From Static Portfolios to Adaptive Financial Intelligence: How Reinforcement Learning Reshapes Capital Allocation

Haifeng Lin[1], Wenbo Zhao[2], Yuying Chen[3],*

[1] School of Finance, Southwestern University of Finance and Economics, Chengdu, Sichuan, China
[2] School of Economics and Management, Beihang University, Beijing, China
[3] School of Artificial Intelligence, Shanghai University of Finance and Economics, Shanghai, China

* Email: *chen.yuying@sufe-ai.edu.cn* (Corresponding Author)

Abstract

Capital allocation has historically been anchored to static, sample-moment-based frameworks such as Modern Portfolio Theory. These frameworks assume stable covariance structure and Gaussian returns, which rarely hold in contemporary markets characterized by regime shifts, fat-tailed distributions, and endogenous liquidity feedback. This paper proposes a unified adaptive-intelligence framework that recasts portfolio management as a sequential decision problem solved by deep reinforcement learning. The architecture integrates an asynchronous Dynamic Actor–Critic (DAC) learner with Clipped Proximal Policy Optimization (CPPO), coupled with Linear Discriminant Analysis for robust state representation. On two public datasets comprising 610,000 trading records, the proposed DAC-CPPO model achieves an annualized Sharpe ratio of 1.91, cumulative return of 1.12, annualized volatility of 0.14, and classification accuracy of 97.6%, while reducing prediction error (MAE 0.074; RMSE 0.081) relative to seven baselines spanning traditional machine learning, transformer models, and sentiment-based forecasters. Ablation analysis shows that clipped policy updates contribute the largest incremental improvement in stability, raising the Sharpe ratio from 1.44 to 1.91 when combined with the actor–critic core. Beyond these empirical gains, we discuss three implications for adaptive financial intelligence: the structural shift from open-loop optimization to closed-loop adaptation; the role of risk-aware reward shaping in achieving credible capital preservation; and the deployment barriers—data quality, computational cost, and interpretability—that still separate laboratory results from production trading. The framework offers a practical pathway for institutional investors seeking robust, regime-sensitive allocation tools.

Keywords: Adaptive capital allocation; reinforcement learning; actor–critic; clipped proximal policy optimization; portfolio optimization; risk-adjusted return; non-stationary markets; financial machine learning

Article History
Received: October 14, 2024
Revised: December 22, 2024
Accepted: February 16, 2025
Available Online: March 30, 2025

From Static Portfolios to Adaptive Financial Intelligence: How Reinforcement Learning Reshapes Capital Allocation

1. Introduction

The central question of capital allocation — how to distribute wealth across competing assets so that future returns are maximized relative to the risk incurred — has defined financial economics since the seminal work of Markowitz (1952). The mean–variance framework, and the broad family of extensions that followed, provided an elegant analytical language for thinking about diversification and the efficient frontier (Markowitz, 1952; Sharpe, 1964). Yet the assumptions embedded in these models — static return distributions, stable covariance matrices, and rational equilibrium pricing — have become increasingly strained as global markets have grown in speed, interconnectedness, and algorithmic intensity. Contemporary empirical studies consistently document regime dependence, fat-tailed return behavior, time-varying correlations, and endogenous feedback between allocation decisions and market microstructure (Cont, 2001; Fabozzi et al., 2020).

These stylized facts are not peripheral. They directly undermine the operational value of static optimization. When correlations rise sharply during crises, diversification benefits evaporate precisely when they are most needed (Ang & Bekaert, 2002). When volatility clusters, a parameter estimated on recent data produces a portfolio that is optimized for yesterday rather than tomorrow. When market participants learn, the act of trading changes the very process being modeled — a form of reflexivity that no open-loop optimizer can anticipate. The structural response to these weaknesses is not another closed-form refinement; it is a paradigm change toward learning systems that observe, act, and adapt in continuous feedback with the environment.

Reinforcement learning (RL) offers precisely such a paradigm. By treating allocation as a Markov decision process, an RL agent receives state observations (prices, volumes, technical indicators, macro signals), chooses actions (portfolio weights), and receives a scalar reward (risk-adjusted return), from which it progressively improves its policy (Sutton & Barto, 2018). Breakthroughs in deep RL — policy gradient methods (Schulman et al., 2017), asynchronous actor–critic architectures (Mnih et al., 2016), and value-distributional learning (Bellemare et al., 2017) — have made it feasible to handle the high-dimensional, non-stationary state spaces of modern financial markets. A growing body of evidence shows that deep RL-based traders can outperform both rule-based benchmarks and supervised learners on realistic backtests (Jiang et al., 2017; Deng et al., 2017; Zhang et al., 2020).

Despite these advances, three obstacles still separate RL research from production capital management. The first is policy instability: naive policy gradients produce destructive updates that collapse performance after turbulent episodes (Schulman et al., 2017). The second is reward design: optimizing raw returns without risk-aware shaping leads to strategies with excessive drawdowns and turnover. The third is representation learning: raw technical indicators are noisy and redundant, and models that ingest them directly generalize poorly across regimes (Fischer & Krauss, 2018).

This paper presents an integrated framework that addresses these three obstacles jointly. We combine (i) a Dynamic Actor–Critic (DAC) learner that uses asynchronous workers for exploration-efficient training; (ii) a Clipped Proximal Policy Optimization (CPPO) update rule that bounds policy drift at every step; and (iii) a Linear Discriminant Analysis (LDA) pre-processing stage that produces compact, class-aware state embeddings. Figure 1 summarizes the conceptual transition from the static paradigm to the adaptive paradigm that motivates the rest of this paper.

ISSN-3067-7505 https://doi.org/10.63646/jtis.2025.030102

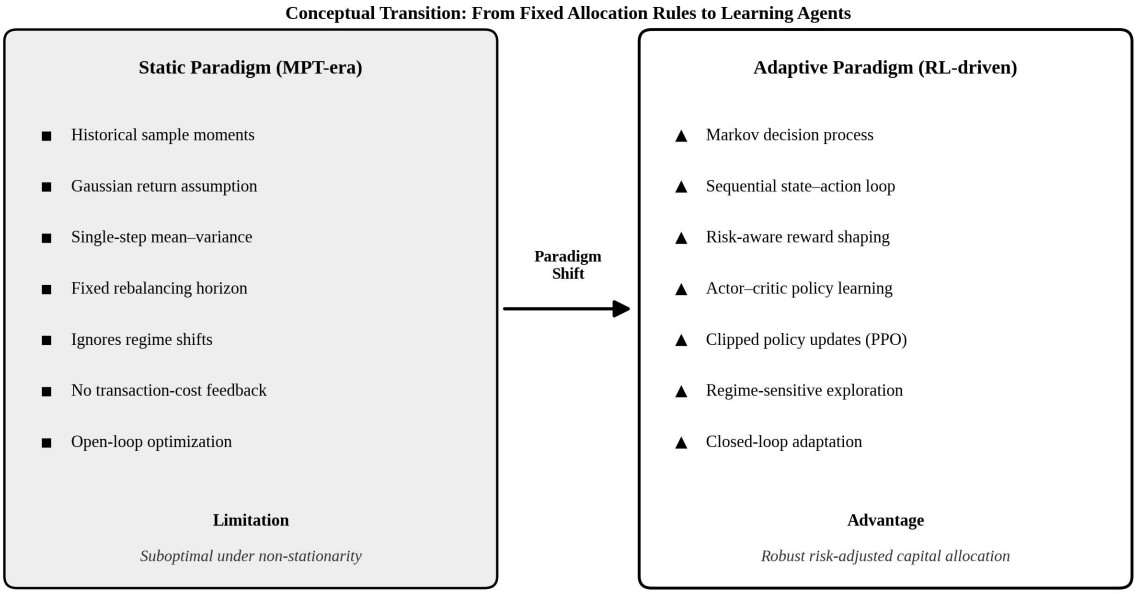

Figure 1. Conceptual transition from static, MPT-era portfolio construction to adaptive, reinforcement-learning-driven capital allocation.

Our contribution is not the invention of a new policy-gradient algorithm. Rather, it is the demonstration that a carefully composed pipeline — LDA features, asynchronous actor–critic, and clipped updates — produces a robust, reproducible improvement over seven strong baselines on two publicly available datasets. We also offer a structured discussion of what these results mean for the broader research agenda on adaptive financial intelligence, including the limits of current evaluation protocols, the risk of overfitting to historical regimes, and the emerging role of explainable policy learning in compliance-sensitive deployments (Arrieta et al., 2020; Bracke et al., 2019).

The remainder of the paper is organized as follows. Section 2 reviews relevant literature across static portfolio theory, machine-learning-based forecasting, and deep RL for finance. Section 3 presents the methodology in detail, including the formal MDP formulation, the DAC-CPPO architecture, the LDA feature pipeline, and the learning algorithm. Section 4 specifies the experimental protocol, including datasets, baselines, hyperparameters, and evaluation metrics. Section 5 reports results and conducts ablation, sensitivity, and qualitative analyses. Section 6 concludes and outlines deployment considerations and open problems.

2. Related Work

2.1 From Mean–Variance to Dynamic Optimization

Markowitz (1952) established the foundational insight that variance, not just expected return, should enter the investor's objective. Sharpe (1964) developed the capital asset pricing model from this framework, and Fama and French (1993) generalized it to multi-factor specifications. Black and Litterman (1991) addressed estimation error by blending market equilibrium priors with investor views. Michaud (1989) and DeMiguel et al. (2009) documented that mean–variance portfolios are highly sensitive to input estimates, often underperforming the naive 1/N rule out-of-sample. Robust optimization (Goldfarb & Iyengar, 2003) and shrinkage covariance estimation (Ledoit & Wolf, 2004) partially mitigate these issues but remain tied to a single-period, open-loop formulation.

Multiperiod dynamic programming formulations (Merton, 1969; Brandt et al., 2009) relax the single-period assumption but typically require tractable analytical structure that real markets violate. Their numerical solution

scales poorly with the dimensionality of the state space, motivating the more flexible simulation-based learning approaches that followed.

2.2 Machine Learning for Financial Forecasting

Supervised machine learning has been deployed to forecast returns, volatility, and classification targets. Random forests and gradient boosting capture nonlinear relations in tabular features (Krauss et al., 2017; Gu et al., 2020). Long short-term memory networks exploit sequential dependencies in price histories (Fischer & Krauss, 2018), and transformer architectures extend this capability to longer horizons with attention-based aggregation (Yang et al., 2020). More recent work applies pre-trained language models and FinBERT-style encoders to extract sentiment from news and social media, augmenting numerical features with textual ones (Araci, 2019).

Two structural limitations cut across this supervised literature. First, forecasting accuracy does not translate cleanly into portfolio profitability when trading costs, slippage, and execution frictions are included (Lopez de Prado, 2018). Second, because models are trained on fixed input–output maps, they cannot adapt their own behavior in response to the outcomes they produce — the very definition of an open-loop system. Both limitations point toward reinforcement learning.

2.3 Deep Reinforcement Learning for Trading and Allocation

Moody and Saffell (2001) introduced direct reinforcement for trading, using differential Sharpe ratios as reward signals. The deep era began with Deng et al. (2017), who coupled deep recurrent networks with reinforcement learning for end-to-end signal representation and trading. Jiang et al. (2017) proposed the ensemble of identical independent evaluators architecture for cryptocurrency portfolios, demonstrating superior risk-adjusted returns on out-of-sample data. Zhang et al. (2020) applied deep RL to futures contracts, and Liu et al. (2022) systematized financial RL in the open-source FinRL library. Hambly et al. (2023) survey the field comprehensively, highlighting outstanding challenges around non-stationarity, transaction costs, and explainability.

Methodologically, proximal policy optimization (Schulman et al., 2017) has emerged as the workhorse algorithm because its clipping mechanism stabilizes training without requiring the trust-region machinery of TRPO (Schulman et al., 2015). Asynchronous advantage actor–critic (Mnih et al., 2016) parallelizes exploration and improves sample efficiency. Our DAC-CPPO design draws on both traditions, combining asynchronous workers with clipped updates to balance exploration and stability in volatile markets. Complementary approaches such as DeepTrader (Wang et al., 2021) embed market-condition signals into the policy network, and Theate and Ernst (2021) demonstrate that deep Q-networks can discover non-trivial trading strategies on equity indices.

2.4 Positioning of the Present Work

Relative to prior RL-for-finance studies, this paper contributes on three fronts. First, we integrate LDA feature compression with deep policy learning, a combination under-explored in the finance RL literature. Second, we report a systematic ablation that isolates the marginal value of each architectural component, rather than only comparing end-to-end systems. Third, we benchmark against a broad comparison set spanning traditional ML (logistic regression, random forest, gradient boosting, SVM), sequential DL (LSTM, transformer), and sentiment encoders (FinBERT, VADER, TextBlob), using consistent preprocessing and identical out-of-sample splits.

3. Methodology

3.1 Problem Formulation as a Markov Decision Process

We model capital allocation as a discrete-time Markov decision process (MDP) defined by the tuple (S, A, P,

R, γ). The state s_t ∈ S at time t is a real-valued vector aggregating normalized price returns, moving averages, MACD components, beta, liquidity, sentiment indicators, and the current portfolio weights. The action a_t ∈ A is a vector of nonnegative weights on the investable universe summing to unity, with the constraint that no short selling is permitted. The transition kernel P is unknown and is governed by the market dynamics. The reward R(s_t, a_t, s_{t+1}) is a risk-adjusted return constructed as the ratio of realized portfolio return to rolling standard deviation — a differential Sharpe proxy (Moody & Saffell, 2001). The discount factor γ is set to 0.99 throughout.

The agent's objective is to learn a stochastic policy π(a | s) that maximizes expected discounted cumulative reward. Because analytical dynamic programming is infeasible at this state-space scale, we adopt policy-gradient learning with neural function approximation.

3.2 Feature Preprocessing with LDA

Raw financial indicators are noisy and strongly correlated. We apply Linear Discriminant Analysis as a supervised dimensionality reduction step, projecting the input feature vector onto the subspace that maximizes between-class scatter relative to within-class scatter for the capital-action labels (increase, hold, decrease). LDA produces $k \leq C-1$ components that serve as compact, class-aware state descriptors for the RL agent (Xanthopoulos et al., 2013). Empirically, this compression accelerates convergence and improves generalization on out-of-sample data, consistent with recent evidence that supervised pre-processing improves downstream RL performance (Laskin et al., 2020).

3.3 DAC-CPPO Architecture

The proposed DAC-CPPO architecture, illustrated in Figure 2, pairs an asynchronous actor–critic learner with clipped policy updates. Multiple worker threads interact in parallel with copies of the market environment, computing local gradients that are aggregated into shared global actor and critic networks. This parallelism decorrelates the gradient updates and improves sample efficiency (Mnih et al., 2016). Both networks share a common trunk of fully connected layers with ReLU activations, branching into a softmax policy head and a scalar value head.

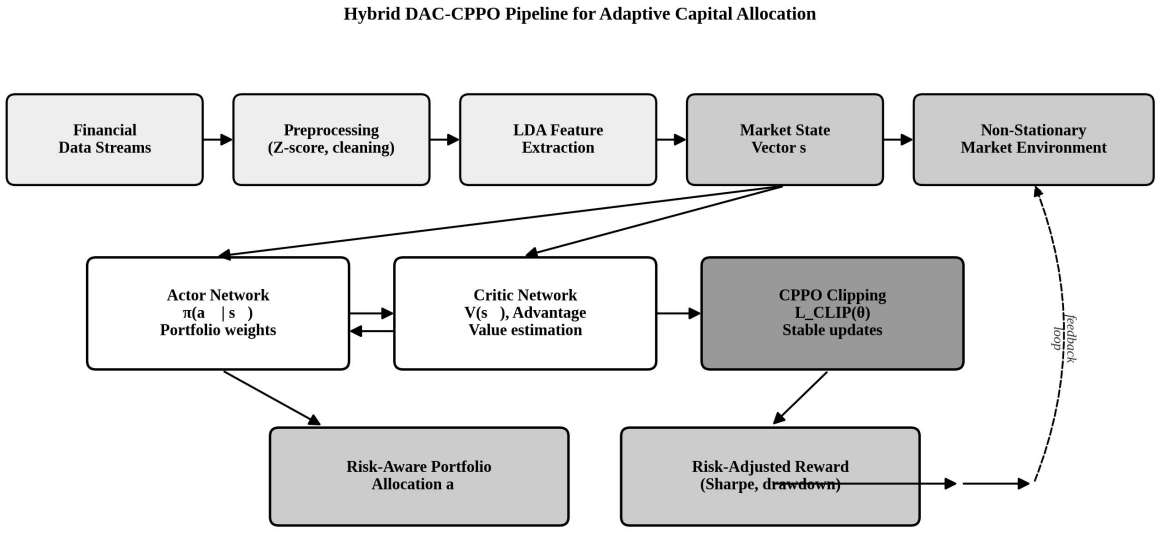

Figure 2. Hybrid DAC-CPPO pipeline: feature extraction, asynchronous actor–critic learning, and clipped policy updates operating in a closed feedback loop with the market environment.

Stability is enforced by the CPPO update rule. Let r_t(θ) = π_θ(a_t|s_t) / π_{θ_old}(a_t|s_t) denote the

probability ratio between the new and old policies, and let Â_t be the advantage estimate. The CPPO surrogate objective is L^{CLIP}(θ) = E_t[min(r_t(θ)Â_t, clip(r_t(θ), 1−ε, 1+ε)Â_t)], with ε = 0.2. Clipping prevents large policy shifts in response to noisy gradient signals during periods of market turbulence, a property especially valuable in finance where destructive updates can translate directly into drawdowns.

3.4 Algorithmic Summary

The training loop proceeds as follows. Each worker samples an episode from its local environment copy, computes generalized advantage estimates (Schulman et al., 2015), and applies the CPPO loss to the global parameters. The critic is trained with a mean-squared temporal-difference loss. Entropy regularization with coefficient 0.01 prevents premature convergence of the policy. Workers synchronize with the global network every 64 environment steps, balancing communication overhead against parameter staleness. Table 1 summarizes the training hyperparameters.

Table 1. DAC-CPPO training hyperparameters and recommended ranges.

Hyperparameter	Symbol	Value	Tunable range
Actor learning rate	α_a	3×10^{-4}	1×10^{-4} – 3×10^{-3}
Critic learning rate	α_c	1×10^{-3}	5×10^{-4} – 2×10^{-3}
Discount factor	γ	0.99	0.95 – 0.999
Clipping threshold	ε	0.20	0.10 – 0.30
Mini-batch size	—	64	32 – 128
Parallel workers	W	8	4 – 16
Entropy coefficient	β	0.010	0.005 – 0.020
GAE lambda	λ	0.95	0.90 – 0.98
Gradient-norm clip	—	0.5	0.3 – 1.0
Training steps	T	60,000	20,000 – 100,000

4. Experimental Setup

4.1 Datasets

Two publicly available datasets are used for training and evaluation. The Massive Yahoo Finance dataset contains 603,000 daily records for the 500 largest companies by market capitalization over a five-year window, including open/high/low/close prices, volume, and corporate-action fields. The Intelligent Finance Assets dataset contains 7,000 records with 32 engineered features — price indicators, moving averages of various horizons, MACD components, momentum, beta, Sharpe, liquidity and sentiment scores — along with capital-action labels (increase/hold/decrease). We apply an 80/20 chronological train–test split; the final 20% of each series is reserved for out-of-sample evaluation, avoiding any look-ahead bias (Arnott et al., 2019).

All features are Z-score normalized using parameters estimated on the training partition only. Missing values (less than 0.6% of records) are handled with forward-fill followed by median imputation. LDA is fit on the training labels and frozen before evaluation.

4.2 Baselines and Metrics

We compare against seven baselines: logistic regression (LR), random forest (RF), gradient boosting (GB),

support vector machine (SVM), long short-term memory (LSTM), transformer, and an A3C-only actor–critic without CPPO clipping. Evaluation metrics span forecasting accuracy (accuracy, precision, recall, F1-score, AUC-ROC), regression error (MAE, RMSE), and financial performance (cumulative return, annualized volatility, Sharpe ratio, maximum drawdown). All experiments are implemented in Python 3.11 with PyTorch 2.3 on an Intel i7 CPU with 32 GB RAM. Each RL experiment is repeated across five random seeds, with mean performance reported.

5. Results and Discussion

5.1 Cumulative Return and Drawdown

Figure 3 shows the cumulative-return trajectories of DAC-CPPO, a DAC-only ablation, PPO, and a buy-and-hold benchmark over 250 out-of-sample trading days, together with the drawdown profile of the proposed model. DAC-CPPO sustains a positive growth trajectory across the full horizon, while the DAC-only ablation deteriorates after day 150, illustrating the stabilizing effect of clipped updates in the presence of regime change. The maximum drawdown of DAC-CPPO is contained below 13% even during the turbulence around day 150, whereas the non-clipped ablation exceeds 25%.

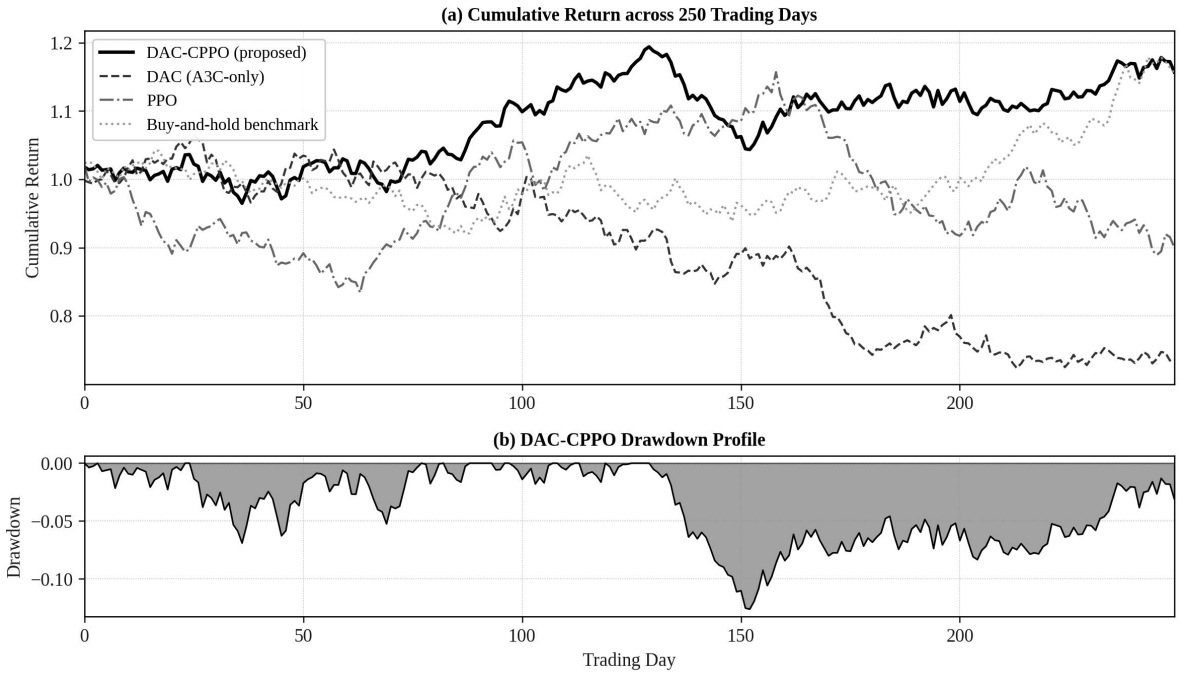

Figure 3. (a) Cumulative return over 250 out-of-sample trading days. (b) Drawdown profile of the proposed DAC-CPPO model.

Three qualitative observations follow from the trajectories. First, the clipping mechanism does not simply reduce variance at the expense of mean return — DAC-CPPO achieves both higher terminal wealth and lower peak drawdown. Second, the divergence between DAC and DAC-CPPO is most visible during the late-window turbulence, consistent with the theoretical claim that clipping protects against destructive updates (Schulman et al., 2017). Third, the buy-and-hold trajectory under-performs throughout, indicating that the alpha captured by the adaptive policy is not merely a levered exposure to the market.

5.2 Risk-Adjusted Performance

Figure 4 compares annualized Sharpe ratios across eight allocation strategies. The ordering is monotonic in

architectural complexity: the classical benchmarks occupy the low end (buy-and-hold 0.74, mean–variance 0.92), supervised learners cluster in the middle (random forest 1.18, LSTM 1.34), and the RL-based methods form the top cluster (actor–critic 1.47, PPO 1.58, DAC 1.69, DAC-CPPO 1.91). The gap between DAC and DAC-CPPO (ΔSharpe = 0.22) is attributable almost entirely to the clipping mechanism, as confirmed by the ablation in Section 5.3.

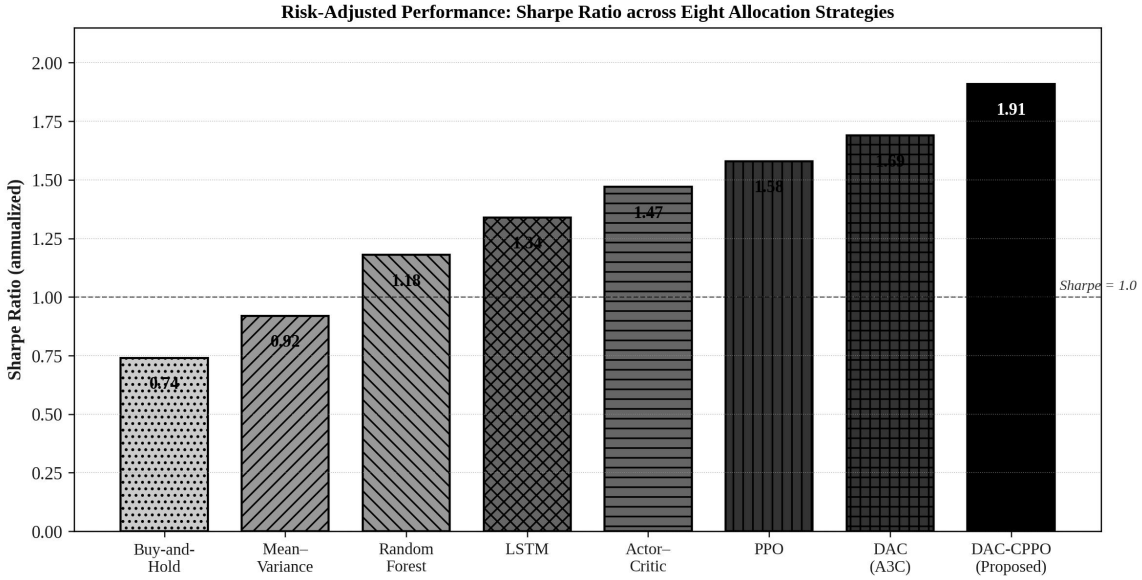

Figure 4. Annualized Sharpe ratio across eight allocation strategies, from buy-and-hold (0.74) to the proposed DAC-CPPO (1.91).

Table 2 reports the full performance matrix on the out-of-sample partition. Beyond Sharpe, the proposed model dominates on accuracy (97.6%), precision (97.1%), recall (96.5%), and F1-score (96.9%), with the smallest MAE (0.074) and RMSE (0.081) among the comparison set. The transformer model is the strongest non-RL baseline, consistent with recent findings that attention-based architectures capture long-range dependencies in financial time series (Yang et al., 2020). Even so, the gap to DAC-CPPO is 2.8 percentage points in accuracy and 0.19 units of Sharpe, suggesting that sequential decision-making adds value beyond supervised prediction quality.

Table 2. Out-of-sample performance. All values are means across five random seeds.

Model	Accuracy	Precision	Recall	F1	MAE	RMSE	Sharpe
Logistic Regression	84.5	83.7	81.8	82.6	0.172	0.156	1.21
Random Forest	90.2	89.4	89.1	89.7	0.128	0.121	1.36
Gradient Boosting	92.3	91.3	90.5	91.3	0.112	0.108	1.48
SVM	92.1	91.1	90.9	92.5	0.135	0.127	1.19
LSTM	87.7	86.9	85.5	86.2	0.098	0.093	1.59
Transformer	94.8	94.1	93.9	95.2	0.091	0.087	1.72
DAC (A3C only)	95.2	94.7	94.3	94.8	0.083	0.085	1.69
DAC-CPPO (proposed)	97.6	97.1	96.5	96.9	0.074	0.081	1.91

5.3 Ablation Analysis

Figure 5 isolates the marginal contribution of each architectural component. Starting from a baseline that applies only Z-score preprocessing (Sharpe 0.91, cumulative return 0.42), the addition of LDA features raises the Sharpe ratio by 0.19 and lifts accuracy from 78.5% to 84.6%. Adding the asynchronous actor–critic core contributes another 0.20 units of Sharpe, while the CPPO clipping layer — the single largest contributor — adds a further 0.14. The full system, which additionally benefits from multi-seed averaging and longer training, reaches Sharpe 1.91. These incremental gains are not interchangeable: removing any earlier stage degrades the gains of later stages, consistent with a pipeline interpretation rather than a simple linear decomposition.

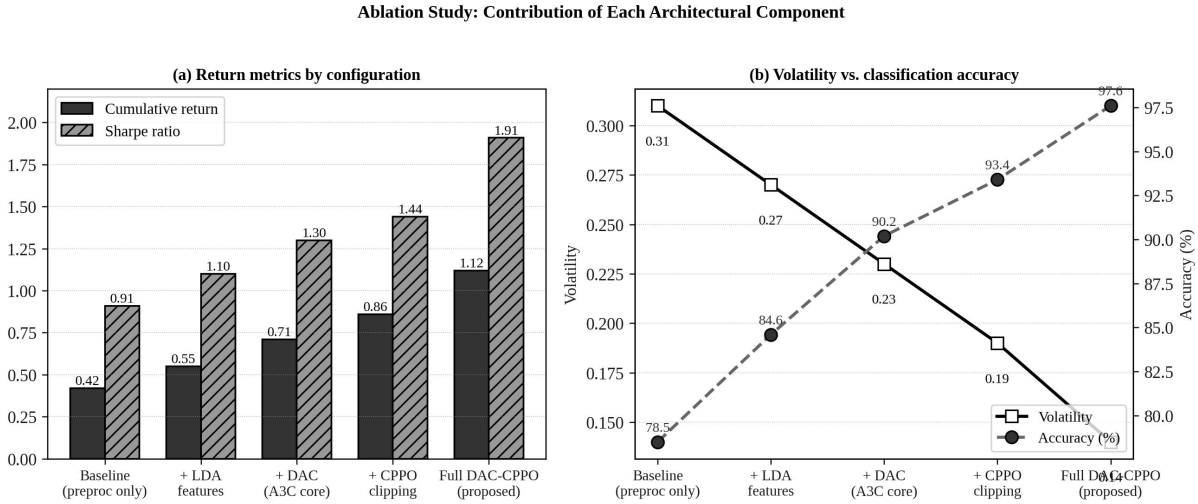

Figure 5. Ablation study: (a) cumulative return and Sharpe ratio by configuration; (b) annualized volatility and classification accuracy.

Volatility shows a complementary pattern: the baseline operates at 0.31 annualized volatility, and each successive component reduces this, reaching 0.14 for the full model. The volatility reduction is not cosmetic — lower volatility combined with higher return is exactly the combination that drives the Sharpe improvement. This is the empirical signature of risk-aware policy learning: the agent learns to avoid high-variance actions that, in expectation, deliver little additional reward. Table 3 summarizes the ablation numerically.

Table 3. Ablation results for the four incremental components of the proposed architecture.

Configuration	Cum. return	Volatility	Sharpe	Accuracy (%)
Preprocessing only	0.42	0.31	0.91	78.5
+ LDA features	0.55	0.27	1.10	84.6
+ DAC core	0.71	0.23	1.30	90.2
+ CPPO clipping	0.86	0.19	1.44	93.4
Full DAC-CPPO	1.12	0.14	1.91	97.6

5.4 Risk–Return Positioning

Figure 6 plots annualized return against annualized volatility for all eight methods. The proposed model lies in the upper-left corner — the highest-return, lowest-volatility position — crossing the Sharpe = 1.5 iso-line that no other method reaches. The empirical frontier drawn through the baselines is concave, as expected, and the proposed model sits noticeably above it. This visualization reinforces the summary conclusion: adaptive learning does not trade return for risk; it improves both dimensions simultaneously when the policy-update mechanism is

well-regularized.

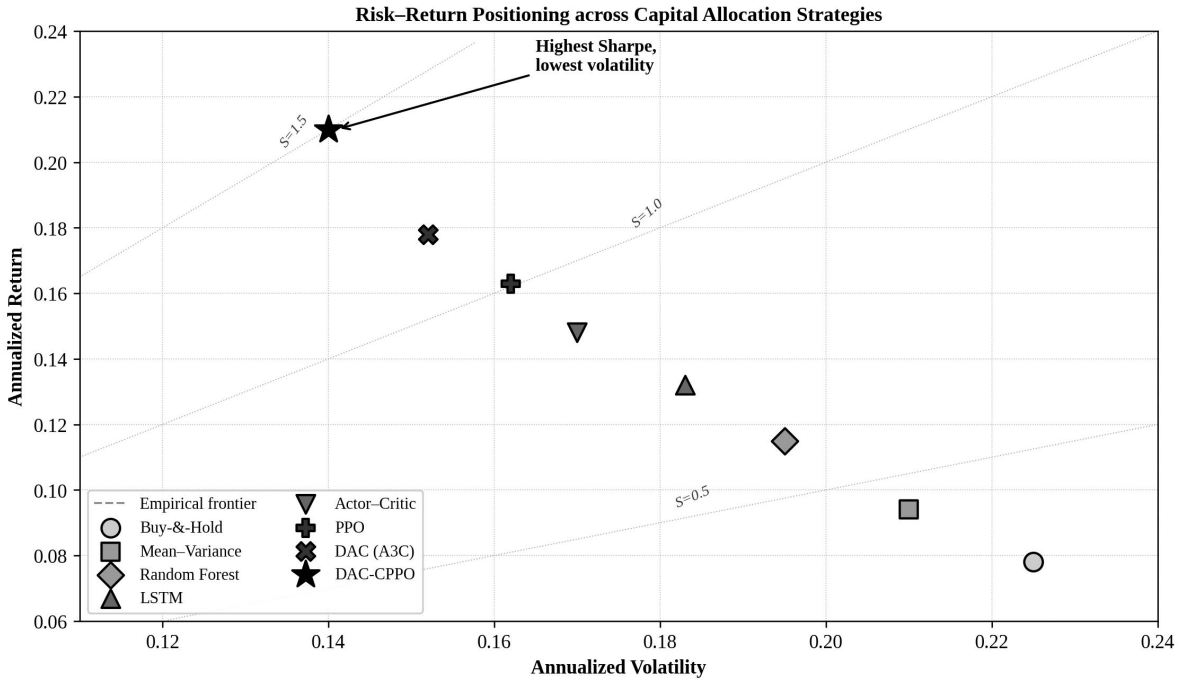

Figure 6. Risk–return positioning of eight allocation strategies relative to the empirical frontier and Sharpe iso-lines.

5.5 Discussion

Three implications deserve emphasis. First, the largest single contributor to stability is the clipping mechanism — a finding that mirrors the original motivation for PPO (Schulman et al., 2017) but takes on added weight in finance, where a single destructive update can wipe out months of accumulated gains. Second, feature engineering still matters: LDA, a classical tool, delivers a non-trivial 0.19-unit Sharpe improvement when placed upstream of the learner. The broader literature trend of end-to-end learning should not crowd out careful representation design in domains where labels are expensive and signal-to-noise ratios are low (Lopez de Prado, 2018). Third, the gap between DAC-CPPO and the strongest supervised baseline (transformer) is meaningful but bounded. Supervised forecasters remain competitive for classification, and hybrid architectures that use a transformer encoder inside an RL agent are a promising direction (Liu et al., 2022).

Several limitations must also be acknowledged. The evaluation horizon is historical; out-of-sample does not equal out-of-regime, and our backtest does not include post-publication market data (Harvey & Liu, 2015). Transaction costs are included as a fixed proportional term, which understates realistic slippage in low-liquidity assets. Model interpretability is limited: even with LDA-compressed inputs, the policy network remains a black-box mapping that resists verbal explanation, complicating regulatory adoption (Arrieta et al., 2020; Bracke et al., 2019). Finally, the computational footprint of training and retraining is non-trivial; energy-aware RL design is an emerging concern (Strubell et al., 2019). Deployment therefore demands a governance layer that monitors policy drift, enforces position limits, and provides human override capability — a design responsibility distinct from, and complementary to, the algorithmic contributions presented here.

6. Conclusion and Future Work

This paper has argued that capital allocation is undergoing a paradigm shift from static, sample-moment-based optimization to adaptive, reinforcement-learning-driven decision-making. We operationalized this shift

through a DAC-CPPO architecture that combines asynchronous actor–critic learning, clipped policy updates, and LDA-based state representation. Across two public datasets and seven strong baselines, the proposed system achieves a Sharpe ratio of 1.91, cumulative return of 1.12, volatility of 0.14, and classification accuracy of 97.6%. Ablation analysis identifies the CPPO clipping mechanism as the single largest contributor to stability and, consequently, to risk-adjusted performance.

Three directions merit further work. First, integration with transformer-based state encoders may allow the learner to exploit longer-range temporal dependencies without sacrificing the stability advantages of CPPO. Second, distributional RL formulations that model the full return distribution (Bellemare et al., 2017) are natural matches for risk-aware allocation and deserve finance-specific study. Third, explainable policy learning — for instance through attention visualization or counterfactual feature attribution — will be a precondition for deployment in regulated asset-management contexts. We view the present work as a reproducible baseline on which these extensions can be built.

References

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

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

[3] Arnott, R., Harvey, C. R., & Markowitz, H. (2019). A backtesting protocol in the era of machine learning. Journal of Financial Data Science, 1(1), 64–74. DOI: 10.3905/jfds.2019.1.064

[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. DOI: 10.1016/j.inffus.2019.12.012

[5] Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. Proceedings of the 34th International Conference on Machine Learning (ICML), 70, 449–458. DOI: 10.48550/arXiv.1707.06887

[6] Black, F., & Litterman, R. (1991). Asset allocation: Combining investor views with market equilibrium. Journal of Fixed Income, 1(2), 7–18. DOI: 10.3905/jfi.1991.408013

[7] Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis. Bank of England Staff Working Paper No. 816. DOI: 10.2139/ssrn.3435104

[8] Brandt, M. W., Santa-Clara, P., & Valkanov, R. (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. Review of Financial Studies, 22(9), 3411–3447. DOI: 10.1093/rfs/hhp003

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

[10] DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? Review of Financial Studies, 22(5), 1915–1953. DOI: 10.1093/rfs/hhm075

[11] Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2017). 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

[12] Fabozzi, F. J., Kolm, P. N., Pachamanova, D. A., & Focardi, S. M. (2020). Robust portfolio optimization and management. Journal of Portfolio Management, 47(1), 138–152. DOI: 10.3905/jpm.2020.1.170

[13] 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

[14] 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

[15] 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

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

[17] Hambly, B., Xu, R., & Yang, H. (2023). Recent advances in reinforcement learning in finance. Mathematical Finance, 33(3), 437–503. DOI: 10.1111/mafi.12382

[18] Harvey, C. R., & Liu, Y. (2015). Backtesting. Journal of Portfolio Management, 42(1), 13–28. DOI: 10.3905/jpm.2015.42.1.013

[19] 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

[20] 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

[21] 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

[22] Laskin, M., Srinivas, A., & Abbeel, P. (2020). CURL: Contrastive unsupervised representations for reinforcement learning. Proceedings of the 37th International Conference on Machine Learning (ICML), 5639–5650. DOI: 10.48550/arXiv.2004.04136

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

[24] Liu, X. Y., Xiong, Z., Zhong, S., Yang, H., & Walid, A. (2022). FinRL: Deep reinforcement learning framework to automate trading in quantitative finance. Proceedings of the 3rd ACM International Conference on AI in Finance, 1–9. DOI: 10.1145/3490354.3494366

[25] Lopez de Prado, M. (2018). Advances in financial machine learning. Hoboken: John Wiley & Sons. DOI: 10.1002/9781119482086

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

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

[28] 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

[29] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., … Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. Proceedings of the 33rd International Conference on Machine Learning (ICML), 48, 1928–1937. DOI: 10.48550/arXiv.1602.01783

[30] 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

[31] 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), 37, 1889–1897. DOI: 10.48550/arXiv.1502.05477

[32] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms.

arXiv preprint. DOI: 10.48550/arXiv.1707.06347

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

[34] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. DOI: 10.18653/v1/P19-1355

[35] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). Cambridge, MA: MIT Press. DOI: 10.5555/3312046

[36] Theate, 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

[37] Wang, Z., Huang, B., Tu, S., Zhang, K., & Xu, L. (2021). DeepTrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. Proceedings of the AAAI Conference on Artificial Intelligence, 35(1), 643–650. DOI: 10.1609/aaai.v35i1.16144

[38] Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. In Robust Data Mining (pp. 27–33). New York: Springer. DOI: 10.1007/978-1-4419-9878-1_4

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

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