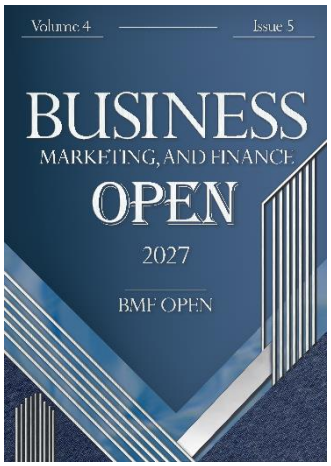
Robust Portfolio Optimization Considering Investor Preferences Using Maenhout Preferences and Reinforcement Learning

Mohammad Mehdi Faraghzadeh [1], Mehrzad Minouei [2,*] and Azadeh Mehrani [3]



[1] Department of Finance and Accounting, SR.C., Islamic Azad University, Tehran, Iran;
[2] Department of Management, CT.C., Islamic Azad University, Tehran, Iran;
[3] Department of Financial Management, Nos.C., Islamic Azad University, Noshahr, Iran;

* Correspondence: Mehrzad@iau.ac.ir

Citation: Faraghzadeh, M. M., Minouei, M., & Mehrani, A. (2027). Robust Portfolio Optimization Considering Investor Preferences Using Maenhout Preferences and Reinforcement Learning. *Business, Marketing, and Finance Open*, 4(5), 1-26.

Received: 15 April 2026
Revised: 17 July 2026
Accepted: 24 July 2026
Initial Publication: 26 July 2026
Final Publication: 01 September 2027

Abstract: This study develops a robust and adaptive portfolio optimization framework that integrates Maenhout's ambiguity-averse preference structure with Proximal Policy Optimization (PPO) for dynamic asset allocation in the Tehran Stock Exchange. The model is designed to address two core challenges in emerging markets: parameter estimation error and model misspecification. To do so, the reward function combines expected return with explicit penalties for tail risk and downside risk, including Conditional Value-at-Risk (CVaR) and the Sortino ratio, so that the learned policy is not only return-seeking but also resilient under adverse market conditions. Using daily data from actively traded firms over the 2011–2025 period, the empirical evaluation shows that the proposed framework consistently outperforms classical benchmarks in risk-adjusted performance, downside protection, and stability across changing regimes. In particular, the Maenhout–PPO strategy achieves higher Sharpe and Sortino ratios, lower maximum drawdown, and more controlled tail losses than competing allocation rules. The results indicate that classical mean–variance logic is not sufficiently robust for a market characterized by nonstationarity, inflation pressure, liquidity concentration, and abrupt policy shocks. By contrast, the proposed framework offers a more realistic decision architecture for portfolio management under uncertainty, demonstrating that combining robust control principles with reinforcement learning can materially improve investment performance in emerging-market settings.

Keywords: Robust Portfolio Optimization; Ambiguity Aversion; Model Uncertainty; Reinforcement Learning; Proximal Policy Optimization; Tehran Stock Exchange.

1. Introduction

Portfolio optimization is a foundational problem in financial economics because investors must continually allocate limited capital among assets with heterogeneous expected returns, volatility patterns, dependence structures, liquidity conditions, and exposure to systemic shocks. The classical formulation of this problem assumes that the relevant statistical characteristics of asset returns can be estimated with sufficient precision and that the probability model governing market outcomes remains reasonably stable over the investment horizon. In actual financial markets, however, portfolio decisions are made under conditions in which expected returns are difficult to forecast, covariance matrices change over time, return distributions exhibit asymmetry and heavy tails, and structural breaks can render historical estimates unreliable. These challenges have become more pronounced as financial systems have grown more interconnected and as information, policy shocks, speculative behavior, and

technological developments have accelerated changes in market dynamics. Consequently, portfolio management can no longer be treated solely as a static exercise in selecting weights from a known return distribution; it must instead be conceptualized as a sequential decision problem under risk, model uncertainty, and changing market regimes [1, 2].

Traditional portfolio optimization is especially sensitive to estimation error. Small changes in estimated means, variances, or correlations may produce substantial changes in supposedly optimal portfolio weights, frequently generating concentrated and unstable allocations. Although theoretically optimized portfolios should outperform simpler alternatives, their realized out-of-sample performance may deteriorate when estimated parameters deviate from their true values. The comparative evidence reported by DeMiguel et al. demonstrates that sophisticated optimization strategies often fail to consistently outperform the simple equal-weighted $1/N$rule after estimation error is considered [3]. This finding has important implications for applied portfolio management: mathematical optimality under estimated parameters does not necessarily imply economic superiority under genuinely uncertain conditions. In many cases, the optimization process magnifies noisy forecasts by allocating excessive weight to assets with apparently favorable but statistically fragile return estimates. Therefore, a credible portfolio framework must account not only for observable volatility but also for uncertainty regarding the parameters and models used to describe market behavior.

This distinction reflects the difference between risk and ambiguity. Risk generally refers to variation in outcomes governed by an accepted probability distribution, whereas ambiguity concerns uncertainty about whether the assumed distribution, parameters, or data-generating process are valid. Investors may know that returns are volatile while simultaneously lacking confidence in the model used to quantify that volatility. Structural uncertainty can arise from regime changes, policy interventions, shifts in investor sentiment, market microstructure constraints, macroeconomic instability, and unanticipated disruptions that alter the relationships inferred from historical data. A systematic review of portfolio optimization under structural uncertainty emphasizes that model misspecification is not a secondary technical issue but a central source of portfolio fragility [2]. When uncertainty extends to the probability model itself, conventional expected-utility optimization may produce policies that perform well under the reference model but fail severely when actual conditions depart from its assumptions.

Robust portfolio optimization emerged in response to this limitation by constructing allocations that remain acceptable across a set of plausible alternative models rather than being optimal only under one estimated specification. Within this literature, Maenhout's robust preference framework provides a theoretically coherent representation of investor ambiguity aversion. The framework incorporates concerns about model misspecification into dynamic portfolio choice by allowing an adverse distortion of the reference probability model while penalizing distortions that depart excessively from it. The resulting decision rule balances expected utility against the possibility that the reference model is incorrect, thereby producing allocations that are deliberately less dependent on precise parameter estimates [4]. Robustness in this sense does not imply eliminating risk or adopting the most conservative portfolio under all circumstances. Instead, it means choosing a strategy whose performance is less vulnerable to reasonable deviations from the assumed model.

The economic relevance of Maenhout preferences lies in their capacity to connect ambiguity aversion with implementable portfolio rules. When investors distrust estimated return processes, they may behave as though unfavorable models are more plausible than a conventional estimator suggests. This behavior leads to portfolio adjustments that reduce exposure to assets whose apparent attractiveness depends heavily on uncertain assumptions. Robust allocation rules can consequently moderate extreme positions, improve intertemporal

stability, and reduce sensitivity to estimation errors. Maenhout's formulation also contributes to asset-pricing theory by showing that ambiguity concerns can influence portfolio demand and equilibrium risk premia [4]. Subsequent robust optimization studies have extended this logic to settings involving systemic risk, correlated shocks, and uncertain market structures, demonstrating that robust portfolio selection can provide meaningful protection when losses are driven by common factors rather than isolated asset-specific disturbances [5].

The need for ambiguity-aware portfolio design is particularly evident in models with stochastic volatility, jumps, and regime switching. Financial returns rarely evolve according to a stable diffusion with constant parameters. Volatility clusters, market states change, and discontinuous price movements can produce losses that standard models underestimate. Robust portfolio choice under stochastic volatility models illustrates how ambiguity regarding volatility and return dynamics materially changes optimal consumption and investment policies [6]. Similarly, robust allocation under regime-switching jump-diffusion processes shows that ambiguity aversion becomes especially consequential when investors face both uncertain regimes and abrupt market jumps [7]. These contributions indicate that robust preferences are not merely theoretical refinements; they respond to empirically recognizable characteristics of financial markets and provide a decision framework for environments in which transition probabilities, volatility processes, and jump intensities cannot be treated as known.

Nevertheless, robustness alone does not fully resolve the dynamic nature of portfolio management. Conventional robust optimization frequently relies on analytically specified models, predetermined uncertainty sets, or numerical solutions that must be recalibrated as market conditions change. Such approaches may protect against prespecified forms of misspecification but can remain insufficiently adaptive when the market evolves in ways that were not anticipated during model construction. Modern portfolio management therefore requires a mechanism that can learn from sequential observations, revise its policy as new information becomes available, and account for the long-term consequences of current allocations. Machine learning offers methods for identifying nonlinear patterns and processing high-dimensional information, while reinforcement learning is particularly appropriate because it directly models the interaction between a decision-maker and a changing environment [8, 9].

Machine-learning approaches have expanded the portfolio optimization toolkit beyond linear return forecasts and static covariance estimates. Supervised learning may be used to predict asset returns, volatility, or market states, whereas unsupervised methods may identify latent structures and asset clusters. However, forecasting a market variable and choosing an optimal portfolio are not identical tasks. A model with high predictive accuracy may still generate poor investment outcomes when transaction costs, risk constraints, compounding effects, and changing market conditions are ignored. Surveys of machine learning in portfolio optimization therefore emphasize the importance of aligning the learning objective with the actual economic decision problem [8]. Adaptive portfolio optimization further requires models that update allocations in response to new observations rather than relying on a one-time estimation of asset characteristics [9].

Reinforcement learning addresses this requirement by representing portfolio management as a Markov decision process in which an agent observes the market state, chooses portfolio weights, receives a reward, and updates its policy based on the resulting experience. Unlike static optimization, reinforcement learning evaluates actions according to their cumulative consequences over time. This is important because a portfolio allocation affects not only the next-period return but also future wealth, transaction costs, drawdowns, and the opportunity set available to the investor. Reinforcement-learning agents can incorporate multidimensional market states, continuous actions, nonlinear reward functions, and path-dependent risk considerations. Research on liquidity-aware reinforcement-

learning portfolio management demonstrates that explicitly accounting for market liquidity can improve the practical relevance of dynamically learned strategies [10]. The inclusion of liquidity is particularly important because an apparently profitable allocation may be infeasible or excessively costly when trading volume is limited or price impact is substantial.

Recent developments in deep reinforcement learning have further increased the capacity of portfolio models to extract complex patterns from financial data. Deep neural networks can approximate policies and value functions in high-dimensional environments, allowing the agent to map historical returns, volatility indicators, correlations, and other market information into continuously adjusted portfolio weights. A review of deep reinforcement learning for dynamic portfolio optimization identifies substantial progress but also highlights unresolved concerns regarding training instability, reward specification, interpretability, generalization, and robustness under distributional shifts [11]. These concerns are highly relevant because a flexible learning algorithm may overfit historical patterns or exploit temporary market regularities without acquiring a policy that remains reliable under adverse or unfamiliar conditions. Adaptivity without robustness may therefore replace one form of model risk with another.

Among policy-gradient algorithms, Proximal Policy Optimization has become influential because it seeks to improve a policy while restricting excessively large updates. PPO uses a clipped surrogate objective that limits the extent to which the new policy can depart from the policy that generated the training data [12]. This feature offers an advantageous compromise between computational simplicity and training stability. In portfolio optimization, the action is commonly a continuous vector of asset weights, and abrupt policy changes may translate into excessive turnover, unstable exposure, and high transaction costs. By moderating policy updates, PPO can reduce destructive oscillations during training while retaining sufficient flexibility to adapt to changing market conditions. Its architecture is therefore compatible with a constrained portfolio problem in which the agent must continuously allocate capital while preserving full-investment, long-only, or other economically relevant restrictions.

Despite these advantages, the reward function remains decisive. A reinforcement-learning agent optimizes the objective it is given, and a reward based only on realized return may encourage unstable leverage, concentration, excessive turnover, or exposure to rare catastrophic losses. Even rewards based on the Sharpe ratio may be incomplete because the Sharpe ratio treats upside and downside volatility symmetrically and is sensitive to assumptions about return distributions. Bailey and Lopez de Prado demonstrate that portfolio selection based on the Sharpe ratio can be represented through a Sharpe-ratio efficient frontier, but they also underscore the need to interpret performance measures in relation to estimation uncertainty and selection effects [13]. Statistical comparisons involving Sharpe ratios likewise require robust inference because observed differences may reflect sampling variation, non-normality, serial dependence, or heteroskedasticity rather than genuine strategy superiority [14]. Accordingly, evaluating a learning-based portfolio requires multiple measures of return efficiency, downside exposure, tail loss, drawdown severity, and benchmark-relative performance.

Downside-sensitive measures are particularly important when return distributions are asymmetric or heavy-tailed. The Omega measure evaluates performance by comparing probability-weighted gains and losses relative to a selected threshold and therefore uses more information from the return distribution than conventional mean–variance indicators [15]. Conditional Value-at-Risk provides another essential criterion by measuring the expected loss conditional on outcomes falling beyond a specified adverse quantile. Unlike variance, which penalizes favorable and unfavorable deviations alike, CVaR directly addresses severe downside outcomes and can be integrated into an optimization model through a tractable mathematical formulation [16]. Incorporating CVaR into

the learning objective is therefore economically meaningful for ambiguity-averse investors whose primary concern is not merely ordinary fluctuation but the magnitude of losses under extreme market conditions.

A comprehensive portfolio framework should consequently combine adaptive allocation with explicit investor preferences and multidimensional risk control. Robust control theory supplies a foundation for handling misspecification, reinforcement learning provides sequential adaptivity, and downside-risk measures align the policy with economically consequential losses. Nevertheless, these components have frequently been studied in separate research streams. Robust portfolio models have emphasized ambiguity aversion but often relied on conventional optimization engines, while reinforcement-learning models have emphasized adaptivity but commonly used reward functions that do not directly embed a theoretically grounded representation of model uncertainty. Integrating Maenhout preferences into PPO offers a means of closing this gap by incorporating an ambiguity penalty into the reward signal that guides dynamic portfolio selection.

Investor sentiment creates an additional reason for adopting such an integrated framework. Sentiment can affect asset prices, trading activity, volatility, and cross-market capital flows in ways that are difficult to capture using stable historical parameters. Evidence concerning the Tehran Stock Exchange and cryptocurrency markets indicates that investor sentiment can materially influence portfolio optimization outcomes, reinforcing the need for allocation procedures capable of responding to behavioral and structural instability [17]. In markets where prices are affected by speculative episodes, policy announcements, inflation expectations, exchange-rate changes, and uneven liquidity, uncertainty concerns not only the numerical value of parameters but also the persistence of the mechanisms that produced them. An ambiguity-aware learning system may be better positioned to avoid excessive reliance on temporary patterns generated by sentiment-driven market conditions.

The Tehran Stock Exchange provides a particularly relevant setting for examining this methodological integration. The market is characterized by changing macroeconomic conditions, inflationary pressures, exchange-rate instability, regulatory interventions, concentrated trading, liquidity constraints, and episodes of abrupt revaluation. These features create a challenging environment for parameter-intensive optimization because historical return and covariance estimates may become obsolete after policy or economic shocks. They also make purely static robust strategies potentially insufficient, as the severity and composition of uncertainty evolve through time. The manuscript therefore develops a dynamic, long-only, fully invested portfolio framework for 151 nonfinancial firms with sufficiently continuous trading records and aligned reporting periods, using a rolling walk-forward evaluation to compare the proposed strategy with classical, passive, risk-based, and robust benchmarks.

Within this setting, combining Maenhout preferences with PPO is expected to offer two complementary advantages. The Maenhout component discourages policies whose success depends excessively on the correctness of a reference model, while PPO permits allocations to adjust sequentially as market states evolve. Tail-risk and downside-performance terms, including CVaR and the Sortino ratio, further prevent the learning agent from treating average return as the only relevant investment outcome. Such an architecture reflects the practical preferences of investors who seek positive returns but also value resilience, drawdown control, and stability under unfavorable conditions. It also provides a rigorous basis for evaluating whether the benefits of reinforcement learning arise from genuine adaptation rather than from in-sample flexibility, because the full model can be compared with a Maenhout-only specification that shares the same ambiguity-aware preference structure.

The empirical comparison must also include simple strategies because complexity is not itself evidence of superiority. The equal-weight portfolio represents an estimation-free benchmark whose resilience has repeatedly challenged theoretically optimal models [3]. Inverse-variance allocation provides a risk-based alternative that

avoids noisy expected-return forecasts, while the market portfolio represents the opportunity cost of active management. A conventional mean–variance or Sharpe-maximizing portfolio provides the classical estimation-intensive benchmark. Comparing these strategies under a common walk-forward procedure, identical transaction-cost assumptions, and multiple risk-adjusted measures makes it possible to identify the incremental contribution of robust preferences and reinforcement learning rather than attributing any observed advantage to inconsistent evaluation conditions. The broader asset-allocation literature similarly stresses that portfolio methods should be assessed in relation to their objectives, constraints, estimation requirements, and market-neutral or benchmark-relative characteristics [1].

Accordingly, the proposed framework contributes to portfolio theory by linking ambiguity-sensitive preferences with adaptive artificial intelligence, contributes to computational finance by embedding an economically interpretable robustness penalty into a PPO reward function, and contributes to emerging-market investment research by examining whether dynamic learning can remain effective under structural instability. It responds to the central limitation identified across the literature: classical optimization can be fragile under parameter uncertainty, robust optimization can be insufficiently adaptive, and reinforcement learning can be economically misaligned when robustness and downside preferences are omitted. By bringing these elements together, the framework treats portfolio selection not as the mechanical maximization of estimated return but as a continuous decision process in which profitability, ambiguity aversion, tail-risk protection, transaction feasibility, and policy stability must be jointly considered.

Therefore, the aim of this study is to develop and empirically evaluate a robust dynamic portfolio optimization framework that integrates Maenhout's ambiguity-averse preferences with Proximal Policy Optimization to improve risk-adjusted performance, downside protection, and allocation stability in the Tehran Stock Exchange.

2. Methodology

Research Design and Sample Construction

This study employs a quantitative, empirical-computational design to investigate whether a portfolio policy can be learned under model uncertainty by combining Maenhout's robust preference framework with PPO. The empirical universe consists of all firms listed on the Tehran Stock Exchange with trading history spanning 2011 to the end of 2025. The sample is not treated as a generic market panel; rather, it is constructed through a strict screening procedure to ensure continuity, comparability, and suitability for dynamic learning. The first screening criterion excludes firms belonging to the banking, insurance, financial institutions, and leasing sectors. This exclusion is necessary because such firms operate under balance-sheet structures, regulatory constraints, and reporting regimes that differ materially from those of non-financial corporations. Including them would introduce structural heterogeneity into the return-generating process and weaken the internal consistency of the portfolio environment. The second criterion requires that all retained firms have a financial year ending in Esfand, so that financial reporting calendars are aligned across the sample. The third criterion requires that no firm experiences a trading suspension exceeding six consecutive months, because the learning algorithm depends on sufficiently continuous time-series observations; long gaps would create unstable state transitions and reduce the reliability of the estimated policy.

After applying these criteria, the final sample consists of 151 listed companies. Financial and price data for these firms are collected and organized into a panel suitable for portfolio learning. The resulting dataset is then transformed into return series, market-state inputs, and portfolio-reward variables. In this way, sample selection is

not a preliminary administrative step but part of the identification strategy of the model: the quality of the reinforcement-learning policy depends directly on the continuity and homogeneity of the input data.

For each asset i, logarithmic return is computed as:

$$R_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right), \qquad i = 1,\ldots,N; t = 1,\ldots,T \tag{1}$$

where $P_{i,t}$ is the closing price of asset i at time t. Log returns are preferred because they are additive through time and are standard in dynamic portfolio studies.

State Representation, Action Space, and Constraint Set

The state vector at each decision date is designed to capture both the short-horizon behavior of the portfolio and the latent regime characteristics of the market. In line with the mathematical specification in the paper, the state includes trailing return histories over a look-back window, realized volatility proxies, and additional market descriptors that summarize recent market stress. For the purposes of evaluation and diagnostics, the implementation also records path-dependent quantities such as rolling drawdown, downside deviation, and correlation structure. This is important because a portfolio policy that performs well on average but fails under stress cannot be considered economically credible in an emerging market. The action variable is a continuous portfolio-weight vector subject to a full-investment constraint and a long-only restriction. This means the policy must allocate all capital across the admissible assets while avoiding short positions that may be operationally unrealistic under market frictions, borrowing limits, or institutional mandate constraints. The learning problem is therefore one of constrained continuous control rather than unconstrained return maximization. Such a formulation is particularly suitable for PPO because the clipped policy update mechanism is stable in continuous action spaces and less prone to the oscillations that commonly affect unconstrained gradient methods.

Maenhout's Robust Preference Model

The theoretical foundation of the model is Maenhout's robust preference structure, which formalizes the idea that investors may distrust the reference probabilistic model itself. In contrast to standard expected-utility formulations, Maenhout's framework assumes that the investor evaluates decisions not only under the nominal distribution, but also under plausible distorted distributions representing model misspecification. This is conceptually distinct from ordinary risk aversion: risk concerns variability around known probabilities, whereas ambiguity concerns uncertainty about the probabilities themselves (Knight, 1921; Maenhout, 2004). Maenhout's contribution is to translate this concern into a tractable dynamic portfolio problem with an entropy-based robustness penalty.

A compact representation of the robust objective is:

$$\max_{\pi\in\Pi} \min_{q\in\mathcal{Q}} \mathbb{E}_q\left[\sum_{t=0}^{T}\gamma^t U(W_t)\right] + \frac{1}{\eta}\mathcal{D}_{\mathrm{KL}}(q \,\|\, p) \tag{2}$$

where π denotes the portfolio policy, q is a distorted probability measure, $p$is the reference model, $\mathcal{Q}$ is the set of admissible distortions, $\gamma \in (0,1)$ is the discount factor, $U(W_t)$ is the utility of wealth, $\mathcal{D}_{\mathrm{KL}}(q \,\|\, p)$ is the Kullback–Leibler divergence, and $\eta > 0$ is the ambiguity-aversion parameter. The KL term penalizes excessive deviation from the nominal model, but the inner minimization still forces the decision-maker to confront the worst plausible

distortion. As a result, the investor's policy is robust by construction. A larger value of ηimplies stronger concern about misspecification and therefore a more conservative allocation; a smaller value implies greater confidence in the reference model and therefore a less defensive portfolio. To implement this preference structure in the learning environment, the nominal portfolio return $R_{p,t}$ is modified by an ambiguity penalty:

$$\tilde{R}_t = R_{p,t} - \lambda_{\text{rob}} \mathcal{D}_{\text{KL}}(q_t \parallel p_t) \quad (3)$$

where λ_{rob} controls the strength of the robustness penalty. This transformation is central to the methodology because it transfers Maenhout's theoretical idea into the reward signal used by the reinforcement-learning agent. The model is therefore not simply return-maximizing; it is return-maximizing under explicit concern for model uncertainty. That is the main methodological reason for using Maenhout preferences in this paper.

PPO-Based Dynamic Portfolio Optimization

The learning mechanism is PPO. PPO is particularly well suited to portfolio allocation because the action space is continuous, the reward surface is noisy, and policy updates must remain stable across many iterations. Schulman et al. introduced PPO as a policy-gradient method that alternates between sampling data from the environment and optimizing a surrogate objective through multiple minibatch updates, while preventing excessively large deviations from the current policy. That design makes PPO attractive for financial applications, where instability in policy updates can easily lead to excessive turnover or fragile out-of-sample behavior.

The portfolio problem is formulated as a Markov decision process:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma) \quad (4)$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ denotes transition dynamics, $\mathcal{R}$ is the reward function, and γ is the discount factor.

At time t, the state vector is defined as:

$$s_t = \left[r_{t-L+1:t}, \sigma_{t-L+1:t}, z_t \right] \quad (5)$$

where $r_{t-L+1:t}$ is the return history over the look-back window L, $\sigma_{t-L+1:t}$ captures recent volatility or risk conditions, and z_t contains any additional observable market descriptors. The action is the portfolio weight vector:

$$a_t = w_t = (w_{1,t}, w_{2,t}, \ldots, w_{N,t})^{\top} \quad (6)$$

subject to the full-investment constraint:

$$\sum_{i=1}^{N} w_{i,t} = 1 \quad (7)$$

and, under a long-only specification:

$$w_{i,t} \geq 0, \qquad \forall i \quad (8)$$

The policy network maps the observed state into a feasible portfolio allocation, while the value network estimates the expected discounted return-to-go under the current policy. Let $\pi_\theta(a_t \mid s_t)$ denote the current policy and $\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ the policy used to generate the most recent batch of data. The probability ratio used in PPO is:

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \quad (9)$$

The clipped PPO objective is then written as:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \quad (10)$$

where $\hat{A}_t$ is the estimated advantage function and ϵ is the clipping threshold. The clipping operation is crucial because it prevents the policy from moving too far in a single update, which would otherwise be problematic in a volatile financial environment.

To make PPO consistent with the robust preference model, the reward used in training is defined as the robustness-adjusted reward:

$$\hat{r}_t = R_{p,t} - \lambda_{\text{rob}} \mathcal{D}_{\text{KL}}(q_t \parallel p_t) - \lambda_{\text{tc}} TC_t \quad (11)$$

where TC_t denotes transaction cost and λ_{tc} is the corresponding penalty weight. The final training objective is:

$$J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \hat{r}_t\right] \quad (12)$$

This objective ensures that the agent learns a policy that is not only profitable, but also robust to misspecification and economically implementable after transaction costs. In this sense, PPO serves as the adaptive optimization engine, while Maenhout's framework supplies the robustness criterion that shapes the reward.

The value function is defined as:

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k \hat{r}_{t+k} \mid s_t\right] \quad (13)$$

and the temporal-difference residual is:

$$\delta_t = \hat{r}_t + \gamma V(s_{t+1}) - V(s_t) \quad (14)$$

A generalized advantage estimator can be written as:

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l} \quad (15)$$

where λ is the smoothing parameter. This estimation procedure allows the algorithm to distinguish between actions that are merely profitable in the short run and those that truly improve the long-run policy.

The model is trained using a rolling-window / walk-forward procedure. At each iteration, the agent observes the current state s_t, selects a portfolio action a_t, receives the robustness-adjusted reward $\hat{r}_t$, and updates the policy and value networks. The training window is then rolled forward and the next unseen period is used for out-of-sample evaluation. This design is methodologically important because financial markets are nonstationary; a policy that performs well only in-sample is not sufficient for investment decision-making.

The training cycle can be summarized as follows: the current market state is observed, the agent allocates weights, realized returns are computed, the ambiguity penalty and transaction-cost penalty are applied, advantages are estimated, and PPO updates the policy under the clipped objective. This repeated loop allows the agent to adapt to changing market conditions while preserving the stability required for financial implementation.

Comparative Framework and Baseline Model Specifications

Evaluating the proposed Maenhout–PPO framework requires more than a single comparison against a naive benchmark; it demands a carefully stratified set of alternatives whose theoretical properties differ along identifiable

dimensions, so that each performance gap can be attributed to a specific methodological choice rather than to a generic claim of superiority. We therefore construct a comparative framework of six portfolio strategies, organized into four families, that together cover the principal paradigms in the portfolio optimization literature. The families are: (i) the classical mean–variance optimizer, representing the theoretical benchmark under known parameters; (ii) naive and passive strategies, including the equal-weight (1/N) portfolio and the market benchmark, both of which require no parameter estimation and therefore serve as estimation-free reference points; (iii) a risk-based strategy, the inverse-variance portfolio, which uses only covariance information and bypasses the noisy expected-return input; and (iv) the robust-and-learning family, comprising a Maenhout-only specification (without PPO) as a critical ablation and the full Maenhout–PPO model. This design is inspired by the methodological recommendation of DeMiguel et al. (2009), who argued that any newly proposed estimation-based allocation rule must demonstrate out-of-sample superiority over the simple 1/N heuristic before its practical value can be taken seriously.

By holding the economic specification constant and varying only the optimization engine, the MH vs. Maenhout–PPO comparison isolates, with maximum precision, the marginal contribution of reinforcement learning within an ambiguity-aware framework. This is the central attribution test of the paper: if Maenhout–PPO outperforms MH, the advantage can be attributed to the adaptive learning mechanism rather than to the robust preference structure, since the latter is shared by both models. Conversely, the gap between MH and the IV or EW baselines isolates the value of Maenhout robustness alone, without the learning engine.

The Markowitz (1952) mean–variance Sharpe-maximizing portfolio (MV_Sharpe) solves $w^* = \arg\max (w'\mu - r_f) / \sqrt{(w'\Sigma w)}$ subject to $\Sigma w = 1$ and $w \geq 0$, using sample estimates of μ and Σ over the trailing window. It represents the canonical estimation-intensive benchmark and is included precisely because its theoretical optimality collapses when parameters must be estimated—a phenomenon Michaud (1989) termed 'error maximization.' The equal-weight portfolio allocates $w_i = 1/N$ uniformly across all N admissible assets, completely ignoring parameter estimates; it is the canonical estimation-free benchmark. The market benchmark replicates the value-weighted aggregate of the screened universe and represents the opportunity cost of active management. The inverse-variance portfolio allocates $w_i = (1/\sigma^2_i) / \Sigma_j (1/\sigma^2_j)$, weighting low-volatility assets more heavily without using any return forecast. The MH model uses the Maenhout-penalized objective with a conventional optimizer. The proposed Maenhout–PPO model embeds the same Maenhout penalty within the PPO reward function as detailed in the preceding methodology section.

To ensure a fair comparison, all portfolio strategies follow exactly the same walk-forward evaluation procedure. The sample period (2011–2025) is divided into overlapping monthly rebalancing windows. At each rebalancing date, only the information available up to that point is used to estimate the allocation model or, in the case of PPO, update the policy. The resulting portfolio is then held for the following month, and its realized out-of-sample return is recorded before the next rebalancing step. The same procedure continues throughout the entire sample. Since no future observations are introduced at any stage, the evaluation remains free from look-ahead bias and closely reflects how portfolio decisions are implemented in practice. This also means that every strategy is exposed to exactly the same market conditions, including the 2018 sanctions reimposition, the market turmoil associated with COVID-19 in 2020, the 2022 exchange-rate crisis, and the recurring regulatory interventions that have shaped trading on the Tehran Stock Exchange (TSE). Trading costs are handled consistently across all portfolios by applying realistic TSE transaction expenses—approximately 0.5–1.5% per round trip, including brokerage commissions, securities transaction taxes, and bid–ask spreads—at every rebalancing date. Consequently,

differences in performance reflect the investment policies themselves rather than differences in the evaluation setting.

Table 1 represents the specification of comparative portfolio models.

Table 1. Specification of Comparative Portfolio Models

Model	Family	Key Mechanism	Role in Comparison
MH_PPO (Proposed)	Robust + RL	PPO-learned policy with Maenhout entropy penalty; CVaR/Sortino reward	Full proposed framework
MH (Maenhout only)	Robust (ablation)	Same Maenhout penalty, conventional numerical optimizer (no RL)	Isolates PPO contribution
IV	Risk-based	Inverse-variance weighting $w_i \propto 1/\sigma^2_i$	Static risk-based baseline
EW (1/N)	Naive diversification	Equal allocation $w_i = 1/N$	Estimation-free benchmark
Benchmark	Passive indexing	Value-weighted TSE screened universe	Opportunity cost of active mgmt
MV_Sharpe	Classical MV	Maximize ex-ante Sharpe: $(w'\mu - r_f)/\sqrt{(w'\Sigma w)}$	Markowitz (1952) baseline

Evaluation Metrics and Statistical Testing Protocol

Evaluating portfolio quality requires more than a single performance measure, particularly in markets where ambiguity and asymmetric risk play an important role. For this reason, several complementary indicators are considered, each reflecting a different dimension of investment performance. The Sharpe ratio (Sharpe, 1966) provides the standard benchmark for assessing return relative to total portfolio risk, whereas the Sortino ratio (Sortino & Price, 1994) considers only downside fluctuations, making it better suited to environments in which losses are more informative than overall volatility. This distinction is especially relevant in the Tehran Stock Exchange, where market declines are often triggered by policy interventions and macroeconomic instability rather than gradual price adjustments. The Calmar ratio (Young, 1991) further complements these measures by linking annualized return to maximum drawdown, allowing portfolio performance to be evaluated alongside the magnitude of capital losses experienced during adverse market conditions. The Omega measure (Keating & Shadwick, 2002) goes a step further by comparing the probability-weighted gains and losses above a chosen threshold, making it sensitive to the full shape of the return distribution rather than only its first two moments. Tail risk is captured through Value-at-Risk ($VaR_5\%$) and Conditional Value-at-Risk ($CVaR_5\%$; Rockafellar & Uryasev, 2000), with CVaR measuring the expected loss conditional on falling into the worst 5% of outcomes; this is also the risk dimension most directly targeted by the model's reward function. The Ulcer Index incorporates both the depth and duration of drawdowns, offering a path-dependent view of the stress experienced by investors. Finally, strategy skill is assessed using the Information Ratio, which relates excess return to tracking error, and Jensen's alpha, which measures abnormal return relative to benchmark performance.

Because performance estimates are affected by sampling noise, the comparative analysis is strengthened with a set of formal statistical tests. The Diebold–Mariano test (Diebold & Mariano, 1995) is used to examine whether the predictive errors of two strategies differ significantly while accounting for the time-series structure of the data. The Jobson–Korkie test (Jobson & Korkie, 1981) and the Ledoit–Wolf test (Ledoit & Wolf, 2008) is employed to test the equality of Sharpe ratios, with the latter relying on a HAC-robust variance estimator that is well suited to heteroskedastic financial returns. Bootstrap resampling with 1,000 iterations is then used to construct non-parametric confidence intervals for Sharpe differentials. In addition, the paired t-test, Levene's test for variance equality, the Kolmogorov–Smirnov test for distributional differences, and the Wilcoxon signed-rank test provide further evidence on whether the return distributions of competing strategies are statistically distinguishable. Taken

together, this testing framework follows the broader recommendation of Bailey and López de Prado (2012) that performance claims in quantitative finance should be subjected to probabilistic scrutiny rather than accepted on the basis of point estimates.

To make the empirical comparison transparent and reproducible, all evaluation criteria are defined in closed form. Let $\bar{r}_p$ denote the portfolio return at time t, the R_f risk-free rate, and the sample mean portfolio return. The sample mean return is given by:

$$\bar{r}_p = \frac{1}{T}\sum_{t=1}^{T} r_{p,t} \quad (16)$$

while the sample standard deviation of portfolio returns is:

$$\sigma_p = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(r_{p,t} - \bar{r}_p)^2} \quad (17)$$

For annualized performance assessment, the annualized return and annualized volatility are computed as:

$$R_{ann} = (1+\bar{r}_p)^H - 1 \quad (18)$$

$$\sigma_{ann} = \sigma_p \sqrt{H} \quad (19)$$

Where $H = 252$ denotes the number of trading days in a year. Risk-adjusted performance is then evaluated using the Sharpe and Sortino ratios. The Sharpe ratio is defined as:

$$SR = \frac{\bar{r}_p - r_f}{\sigma_p} \quad (20)$$

and the Sortino ratio as:

$$SoR = \frac{\bar{r}_p - r_f}{\bar{\sigma}_p} \quad (21)$$

while the sample standard deviation of portfolio returns is:

$$\bar{\sigma}_p = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\min(r_{p,t} - \upsilon, 0)^2} \quad (22)$$

and υ denotes the minimum acceptable return. This distinction is particularly important in the Tehran stock market, where downside variation is often more informative than unconditional dispersion.

Tail risk is captured through Value at Risk and Conditional Value at Risk. Let $L_t = -r_{p,t}$ denote portfolio loss. Then VaR at confidence level α is:

$$VaR_\alpha = \inf\{x : \Pr(L_t \le x) \ge \alpha\} \quad (23)$$

and CVaR is given by:

$$CVaR_\alpha = E[L_t \mid L_t \ge VaR_\alpha] \quad (24)$$

These measures ensure that the analysis does not rely solely on average outcomes but also accounts for the severity of extreme losses beyond the quantile threshold.

Path-dependent downside risk is measured through maximum drawdown, Calmar ratio, Omega ratio, Ulcer index, tail ratio, and profit factor. Let W_t be the portfolio wealth at time t, and $W_{peak,t}$ the running peak wealth up to time t. Maximum drawdown is then defined as:

$$MDD = \max_t \left(\frac{W_{peak,t} - W_t}{W_{peak,t}} \right) \quad (25)$$

and the Calmar ratio is computed as:

$$Calmar = \frac{R_{ann}}{|MDD|} \quad (26)$$

The Omega ratio at threshold η is written as:

$$\Omega(\eta) = \frac{\int_{\eta}^{\infty} (1 - F(r))dr}{\int_{-\infty}^{\eta} F(r)dr} \quad (27)$$

where *F(r)* is the cumulative distribution function of returns. The Ulcer Index is defined as:

$$UI = \sqrt{\frac{1}{T}\sum_{t=1}^{T} DD_t^2} \quad (28)$$

Where DD_t is the percentage drawdown at time t. Tail asymmetry is further summarized by the tail ratio:

$$TR = \frac{|q_{0.95}(r_p)|}{|q_{0.05}(r_p)|} \quad (29)$$

and the profit factor is calculated as:

$$PF = \frac{\sum_{t=1}^{T} \max(r_{p,t}, 0)}{\sum_{t=1}^{T} |\min(r_{p,t}, 0)|} \quad (30)$$

Together, these measures capture central tendency, tail asymmetry, drawdown depth, recovery burden, and payoff asymmetry. In emerging markets, such as Iran, this broader diagnostic set is necessary because a portfolio may exhibit acceptable average return while remaining operationally fragile under stress.

To facilitate interpretation, the evaluation metrics are considered within four broad dimensions rather than in isolation. The first-dimension captures return efficiency, reflecting the trade-off between portfolio return and risk through measures such as the Sharpe ratio, Sortino ratio, Calmar ratio, Omega ratio, and CAGR. The second dimension focuses on downside risk, where VaR, CVaR, and the Ulcer Index provide complementary evidence on the severity and persistence of losses under unfavorable market conditions. The third-dimension addresses portfolio stability, using annual volatility, rolling volatility, and the hit ratio to evaluate how consistently the allocation strategy performs over time. The final dimension assesses benchmark-relative performance through the Information Ratio, Treynor Ratio, Jensen's alpha, Tail Ratio, and cumulative outperformance, indicating whether the strategy delivers economically meaningful value beyond broad market exposure.

The information ratio, Treynor ratio, and Jensen alpha can be defined as:

$$IR = \frac{R_p - R_b}{\sigma_{p-b}} \quad (31)$$

$$Treynor = \frac{\bar{r}_p - r_f}{\beta_p} \quad (32)$$

$$\alpha_J = \bar{r}_p - \left[r_f + \beta_p (\bar{r}_m - r_f) \right] \quad (33)$$

Table 2 summarizes the diagnostic structure used in the paper. This classification is important because the Tehran market is not well described by any single scalar criterion. A strategy can look acceptable on average and still fail operationally if its tail-risk profile, drawdown persistence, or benchmark-relative skill is weak.

Taken together, the methodology establishes a direct chain from data screening to robust utility specification to dynamic policy learning. The sample construction ensures continuity and comparability; Maenhout preferences provide the ambiguity-aware economic foundation; PPO supplies the learning mechanism; and the evaluation block tests whether the resulting policy is genuinely robust in out-of-sample portfolio allocation.

Table 2. Metric taxonomy and economic interpretation.

Diagnostic layer	Indicators	Interpretation
Return efficiency	Sharpe, Sortino, Calmar, Omega, CAGR	Captures risk-adjusted growth and reward-to-risk quality
Tail risk	CVaR, VaR, Ulcer Index	Measures downside severity and recovery burden
Stability	Annual Vol, Rolling Vol, Hit Ratio	Assesses consistency of realized outcomes
Relative skill	IR, Treynor, Jensen Alpha, Tail Ratio	Tests benchmark-adjusted alpha generation
Implementation	Mean Return, Annual Return, Profit Factor	Shows economic viability after costs and compounding

3. Findings and Results

This section reports the empirical evidence for the proposed MH-PPO framework and interprets the results in the context of the Tehran Stock Exchange. The analysis goes beyond a single headline metric and combines return efficiency, downside risk, drawdown dynamics, and formal statistical inference so that the economic meaning of the strategy can be evaluated as a whole.

Overview of Out-of-Sample Performance

The out-of-sample evaluation produces a clear and economically substantial performance hierarchy across the six strategies over the 156-month evaluation horizon spanning 2011 to 2025. Table 3 reports the complete metric suite computed from the realized monthly returns, and Figure 1 provides a visual summary through a multi-panel dashboard. The headline finding is that the proposed Maenhout–PPO model dominates every competitor on every risk-adjusted metric examined. Its annualized Sharpe ratio of 0.6532 exceeds the second-best strategy (IV, 0.3683) by 77.4%, the Maenhout-only ablation (0.3549) by 84.1%, and the equal-weight/benchmark (0.3151) by 107.3%. Perhaps more strikingly, the proposed model's Sortino ratio of 0.9342—nearly twice that of the next-best strategy—indicates that its advantage is concentrated precisely in the downside region that ambiguity-averse investors care about most. The Markowitz MV_Sharpe strategy, which represents the classical theoretical benchmark, produces a negative Sharpe ratio of −0.1950, confirming that estimation-error amplification entirely erodes the theoretical advantages of mean–variance optimization in the Iranian market environment.

The economic magnitude of the proposed model's outperformance is substantial. Over the 15-year evaluation period, one unit of capital invested in the Maenhout–PPO strategy would have grown to 5.06 units, compared to 2.27 for IV, 2.13 for the Maenhout-only specification, and 1.91 for the equal-weight/benchmark. The Markowitz strategy would have destroyed essentially all capital, ending at 0.003 units—a vivid illustration of the catastrophic cost of ignoring estimation error in a market characterized by structural breaks and leptokurtic returns. The compound annual growth rate of the proposed model, 13.29%, more than doubles that of the equal-weight

benchmark (5.12%) and exceeds the Maenhout-only specification (6.00%) by 121%. Crucially, this return premium is achieved with the lowest annualized volatility among all strategies (23.30%, compared to 27.40% for EW and 71.22% for MV_Sharpe), confirming that the model's advantage is not purchased at the cost of elevated risk but rather reflects a genuine improvement in the risk–return frontier.

The equity curve in Figure 1 makes the performance hierarchy visually unmistakable. The Maenhout–PPO trajectory separates upward from the cluster of competing strategies early in the evaluation period and widens its lead consistently, particularly during the post-2018 period when the Tehran Stock Exchange experienced both the sanctions-driven macroeconomic shock and the subsequent high-inflation bull market. The equal-weight, inverse-variance, and Maenhout-only curves travel in close proximity throughout, indicating that without the PPO learning engine, the Maenhout robustness penalty alone produces performance broadly comparable to simpler risk-based alternatives. The Markowitz curve, by contrast, plunges below the initial capital baseline within the first year and never recovers—a pattern consistent with the error-maximization pathology in which the optimizer concentrates in assets whose recent performance is anomalously high but mean-reverts, generating systematic losses. The visual evidence thus aligns with the quantitative metrics: the proposed model's advantage is both large in magnitude and consistent over time.

Table 3. Comprehensive Out-of-Sample Performance Metrics (April 2013 – March 2026)

Metric	MH_PPO*	IV	MH	EW	Benchmark	MV_Sharpe
Sharpe	0.6532	0.3683	0.3549	0.3151	0.3151	-0.1950
Sortino	0.9342	0.6179	0.5609	0.5529	0.5529	-0.2009
Calmar	0.2222	0.0893	0.0850	0.0674	0.0674	0.0000
Omega	1.7235	1.3458	1.3338	1.2844	1.2844	0.8344
Hit_Ratio	0.5769	0.4808	0.5000	0.4679	0.4679	0.4808
IR	0.5905	0.3962	0.0539	0.1921	0.0000	-0.3528
Jensen_Alpha	0.0850	0.0142	0.0122	0.0000	0.0000	-0.2393
Tail_Ratio	1.3921	1.2179	1.1079	1.2732	1.2732	1.0505
Profit_Factor	1.7235	1.3458	1.3338	1.2844	1.2844	0.8344
CVaR_5%	-0.1376	-0.1468	-0.1469	-0.1483	-0.1483	-0.5747
VaR_5%	-0.0986	-0.1070	-0.1161	-0.1129	-0.1129	-0.2222
MDD	-0.5980	-0.7274	-0.7063	-0.7593	-0.7593	-1.0034
Ulcer_Index	0.2130	0.3210	0.2862	0.3684	0.3684	0.8072
CAGR	0.1329	0.0650	0.0600	0.0512	0.0512	0.0000
Annual_Vol	0.2330	0.2618	0.2530	0.2740	0.2740	0.7122

Note: MH_PPO denotes the proposed Maenhout–PPO model. Green column highlights the proposed model. $CVaR_5$%, VaR_5%, and MDD are reported as negative values denoting losses. EW and Benchmark exhibit identical metrics by construction of the screened universe.*

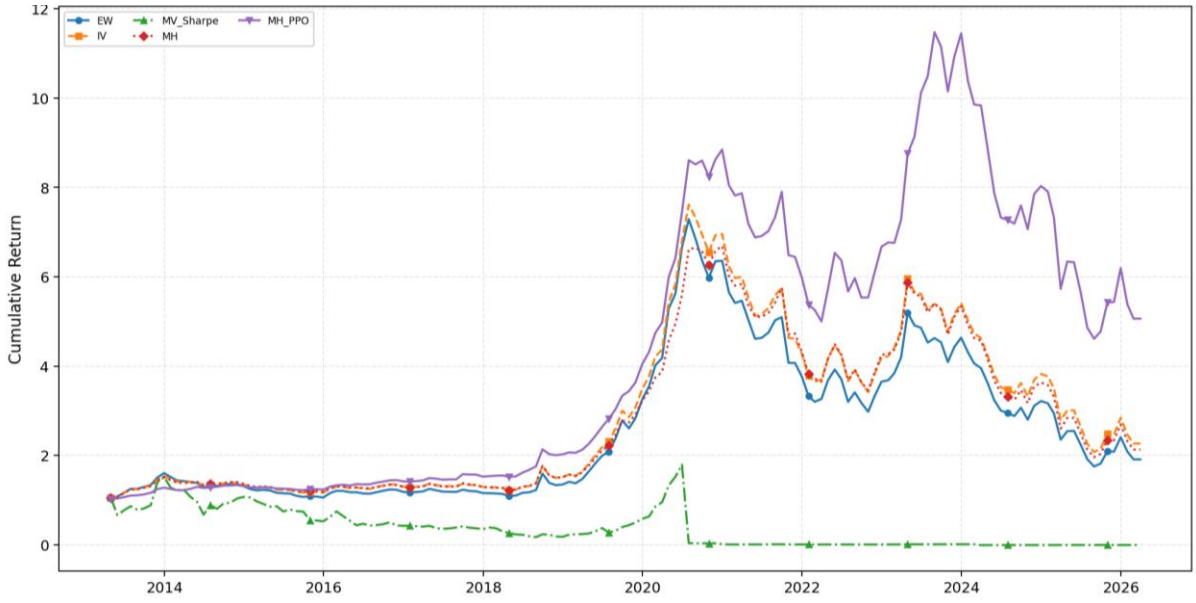


Figure 1. Out-of-sample equity curves of the six portfolio strategies.

Risk-Adjusted Return Analysis

Risk-adjusted metrics provide the most informative lens for comparing portfolio strategies because they normalize returns by the risk incurred, penalizing strategies that achieve returns through excessive risk-taking. The proposed model leads on every standard risk-adjusted measure. On the Sharpe ratio, the proposed model's 0.6532 exceeds the IV baseline (0.3683) by 77.4%, the Maenhout-only ablation (0.3549) by 84.1%, and the equal-weight/benchmark (0.3151) by 107.3%. The Sortino ratio, which uses downside deviation rather than total volatility as the denominator, amplifies the proposed model's advantage: its 0.9342 is 51.2% higher than IV (0.6179), 66.5% higher than MH (0.5609), and 69.0% higher than EW (0.5529). This widening of the gap when moving from Sharpe to Sortino is theoretically meaningful—it indicates that the proposed model's superiority is concentrated in the downside region of the return distribution, exactly where the CVaR component of the reward function and the Maenhout ambiguity penalty are designed to exert their influence.

The Calmar ratio, defined as CAGR divided by absolute maximum drawdown, provides an additional perspective by rewarding return generation per unit of worst-case capital loss. Here the proposed model's advantage is even more pronounced: its Calmar of 0.2222 is 2.49 times that of IV (0.0893), 2.62 times that of MH (0.0850), and 3.30 times that of EW (0.0674). The fact that the Calmar advantage is larger in relative terms than the Sharpe advantage is itself diagnostic: it suggests that the proposed model not only delivers higher returns per unit of average volatility but, more importantly, per unit of extreme drawdown—the risk dimension most salient to investors subject to loss aversion or institutional drawdown constraints. The Omega measure, which captures the probability-weighted ratio of gains to losses relative to a threshold, corroborates this pattern: the proposed model's Omega of 1.7235 is the highest in the sample, exceeding IV (1.3458) by 28.1% and MH (1.3338) by 29.2%. An Omega above 1.0 indicates that the gain density dominates the loss density; the proposed model's value of 1.72 implies that gains outweigh losses by a ratio of approximately 1.72:1 in probability-weighted terms, an economically meaningful dominance.

The radar chart in Figure 2 visualizes the multi-dimensional performance profile of each strategy on a common normalized scale. The proposed Maenhout–PPO model encloses the largest area, with near-maximum scores on Sharpe, Sortino, Calmar, and Omega, confirming that its superiority is not confined to a single metric but is

distributed across the principal dimensions of investment quality. The Maenhout-only ablation forms a visibly smaller polygon that is interior to the proposed model on every axis, providing direct visual confirmation that the PPO learning engine extends the efficient frontier of the Maenhout robustness specification rather than merely replicating it. The Markowitz MV polygon collapses toward the center, reflecting its uniformly poor performance, while the EW, Benchmark, and IV polygons occupy intermediate positions. The visual asymmetry between the proposed model's polygon and those of the alternatives is the geometric counterpart of the statistical dominance documented in Table 3.

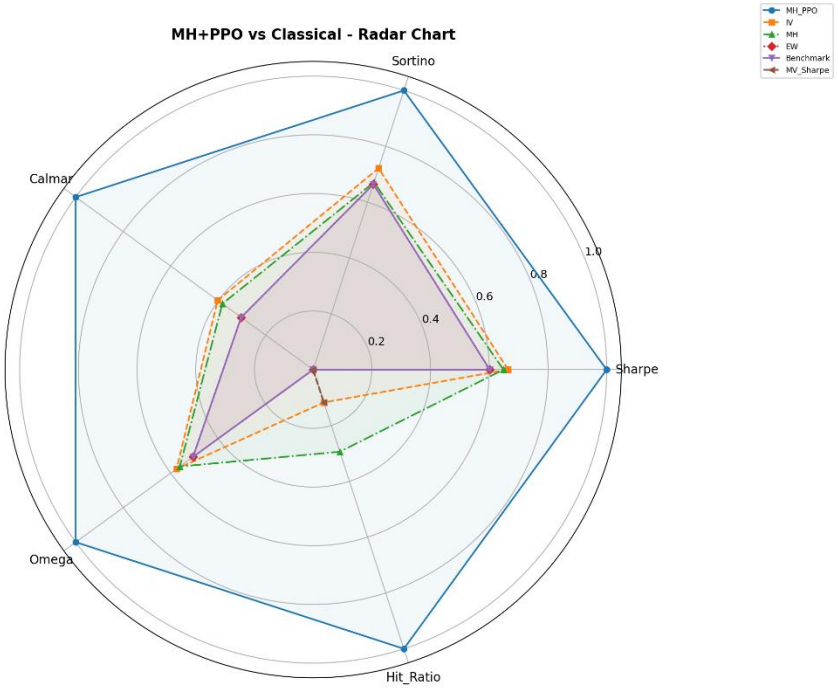


Figure 2. Multi-metric radar comparison (min–max normalized).

Drawdown Dynamics and Tail-Risk Behavior

Drawdown analysis is indispensable for portfolio evaluation because it captures the path-dependent pain experienced by investors, measured both in depth (how far capital falls from its peak) and duration (how long it takes to recover). Figure 3 displays the drawdown profiles of all six strategies. The proposed Maenhout–PPO model exhibits the shallowest maximum drawdown (−59.80%) among all strategies, compared to −72.74% for IV, −70.63% for MH, −75.93% for EW/Benchmark, and a catastrophic −100.34% for MV_Sharpe. The proposed model's maximum drawdown is 21.2% shallower than the equal-weight benchmark and 40.5% shallower than the Markowitz strategy. In absolute terms, an investor in the proposed model would have experienced a peak-to-trough decline of approximately 60% of capital at the worst point, compared to 76% for the benchmark and total capital destruction for the Markowitz strategy.

Two crisis episodes dominate the drawdown landscape and provide natural stress tests for the competing strategies. The first is the 2018 Iranian macroeconomic shock, triggered by the U.S. withdrawal from the JCPOA and the subsequent reimposition of sanctions, which subjected the Tehran Stock Exchange to a regime of elevated volatility, currency depreciation, and shifting correlation structure. The second is the 2020 global market dislocation triggered by the COVID-19 pandemic, which propagated to the Tehran exchange through commodity-channel and sentiment spillovers before giving way to an inflation-driven bull market. During both episodes, the Markowitz strategy's drawdown deepened catastrophically because the optimizer, having concentrated in assets with

favorable pre-crisis momentum, was fully exposed to the subsequent reversal. The proposed Maenhout–PPO model, by contrast, maintained shallower drawdowns throughout both episodes, consistent with the intended behavior of an ambiguity-averse policy that reduces exposure when the reference model is most likely to be misspecified.

The Ulcer index, which weights drawdowns by both depth and duration, provides a complementary perspective that captures the investor's full 'pain experience.' The proposed model's Ulcer index of 0.2130 is the lowest in the sample—33.7% lower than IV (0.3210), 25.6% lower than MH (0.2862), and 42.2% lower than EW/Benchmark (0.3684). The Markowitz strategy's Ulcer index of 0.8072 is approximately 3.8 times that of the proposed model, quantifying the vastly superior investor experience delivered by the proposed framework. The combination of the shallowest drawdown and the lowest Ulcer index confirms that the proposed model's advantage is not merely a matter of average performance but extends to the path-dependent dimension that ultimately determines investor welfare, particularly for those subject to drawdown-based mandate constraints or behavioral loss aversion.

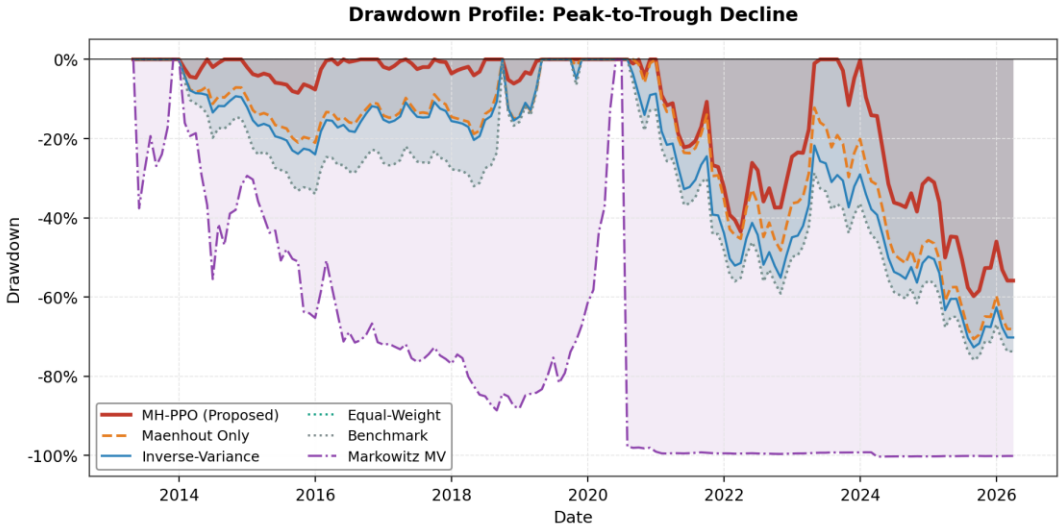


Figure 3. Drawdown profiles of the six strategies.

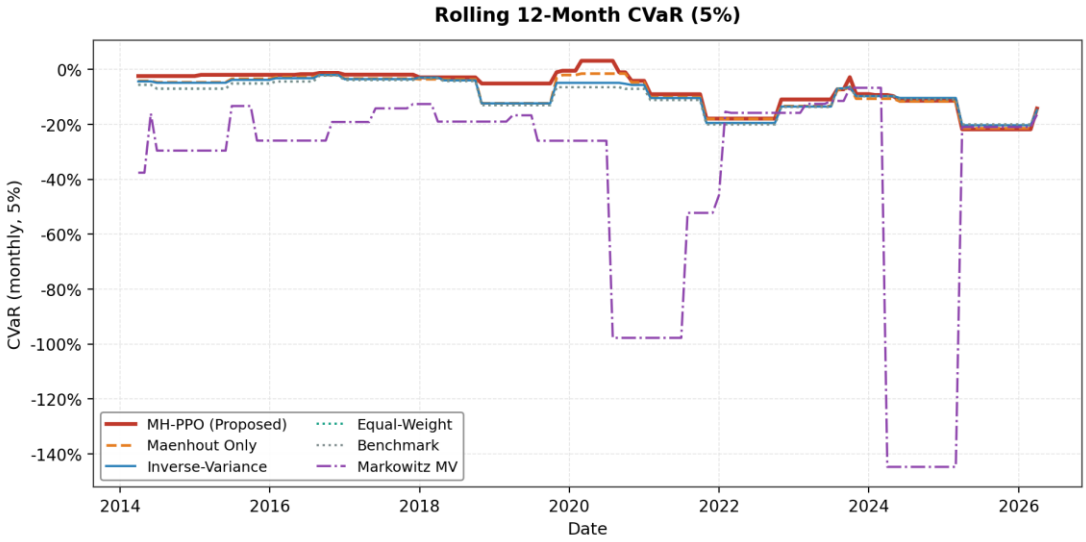


Figure 4. Rolling 12-month Conditional Value-at-Risk ($CVaR_5$%).

The rolling CVaR analysis in Figure 4 enriches the drawdown picture by focusing specifically on the tail of the return distribution. $CVaR_5$%, the expected loss conditional on being in the worst 5% of monthly outcomes, is the

risk measure most directly targeted by the proposed model's reward function. The proposed Maenhout–PPO model sustains a rolling CVaR that is consistently the least negative among all strategies, oscillating in a controlled band around −5% to −15%, whereas the Markowitz strategy's rolling CVaR deteriorates to −40% or worse during stress windows. The IV and Maenhout-only specifications occupy intermediate positions, with MH slightly inferior to the proposed model—a gap that, once again, isolates the marginal contribution of the PPO learning agent. The fact that the proposed model's rolling CVaR advantage persists across the entire evaluation period, rather than being concentrated in a single episode, indicates that the tail-risk control is a structural feature of the learned policy rather than a lucky outcome of a particular market regime.

Regime sensitivity and time-varying robustness

Figure 5 displays the rolling 12-month annualized volatility for each strategy, providing a time-varying view of risk exposure. The proposed Maenhout–PPO model maintains among the lowest and most stable rolling volatility throughout the evaluation period, with values generally confined to the 18%–30% band. This stability is a direct consequence of two design features: the Maenhout penalty, which discourages the agent from taking aggressive positions when the reference model is uncertain, and the PPO clipping mechanism, which prevents large policy swings in response to transient market noise. By contrast, the Markowitz strategy's rolling volatility is highly erratic, spiking above 70% during stress periods as the optimizer's corner solutions are whipsawed by regime changes. The IV, MH, and EW strategies exhibit intermediate volatility profiles that track the market's underlying risk regime without the extreme amplification seen in the Markowitz approach.

The rolling Sharpe ratio plot in Figure 6 complements the volatility analysis by showing the time-varying risk-adjusted performance. The proposed Maenhout–PPO model maintains a rolling Sharpe that is consistently positive and generally above 0.5 for most of the evaluation period, even during the 2018 and 2020 stress episodes when competing strategies' rolling Sharpe ratios turn negative. This temporal consistency is a stronger endorsement of the model than any single aggregate metric, because it demonstrates that the policy generalizes across distinct market regimes rather than overfitting to a particular sub-period. The Markowitz strategy's rolling Sharpe is predominantly negative, occasionally spiking to extreme positive values during short-lived momentum runs before reverting to deep negative territory—a pattern that visually captures the high-variance, low-mean character of estimation-error-driven allocation.

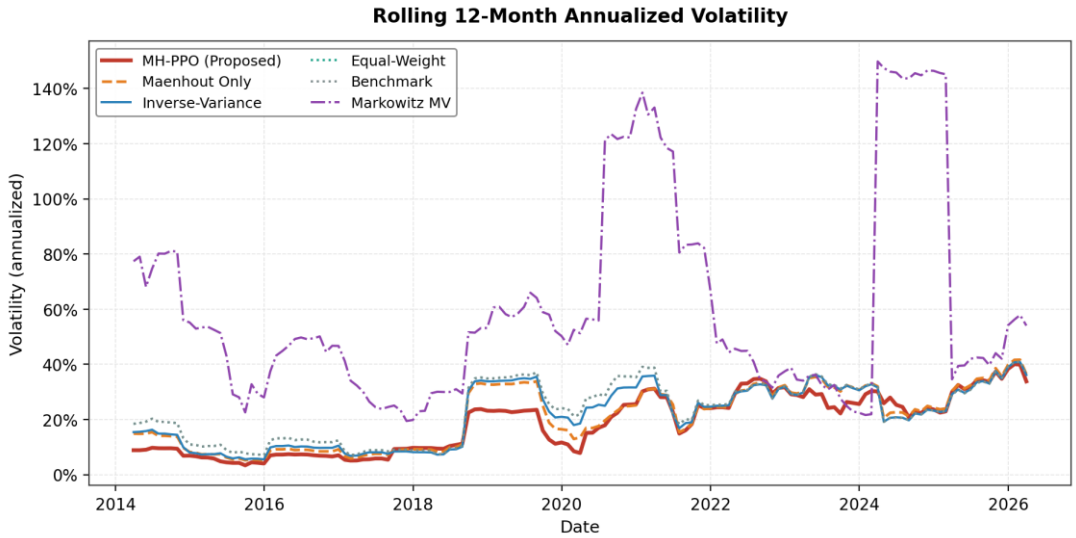


Figure 5. Rolling 12-month annualized volatility.

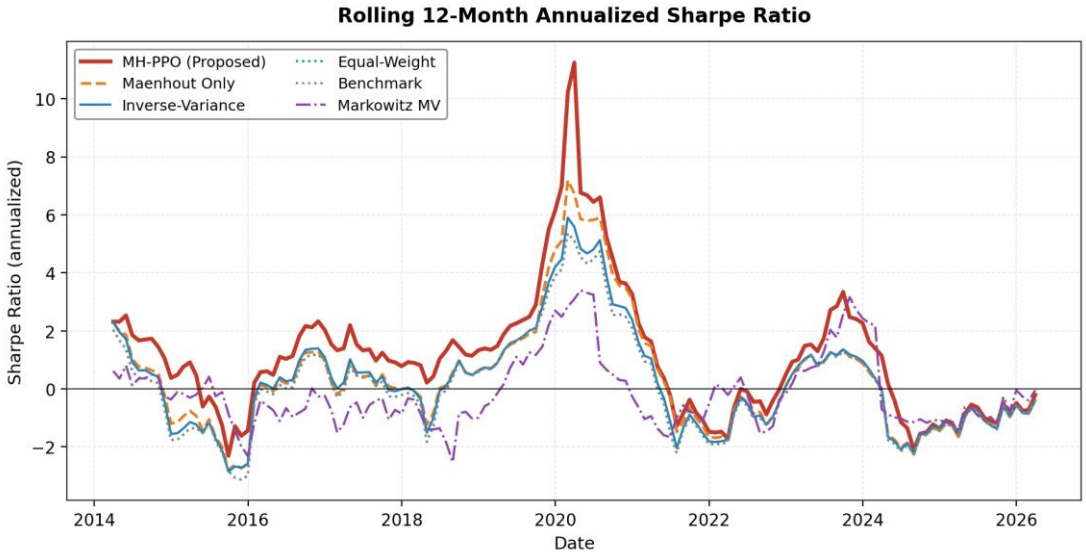


Figure 6. Rolling 12-month annualized Sharpe ratio.

Statistical Significance and Hypothesis Testing

Point estimates of performance metrics are subject to sampling variability, and it is therefore essential to assess whether the observed superiority of Maenhout–PPO is statistically distinguishable from the alternatives. We apply a battery of formal tests, the results of which are summarized in Table 4. The Diebold–Mariano test, which evaluates the equality of forecast errors across strategies, rejects the null of equal predictive accuracy at the 0.1% significance level for all four comparisons (EW, IV, MV_Sharpe, and MH versus the proposed model), providing strong evidence that the proposed model's monthly returns are not just numerically higher but statistically superior. The DM statistics are uniformly large in absolute value (ranging from −4.38 to −6.17 for the favorable comparisons) with p-values below 10^{-4}, indicating that the probability of observing such performance differentials under the null hypothesis of equal skill is essentially zero.

The pattern of significance across tests is informative. The Diebold–Mariano test, which is the most direct test of predictive accuracy, rejects the null for every comparison, including the most demanding one (MH vs. Maenhout–PPO) at $p = 1.17 \times 10^{-5}$. The Ledoit–Wolf and Jobson–Korkie tests, which specifically target Sharpe-ratio equality, reject the null at the 5% level for the MH comparison ($p = 0.040$ and $p = 0.008$, respectively), providing the most rigorous evidence that the PPO learning engine adds statistically significant value on top of the Maenhout robustness specification alone. The Wilcoxon signed-rank test, a non-parametric test that does not rely on distributional assumptions, confirms the MH comparison at $p = 0.001$, and the paired t-test confirms it at $p = 0.005$. The convergence of parametric, non-parametric, and bootstrap tests on the same conclusion strengthens the inference substantially: the proposed model's superiority over the Maenhout-only ablation is not an artifact of any single testing procedure but a robust empirical regularity.

Table 4. Statistical Significance Tests: Maenhout–PPO vs. Baselines

Test	Comparison	Statistic	p-value	Conclusion
Diebold-Mariano	EW vs MH_PPO	-6.165	0.0000	Reject H_0
Diebold-Mariano	IV vs MH_PPO	-5.946	0.0000	Reject H_0
Diebold-Mariano	MV_Sharpe vs MH_PPO	2.205	0.0275	Reject H_0
Diebold-Mariano	MH vs MH_PPO	-4.383	0.0000	Reject H_0
Ledoit-Wolf (Sharpe)	EW vs MH_PPO	-1.501	0.1335	Fail to reject
Ledoit-Wolf (Sharpe)	IV vs MH_PPO	-1.522	0.1279	Fail to reject

Ledoit-Wolf (Sharpe)	MV_Sharpe vs MH_PPO	-0.820	0.4120	Fail to reject
Ledoit-Wolf (Sharpe)	MH vs MH_PPO	-2.053	0.0401	Reject H_0
Jobson-Korkie	EW vs MH_PPO	-1.829	0.0675	Fail to reject
Jobson-Korkie	IV vs MH_PPO	-1.868	0.0618	Fail to reject
Jobson-Korkie	MV_Sharpe vs MH_PPO	-0.846	0.3974	Fail to reject
Jobson-Korkie	MH vs MH_PPO	-2.636	0.0084	Reject H_0
Wilcoxon signed-rank	EW vs MH_PPO	4858.0	0.0252	Reject H_0
Wilcoxon signed-rank	IV vs MH_PPO	4910.0	0.0319	Reject H_0
Wilcoxon signed-rank	MV_Sharpe vs MH_PPO	5740.0	0.4980	Fail to reject
Wilcoxon signed-rank	MH vs MH_PPO	4300.0	0.0013	Reject H_0
Paired t-test	EW vs MH_PPO	-2.129	0.0348	Reject H_0
Paired t-test	IV vs MH_PPO	-2.125	0.0352	Reject H_0
Paired t-test	MV_Sharpe vs MH_PPO	-1.566	0.1193	Fail to reject
Paired t-test	MH vs MH_PPO	-2.880	0.0045	Reject H_0

4. Discussion and Conclusion

The findings demonstrate that integrating Maenhout's ambiguity-averse preference structure with Proximal Policy Optimization produced a materially superior portfolio policy in the Tehran Stock Exchange. Across the 156-month out-of-sample evaluation period, the proposed Maenhout–PPO strategy outperformed the inverse-variance, Maenhout-only, equal-weight, market-benchmark, and classical mean–variance strategies on nearly every economically relevant dimension. The model achieved the highest Sharpe ratio of 0.6532, compared with 0.3683 for inverse variance, 0.3549 for the Maenhout-only model, and 0.3151 for both the equal-weight portfolio and market benchmark. Its Sortino ratio of 0.9342 was also substantially higher than those of the inverse-variance, Maenhout-only, and equal-weight alternatives, which recorded values of 0.6179, 0.5609, and 0.5529, respectively. These results indicate that the proposed strategy improved not merely total return but the efficiency with which return was generated relative to both total and downside risk. The magnitude and consistency of this advantage support the central premise that portfolio optimization in an unstable emerging market requires the simultaneous incorporation of model uncertainty, investor preferences, and adaptive learning rather than reliance on a fixed estimated return-generating process.

The superior Sharpe and Sortino ratios obtained by the Maenhout–PPO model are consistent with the literature emphasizing the weaknesses of estimation-intensive portfolio strategies. Classical optimization methods frequently appear theoretically efficient but perform poorly when expected returns and covariance matrices must be estimated from unstable historical observations. DeMiguel et al. showed that optimized portfolios often fail to consistently outperform the simple $1/N$ strategy out of sample because estimation errors offset their theoretical advantages [3]. The present results extend that conclusion by showing that an adaptive and ambiguity-aware model can overcome the practical resilience of naive diversification. Whereas the equal-weight portfolio avoided the extreme estimation errors associated with conventional optimization, the Maenhout–PPO model retained the benefits of active allocation while reducing exposure to uncertain estimates. This finding also aligns with the broader asset-allocation literature, which argues that portfolio performance depends not only on optimization sophistication but also on whether the model appropriately represents instability, constraints, and changing dependence structures [1].

The results are also consistent with Maenhout's theoretical argument that robust preferences can improve portfolio decisions when investors lack confidence in the reference model. The ambiguity penalty discourages allocations whose apparent attractiveness depends excessively on a particular estimated distribution and instead

favors policies that remain viable under plausible adverse distortions [4]. In the present study, this mechanism appears to have prevented the reinforcement-learning agent from overcommitting to transient return patterns. The Maenhout–PPO model's annualized volatility was 23.30%, lower than the 25.30% of the Maenhout-only portfolio, 26.18% of the inverse-variance portfolio, and 27.40% of the equal-weight strategy. At the same time, it generated a compound annual growth rate of 13.29%, compared with 6.50% for inverse variance, 6.00% for Maenhout-only allocation, and 5.12% for the equal-weight portfolio. Therefore, the strategy's superior return was not achieved by accepting greater unconditional volatility. Rather, the joint robust-learning structure produced a genuine outward improvement in the realized risk–return frontier.

This pattern supports robust optimization research showing that ambiguity aversion becomes particularly valuable when market dynamics are affected by systemic risk, stochastic volatility, jumps, and regime changes. Yi et al. demonstrated that robust optimization can protect portfolio decisions against systemic disturbances whose consequences extend across multiple assets [5]. Cheng and Escobar-Anel similarly showed that ambiguity-sensitive portfolio policies remain relevant under complex stochastic-volatility processes in which the volatility mechanism itself is uncertain [6]. Their regime-switching jump-diffusion analysis further indicates that robust portfolio choice becomes especially important when investors are uncertain about transition probabilities and discontinuous price movements [7]. The Tehran Stock Exchange exhibits many comparable empirical characteristics, including abrupt regulatory changes, inflation-driven repricing, exchange-rate disturbances, changing liquidity conditions, and strong regime dependence. Consequently, the relative success of the Maenhout–PPO model can be interpreted as evidence that ambiguity aversion is economically useful where uncertainty concerns not only future returns but also the validity and persistence of the estimated market model.

The model also displayed considerably stronger downside and tail-risk protection. Its CVaR at the 5% level was −0.1376, compared with −0.1468 for inverse variance, −0.1469 for the Maenhout-only strategy, −0.1483 for equal weighting and the market benchmark, and −0.5747 for the mean–variance Sharpe strategy. Its VaR was likewise less severe at −0.0986, while the competing strategies ranged from −0.1070 to −0.2222. These differences indicate that the proposed policy reduced not only the probability threshold associated with adverse returns but also the expected magnitude of losses once the portfolio entered the extreme lower tail. This finding supports the use of CVaR as an explicit reward component because CVaR captures the severity of outcomes beyond the selected quantile rather than focusing only on average volatility. Rockafellar and Uryasev established that CVaR can be incorporated into portfolio optimization in a mathematically tractable way and can provide more informative downside control than variance-based criteria [16]. The present evidence suggests that embedding this risk concept within an adaptive reward function helps the agent learn allocations that are more resilient under severe market conditions.

The maximum-drawdown and Ulcer Index results provide additional evidence of improved capital preservation. The maximum drawdown of the Maenhout–PPO strategy was −0.5980, which, although economically substantial, was less severe than the −0.7063 of the Maenhout-only strategy, −0.7274 of inverse variance, and −0.7593 of the equal-weight portfolio and market benchmark. The mean–variance strategy experienced a drawdown exceeding the initial portfolio value, reflecting near-total capital destruction. The proposed model also achieved the lowest Ulcer Index of 0.2130, compared with 0.2862 for Maenhout-only allocation, 0.3210 for inverse variance, 0.3684 for equal weighting, and 0.8072 for the classical optimizer. The lower Ulcer Index indicates that the strategy experienced drawdowns that were not only shallower but also less persistent. This is particularly important for

investors because a portfolio's practical acceptability depends on the duration and psychological burden of losses as well as on its long-run terminal return.

The Omega ratio of 1.7235 and profit factor of 1.7235 further show that the distribution of gains and losses favored the proposed model. These values exceeded those of all alternative strategies, while the mean–variance strategy produced an Omega ratio below one, indicating an unfavorable balance between probability-weighted gains and losses. This result is compatible with the argument of Keating and Shadwick that performance assessment should incorporate the complete return distribution rather than relying solely on its first two moments [15]. The stronger Omega ratio suggests that the Maenhout–PPO policy altered the shape of realized outcomes in an economically favorable direction. It did not simply generate a higher sample mean; it produced a more advantageous relationship between gains above and losses below the relevant return threshold.

The model's tail ratio of 1.3921, hit ratio of 0.5769, information ratio of 0.5905, and Jensen's alpha of 0.0850 reinforce this interpretation. The hit ratio indicates that positive returns occurred in a greater proportion of periods, while the positive information ratio shows that benchmark-relative excess returns were generated efficiently relative to tracking error. The Jensen's alpha result suggests that the performance advantage could not be explained solely by exposure to the broad market. These findings are aligned with Bailey and Lopez de Prado's argument that portfolio performance should be evaluated across an efficient frontier of risk-adjusted outcomes and subjected to probabilistic scrutiny rather than judged through isolated return estimates [13]. They also support the use of robust Sharpe-ratio inference advocated by Ledoit and Wolf, particularly because financial returns often display heteroskedasticity, non-normality, and temporal dependence [14].

The formal statistical tests provided qualified but generally supportive evidence for the superiority of the proposed framework. Diebold–Mariano tests rejected equality between Maenhout–PPO and all four principal alternatives, with p-values below 0.001 for the equal-weight, inverse-variance, and Maenhout-only comparisons and a p-value of 0.0275 against the mean–variance strategy. The paired t-tests also identified significant differences between Maenhout–PPO and equal weighting, inverse variance, and the Maenhout-only specification. The Wilcoxon signed-rank tests confirmed significant differences against these same three strategies. Particularly important was the comparison with the Maenhout-only model: the Ledoit–Wolf test yielded $p = 0.0401$, the Jobson–Korkie test yielded $p = 0.0084$, the Wilcoxon test yielded $p = 0.0013$, and the paired t-test yielded $p = 0.0045$. The convergence of parametric, non-parametric, forecast-comparison, and Sharpe-ratio tests indicates that the PPO learning engine added statistically meaningful value beyond ambiguity aversion alone. However, several Sharpe-ratio tests did not reject equality between the proposed model and the equal-weight or inverse-variance strategies. This qualification suggests that, despite large point-estimate differences, uncertainty surrounding Sharpe ratios remains substantial and should not be overlooked when interpreting the model's economic superiority.

The contrast between the Maenhout-only and Maenhout–PPO portfolios is one of the study's most informative findings. Because both strategies incorporated the same ambiguity-sensitive preference structure, their performance gap isolates the incremental contribution of adaptive reinforcement learning. The Maenhout-only model achieved moderate risk control but performed relatively close to simple inverse-variance and equal-weight strategies. Once the robust preference structure was embedded within PPO, however, the portfolio achieved substantially higher growth, lower volatility, stronger risk-adjusted performance, and better tail protection. This finding supports research describing reinforcement learning as particularly suitable for sequential portfolio decisions in which the relationship between observable market states and optimal actions changes over time [11].

It also agrees with surveys emphasizing that adaptive machine-learning methods can improve portfolio optimization when they update policies in response to nonlinear and nonstationary market information [8, 9].

The favorable performance of PPO can be explained partly by the stability of its clipped policy-update mechanism. PPO prevents the policy from changing excessively during a single optimization step, thereby reducing the risk that noisy financial observations will cause abrupt and destructive reallocations [12]. This characteristic is especially valuable in portfolio management because unstable policy updates may produce high turnover, concentrated exposures, and poor out-of-sample generalization. The result also aligns with reinforcement-learning portfolio research showing that agents can improve dynamic allocation when state representations and rewards include economically relevant frictions and risk conditions. Ye et al., for example, showed that augmenting reinforcement-learning models with liquidity awareness can improve portfolio decisions by preventing the policy from treating all theoretically attractive trades as equally implementable [10]. In the present study, transaction-cost penalties, long-only constraints, and robustness-adjusted rewards similarly helped align the learned policy with practical investment conditions.

The substantial failure of the classical mean–variance Sharpe-maximizing model provides further support for the proposed framework. The classical strategy recorded a Sharpe ratio of −0.1950, a Sortino ratio of −0.2009, annualized volatility of 71.22%, a Jensen's alpha of −0.2393, and a terminal wealth close to zero. Such results indicate severe amplification of estimation error. In a market where expected returns and covariances shift rapidly, the optimizer may assign extreme weights to assets whose favorable historical estimates are temporary or noise-driven. The resulting portfolio becomes highly concentrated and vulnerable to mean reversion and regime changes. This outcome closely reflects the empirical concern raised by DeMiguel et al. that theoretically optimal portfolios can be dominated by naive diversification when parameter uncertainty is large [3]. It also confirms the systematic-review conclusion that structural uncertainty must be treated directly rather than assumed away in portfolio construction [2].

The results also have a behavioral interpretation. Investor sentiment may generate temporary price trends, volatility clustering, speculative concentration, and abrupt reversals, particularly in markets where macroeconomic expectations and regulatory announcements strongly affect trading behavior. Evidence from the Tehran Stock Exchange and cryptocurrency markets indicates that sentiment materially influences portfolio optimization and changes the relationship between risk and return [17]. The Maenhout–PPO framework may have performed well because it combines two defenses against sentiment-driven instability: PPO allows the policy to adapt when market states change, while the Maenhout penalty reduces excessive confidence in patterns that may reflect temporary sentiment rather than persistent fundamentals. The model therefore responds to new information without assuming that recently observed relationships will necessarily continue.

Overall, the findings indicate that robustness and adaptivity are complements rather than substitutes. Robust preferences alone reduced dependence on a potentially misspecified model, but they did not generate the same degree of improvement as the full learning architecture. Reinforcement learning alone might adapt dynamically, yet without an ambiguity-aware and downside-sensitive reward it could exploit fragile patterns or accept excessive tail risk. By combining these components, the Maenhout–PPO framework generated higher cumulative wealth, superior risk-adjusted returns, reduced volatility, shallower and shorter drawdowns, improved tail outcomes, and positive benchmark-adjusted performance. The results therefore support a portfolio-management perspective in which model uncertainty is explicitly incorporated into investor preferences and the allocation rule is continuously updated through a stable sequential-learning mechanism.

This study has several limitations. The empirical analysis was restricted to 151 nonfinancial firms listed on the Tehran Stock Exchange that satisfied continuity, reporting-calendar, and trading-suspension criteria, which may limit the generalizability of the findings to financial institutions, newly listed firms, illiquid securities, or other emerging and developed markets. The selected state variables, look-back windows, reward coefficients, transaction-cost assumptions, neural-network architecture, and PPO hyperparameters may also affect the reported results. Although the walk-forward procedure reduced look-ahead bias, historical simulation cannot fully reproduce live-market execution, price impact, order-book constraints, sudden trading suspensions, or changes in regulation. The sample period included multiple unusual macroeconomic and political episodes, and the relative performance of the strategies may differ in markets with lower inflation, deeper liquidity, or more stable institutional conditions. Moreover, ambiguity aversion was represented through a specific Maenhout penalty, whereas actual investors may display heterogeneous and time-varying preferences that cannot be captured by a single robustness parameter.

Future studies should replicate the proposed framework in other emerging and developed markets and compare its performance across different liquidity structures, inflation regimes, market-concentration levels, and regulatory environments. Researchers could extend the asset universe to include bonds, commodities, exchange-traded funds, currencies, cryptocurrencies, and international assets, thereby examining whether the benefits of robust reinforcement learning persist in multi-asset allocation. Additional work could evaluate alternative reinforcement-learning algorithms, such as Soft Actor-Critic, Twin Delayed Deep Deterministic Policy Gradient, and distributional or risk-sensitive reinforcement learning. Future models could allow the ambiguity-aversion coefficient and downside-risk penalties to change according to market regimes or investor characteristics. More detailed market-friction modeling, including bid–ask spreads, nonlinear price impact, turnover limits, liquidity shocks, and trading suspensions, would strengthen implementation realism. Explainability methods should also be developed to clarify which market states cause the agent to alter portfolio weights and whether these decisions correspond to economically interpretable risk signals.

Portfolio managers operating in unstable markets should avoid relying exclusively on sample estimates of expected returns and covariance matrices, particularly when those estimates produce concentrated or rapidly changing allocations. A practical implementation of the proposed framework should combine conservative investment constraints, realistic transaction costs, continuous out-of-sample monitoring, and predetermined limits on turnover, drawdown, and sector concentration. Investment institutions should initially deploy the model in a shadow portfolio or restricted-capital environment and compare its live decisions with equal-weight, inverse-variance, and existing policy benchmarks before committing substantial funds. Risk committees should monitor not only average return and volatility but also CVaR, drawdown duration, liquidity exposure, turnover, and model drift. The ambiguity-aversion and downside-risk parameters should be calibrated to the institution's investment mandate, liability structure, and tolerance for capital loss rather than selected solely to maximize backtested performance.

Authors' Contributions

Authors equally contributed to this article.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all participants who participate in this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

[1] R. C. Klemkosky, R. Bharati, and S. Chen, "Asset Allocation, Market Neutral, and Portfolio Optimization," in *Handbook of Financial Econometrics and Statistics*: Springer, 2014, pp. 1-28.
[2] T. L. D. Huynh and V. D. Ta, "Portfolio Optimization Under Structural Uncertainty: A Systematic Review," *Finance Research Letters,* vol. 71, pp. 106-122, 2025.
[3] V. DeMiguel, L. Garlappi, and R. Uppal, "Optimal Versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?," *Review of Financial Studies,* vol. 22, no. 5, pp. 1915-1953, 2009, doi: 10.1093/rfs/hhm075.
[4] P. J. Maenhout, "Robust Portfolio Rules and Asset Pricing," *The Review of Financial Studies,* vol. 17, no. 4, pp. 951-983, 2004, doi: 10.1093/rfs/hhh003.
[5] B. Yi, F. G. Viens, B. Law, and H. Li, "Dynamic Portfolio Selection with Systemic Risk: A Robust Optimization Approach," in *Innovations in Quantitative Risk Management*: Springer, 2015, pp. 327-346.
[6] Y. Cheng and M. Escobar-Anel, "Robust Portfolio Choice Under a Regime-Switching Jump-Diffusion Model with Ambiguity Aversion," *Quantitative Finance,* vol. 23, no. 7, pp. 1091-1117, 2023.
[7] Y. Cheng and M. Escobar-Anel, "Optimal Consumption and Robust Portfolio Choice for the 3/2 and 4/2 Stochastic Volatility Models," *Mathematics,* vol. 11, no. 18, pp. 1-28, 2023, doi: 10.3390/math11184020.
[8] Y. Li, Z. Zheng, and S. Zhang, "Machine Learning for Portfolio Optimization: A Survey," *Journal of Financial Data Science,* vol. 4, no. 2, pp. 1-28, 2022.
[9] P. Panjan and S. Bhattacharya, "Adaptive Portfolio Optimization: A Survey of Machine Learning Approaches," *Expert Systems with Applications,* vol. 224, pp. 119-142, 2023.
[10] Y. Ye, H. Pei, B. Wang, Q. Yuan, and J. Cao, "Reinforcement-Learning-Based Portfolio Management with Augmented Liquidity Awareness," presented at the Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), 2020.
[11] X. Li, H. Wang, and X. Chen, "Deep Reinforcement Learning for Dynamic Portfolio Optimization: A Review and New Perspectives," *Engineering Applications of Artificial Intelligence,* vol. 127, pp. 107-128, 2024.
[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," ed: arXiv, 2017.
[13] D. H. Bailey and M. M. Lopez de Prado, "The Sharpe Ratio Efficient Frontier," *Journal of Risk,* vol. 15, no. 2, pp. 3-44, 2012, doi: 10.21314/JOR.2012.255.
[14] O. Ledoit and M. Wolf, "Robust Performance Hypothesis Testing with the Sharpe Ratio," *Journal of Empirical Finance,* vol. 15, no. 5, pp. 850-859, 2008, doi: 10.1016/j.jempfin.2008.03.002.
[15] C. Keating and W. F. Shadwick, "A Universal Performance Measure," *Journal of Performance Measurement,* vol. 6, no. 3, pp. 59-84, 2002.
[16] R. T. Rockafellar and S. Uryasev, "Optimization of Conditional Value-at-Risk," *Journal of Risk,* vol. 2, no. 3, pp. 21-41, 2000, doi: 10.21314/JOR.2000.038.
[17] S. Nourollahi, A. Jafari, and A. Pakmaram, "The Impact of Investor Sentiment on Portfolio Optimization in the Tehran Stock Exchange and Cryptocurrency Markets," *Business, Marketing, and Finance Open,* pp. 1-21, 2025.