0001

Goldfarb, D., & Iyengar, G. (2003). Robust portfolio selection problems. Mathematics of Operations Research, 28(1), 1–38. DOI: 10.1287/moor.28.1.1.14260

Ceria, S., & Stubbs, R. A. (2006). Incorporating estimation errors into portfolio selection: Robust portfolio construction. Journal of Asset Management, 7(2), 109–127. DOI: 10.1057/palgrave.jam.2240207

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987–1007. DOI: 10.2307/1912773

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics,

31(3), 307–327. DOI: 10.1016/0304-4076(86)90063-1

Longin, F., & Solnik, B. (2001). Extreme correlation of international equity markets. The Journal of Finance, 56(2), 649–676. DOI: 10.1111/0022-1082.00340

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273. DOI: 10.1093/rfs/hhaa009

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3–12. DOI: 10.1002/asmb.2209

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689 – 702. DOI: 10.1016/j.ejor.2016.10.031

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654 – 669. DOI: 10.1016/j.ejor.2017.11.054

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005 – 2019. Applied Soft Computing, 90, 106181. DOI: 10.1016/j.asoc.2020.106181

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. The Journal of Finance, 75(3), 1327–1370. DOI: 10.1111/jofi.12883

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. DOI: 10.48550/arXiv.1706.03762

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint. DOI: 10.48550/arXiv.1908.10063

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), 2327–2333. DOI: 10.5555/2832415.2832572

Ban, G. Y., El Karoui, N., & Lim, A. E. B. (2018). Machine learning and portfolio optimization. Management Science, 64(3), 1136–1154. DOI: 10.1287/mnsc.2016.2644

Elmachtoub, A. N., & Grigas, P. (2022). Smart 'predict, then optimize'. Management Science, 68(1), 9–26. DOI: 10.1287/mnsc.2020.3922

Amos, B., & Kolter, J. Z. (2017). OptNet: Differentiable optimization as a layer in neural networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 136 – 145. DOI: 10.48550/arXiv.1703.00443

Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., & Kolter, J. Z. (2019). Differentiable convex optimization layers. Advances in Neural Information Processing Systems, 32. DOI: 10.48550/arXiv.1910.12430

Zhang, Z., Zohren, S., & Roberts, S. (2020). Deep learning for portfolio optimization. The Journal of Financial Data Science, 2(4), 8–20. DOI: 10.3905/jfds.2020.1.042

Butler, A., & Kwon, R. H. (2023). Integrating prediction in mean-variance portfolio optimization. Quantitative Finance, 23(3), 429–452. DOI: 10.1080/14697688.2022.2162432

Hull, J., & White, A. (2017). Optimal delta hedging for options. Journal of Banking & Finance, 82, 180–190. DOI: 10.1016/j.jbankfin.2017.05.006

Nevmyvaka, Y., Feng, Y., & Kearns, M. (2006). Reinforcement learning for optimized trade execution.

Proceedings of the 23rd International Conference on Machine Learning (ICML), 673 – 680. DOI: 10.1145/1143844.1143929

Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2016). Deep direct reinforcement learning for financial signal representation and trading. IEEE Transactions on Neural Networks and Learning Systems, 28(3), 653–664. DOI: 10.1109/TNNLS.2016.2522401

Moody, J., & Saffell, M. (2001). Learning to trade via direct reinforcement. IEEE Transactions on Neural Networks, 12(4), 875–889. DOI: 10.1109/72.935097

Jiang, Z., Xu, D., & Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. arXiv preprint. DOI: 10.48550/arXiv.1706.10059

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. Proceedings of the 33rd International Conference on Machine Learning (ICML), 1928–1937. DOI: 10.48550/arXiv.1602.01783

Babaeizadeh, M., Frosio, I., Tyree, S., Clemons, J., & Kautz, J. (2017). Reinforcement learning through asynchronous advantage actor-critic on a GPU. International Conference on Learning Representations (ICLR). DOI: 10.48550/arXiv.1611.06256

Liu, X. Y., Yang, H., Chen, Q., Zhang, R., Yang, L., Xiao, B., & Wang, C. D. (2020). FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance. arXiv preprint. DOI: 10.48550/arXiv.2011.09607

Théate, T., & Ernst, D. (2021). An application of deep reinforcement learning to algorithmic trading. Expert Systems with Applications, 173, 114632. DOI: 10.1016/j.eswa.2021.114632

Yang, H., Liu, X. Y., Zhong, S., & Walid, A. (2020). Deep reinforcement learning for automated stock trading: An ensemble strategy. Proceedings of the First ACM International Conference on AI in Finance (ICAIF), 1–8. DOI: 10.1145/3383455.3422540

Betancourt, C., & Chen, W. H. (2021). Deep reinforcement learning for portfolio management of markets with a dynamic number of assets. Expert Systems with Applications, 164, 114002. DOI: 10.1016/j.eswa.2020.114002

Sun, S., Wang, R., & An, B. (2023). Reinforcement learning for quantitative trading. ACM Transactions on Intelligent Systems and Technology, 14(3), 1–29. DOI: 10.1145/3582560

Zhang, Z., Zohren, S., & Roberts, S. (2020). Deep reinforcement learning for trading. The Journal of Financial Data Science, 2(2), 25–40. DOI: 10.3905/jfds.2020.1.030

Moody, J., & Wu, L. (1997). Optimization of trading systems and portfolios. Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr), 300 – 307. DOI: 10.1109/CIFER.1997.618944

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. Journal of Risk, 2, 21–42. DOI: 10.21314/JOR.2000.038

Tamar, A., Glassner, Y., & Mannor, S. (2015). Optimizing the CVaR via sampling. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1). DOI: 10.1609/aaai.v29i1.9561

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139–1168. DOI: 10.1111/j.1540-6261.2007.01232.x

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). Predicting returns with text data. NBER Working Paper No. 26186. DOI: 10.3386/w26186

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). High-dimensional continuous control using generalized advantage estimation. arXiv preprint. DOI: 10.48550/arXiv.1506.02438

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. Proceedings of the 35th International Conference on Machine Learning (ICML), 1861–1870. DOI: 10.48550/arXiv.1801.01290

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., & Madry, A. (2020). Implementation matters in deep policy gradients: A case study on PPO and TRPO. International Conference on Learning Representations (ICLR). DOI: 10.48550/arXiv.2005.12729

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). DOI: 10.1609/aaai.v32i1.11694

Sharpe, W. F. (1994). The Sharpe ratio. The Journal of Portfolio Management, 21(1), 49 – 58. DOI: 10.3905/jpm.1994.409501

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291. DOI: 10.2307/1914185

Ait-Sahalia, Y., & Brandt, M. W. (2001). Variable selection for portfolio choice. The Journal of Finance, 56(4), 1297–1351. DOI: 10.1111/0022-1082.00369

Guasoni, P., & Robertson, S. (2012). Portfolios and risk premia for the long run. The Annals of Applied Probability, 22(1), 239–284. DOI: 10.1214/11-AAP767