Regime-Switching ESG Portfolio Optimization: Integrating Temporal Fusion Transformers and Reinforcement Learning for Enhanced Risk-Adjusted Returns

Abdul Kadar Muhammad Masum[1], Md. Abul Kalam Azad[2]*,
Chanda Rani Debi[3], ABM Yasir Arafat[4], Ramona Birau[5], Virgil Popescu[6],
Gabriela Ana Maria Lupu (Filip)[7], Roxana Mihaela Nioata (Chireac)[8], Stefan Margaritescu[9]
Dept. of CSE, Southeast University, Dhaka, Bangladesh[1]
Dept. of Business and Tech. Management, Islamic University of Technology, Dhaka, Bangladesh[2]
Computational Intelligence Lab, Southeast University, Dhaka, Bangladesh[3]
Department of CSE, International Islamic University Chittagong, Chittagong, Bangladesh[4]
University of Craiova, "Eugeniu Carada" Doctoral School of Economic Sciences, Craiova, Romania[5,7,8,9]
University of Craiova, Faculty of Economics and Business Administration, Craiova, Romania[6]

***Abstract*—Integrating ESG criteria into institutional portfolios creates tension with financial objectives under changing market conditions. This paper proposes an architecture combining Temporal Fusion Transformer (TFT)–XGBoost forecasts, Gaussian Hidden Markov Model (HMM) inference, Proximal Policy Optimization (PPO), and constrained portfolio refinement. An executed annual chronological experiment separately evaluates a tractable allocation implementation using Ridge forecasts and an observable prior-return risk indicator. The annual experiment uses 4,837 source firm records, a lagged-data candidate rule, and eight independently solved allocation strategies. Over the retrospective 2023–2024 evaluation, annual adaptive ESG allocation earns 19.00% cumulative net return, compared with 54.29% for constrained equal weight and 62.91% for constrained capitalization weighting. Its mean lagged ESG score is 74.58. On 100 firm-year forecasts across two market dates, Ridge RMSE is 51.25 percentage points, compared with 42.21 for historical means. All 240 allocations across six scenarios satisfy the specified annual constraints within 10^{-7}. These findings establish conditional pipeline feasibility rather than a fully validated TFT–HMM–PPO architectural outperformance or a statistically robust algorithmic edge. ESG Beta remains an exploratory co-movement metric within the framework. Unverified data vintages and low-frequency annual constraints highlight practical deployment bottlenecks for automated financial systems.**

***Keywords*—ESG investing; reinforcement learning; temporal fusion transformer; Hidden Markov Model; portfolio optimization; regime detection, multi-objective optimization; ESG Beta**

I. Introduction

ESG integration into investment decision-making represents one of the most significant structural shifts in contemporary financial markets. By 2020, global sustainable investment assets reached $35.3 trillion, representing a 55% cumulative growth over four years [1]. This expansion is not merely ethically driven; empirical evidence increasingly demonstrates that ESG attributes capture financially material risks and opportunities inadequately reflected in traditional valuation metrics [2], [3]. The COVID-19 pandemic further validated this perspective, with ESG-focused portfolios demonstrating superior downside resilience and outperforming traditional benchmarks by 2.5–3.8% during peak market volatility [4]. These findings indicate that ESG factors act as dynamic, state-specific determinants of portfolio returns rather than static screening filters.

Despite rapid methodological progress, existing ESG portfolio optimization models face substantial structural constraints. The majority rely on static mean-variance paradigms or their extensions [5], [6], which implicitly assume stationary return distributions and time-invariant relationships between ESG characteristics and financial performance. However, financial markets exhibit non-stationary behavior characterized by abrupt regime switches between bull and bear states with distinct volatility, correlation structures, and investor risk preferences [7], [8]. Empirical evidence demonstrates that the return implications of ESG performance vary systematically across regimes, capturing upside potential during expansionary phases while mitigating downside risk during contractionary periods [2], [9]. Consequently, static ESG portfolio approaches fail to exploit regime-contingent market dynamics.

Although recent research has leveraged machine learning to model non-linearity and high-dimensionality in financial data [10], [11], existing applications remain methodologically fragmented. Machine learning models are typically deployed in isolation, either as standalone return forecasters or unconstrained allocation engines, failing to combine adaptive learning with rigorous mathematical optimization. Furthermore, hard ESG screening thresholds often induce empty feasible decision spaces under market distress, while assuming firm-level ESG scores remain time-independent relative to broader market conditions [12].

*Corresponding author.

Parallel advancements in Deep Reinforcement Learning (DRL) and temporal deep learning offer promising avenues to address these challenges. Actor-critic architectures such as Proximal Policy Optimization (PPO) [13] exhibit stability in sequential decision-making tasks like dynamic asset allocation [14], [15]. Concurrently, temporal architectures including Long Short-Term Memory (LSTM) networks and Temporal Fusion Transformers (TFT) achieve state-of-the-art accuracy in financial forecasting by capturing long-range interactions and multi-horizon statistics [16], [17]. However, these techniques have not been systematically integrated into an ESG-aware portfolio optimization framework that explicitly models market regimes, tail risk, and continuous penalty structures.

While Conditional Value-at-Risk (CVaR) multi-objective optimization models effectively manage downside tail risk [18], [19], [20], current ESG implementations typically apply fixed scalarization weights [12], [21]. This implicitly assumes constant investor risk preferences across time—contradicting empirical behavior where investors prioritize return maximization during expansions and capital preservation during downturns.

To address these gaps, this paper introduces a unified, regime-adaptive hybrid reinforcement learning framework for ESG portfolio optimization. The proposed two-stage pipeline extracts features and market regimes from a broad asset universe before optimizing a target portfolio of 50 assets. The supplied export contains 4,837 distinct firm identifiers with annual fields for 2017–2024; it is not a verified historical S&P constituent panel. The proposed architecture combines HMM regime detection, TFT–XGBoost return forecasting, and PPO allocation proposals followed by constrained optimization. ESG Beta measures co-movement of changes in firm and aggregate ESG ratings. Soft ESG preferences may reduce screening pressure, but do not guarantee feasible portfolios, liquidity, or capital preservation. The proposed architecture, historical notebook diagnostics, and executed annual allocation experiment are distinguished throughout the empirical analysis.

The main contributions of this study are as follows:

- Theoretical Integration: A unified framework combines regime-switching dynamics, deep reinforcement learning, and ESG portfolio optimization; ESG Beta is an exploratory rating co-movement measure.
- Methodological Innovation: The proposed forecast–optimization pipeline adapts objective weights to filtered regime probabilities and checks allocation feasibility after policy proposal and solver refinement.
- Empirical Evaluation: A separately specified annual chronological experiment compares eight constrained allocation rules. Adaptive ESG earns 19.00% cumulative net return in 2023–2024, below constrained equal weight at 54.29%; no statistically significant advantage or full-architecture validation is claimed.

II. Literature Review

This section critically reviews the theoretical and empirical foundations underlying the regime-adaptive ESG portfolio optimization framework. The literature is organized into six thematic streams: 1) ESG integration in portfolio construction, 2) multi-objective portfolio optimization with tail risk measures, 3) machine learning and deep learning for financial forecasting, 4) reinforcement learning in portfolio management, 5) regime-switching models in finance, and 6) hybrid methodologies combining learning and optimization. For each stream, the discussion identifies methodological limitations that motivate the proposed framework design.

A. ESG Integration in Portfolio Construction

The incorporation of ESG factors into investment decisions has evolved from ethical screening to sophisticated integration frameworks. Initial studies were devoted to the paradox of ESGperformance: does sustainable investment require financial sacrifice? In a metaanalysis study of more than 2,000 empirical studies, Friede et al. [22] discovered that 62.6% identified positive ESG-return relationships, 10% identified negative relationships and 27.4% identified neutral relationships. This heterogeneity indicates a context specificity, but not a universal effect as common with a non-dynamical optimization model.

1) Static ESG portfolio optimization: Branch et al. [6] offered the seminal advice regarding the construction of ESG portfolio, including a multi-step framework: 1) the normalisation of the ESG data by the providers, 2) analysis of the factor exposure, and 3) the mean-variance optimisation with ESG constraints. Nevertheless, their final procedure presupposes the parity of the regimes of the ESG-return curve and uses hard ESG thresholds (e.g. minimum score = 70), which are not optimal to use in case of a transition between regimes when ESG premiums vary. This work was expanded by Alessandrini and Jondeau [5], who assessed four strategies of ESG integration namely exclusion screening, best-in-class selection, ESG tilting, and ESG momentum. Based on European equity data (2007-2019), they showed that ESG momentum strategies (with an overweight to firms with rising ESG scores) performed the best, out of static screens, by up to 1.2- 1.8% per year. The proposed model is based on this observation by establishing ESG momentum characteristics and ESG Beta to reflect systematic ESG trends. Recently, Shushi [23] introduced a multi-objective optimization model that separates the Environment, Social, and Governance preferences so that the investors can define the utility functions that each pillar would be specified with. Even though this method is conceptually beautiful, it does not imply the temporal dynamics: the relationship between the ESG pillars differs given different market regimes (e.g., the environmental issues gain significance during bull markets when the regulatory pressure is stricter).

2) ESG Performance during market stress: The natural experimental venue concerning ESG resilience was the COVID-19 pandemic. A study by Diaz et al. [2] of the 1,200 equities in the world over the Q1 2020 period revealed that high-ESG portfolios have 15-20% lower volatility and recover faster than the low-ESG portfolios. Zhang et al. [4] confirmed them in Chinese markets citing high performance of stakeholder trust and resilience in the operations. Ouchen [9] proved that ESG portfolios have lower turbulence (quantified by realized volatility and drawdown period).

These results encourage the regime-conditioned model: when ESG variables offer asymmetric protection (i.e., can be stronger in bear market), optimization objectives should

modify ESG weights dynamically depending on the probability of different regimes. The presence of the existing literature agrees to this phenomenon without providing formal integration frameworks.

3) Limitations of current ESG frameworks: Although significant advances have been made, existing ESG portfolio models suffer three fundamental drawbacks namely: 1) Static optimization horizons, which do not take into account ESG changes with time, 2) Univariate ESG measures, which do not address pillar-specific effects and ESG motion, and 3) No systematic ESG risk factors, such as those in traditional factor models. The ESG Beta measure controls three limitations by assigning a quantitative measure of sensitivity of firms to aggregate ESG trends in the market.

B. Multi-Objective Portfolio Optimization and Tail Risk Measures

Modern portfolio theory found in different formulations and meanvariance optimization has been enhanced to include the multiplicity of, mostly competing, objectives. The multi-objective optimization (MOO) models produce pareto-efficient frontiers in which no objective can be enhanced at the expense of another [12].

1) Multi-objective ESG portfolios: According to the study by Xidonas and Essner [12], a tri-objective ESG portfolio model was developed that would maximize the returns and risk reduction and the weighted ESG scores. They collected results that are based on the evolutionary algorithms (NSGA-II) and proved that Pareto-optimal ESG portfolios could be 85-90 percent of maximum-return portfolio performance with a 25+ point greater ESG score. Nevertheless, their approach to scalarization relies on set weights of objective ($w_{\text{return}} = 0.4$, $w_{\text{risk}} = 0.3$, $w_{\text{ESG}} = 0.3$), which is not optimal when the interest of investors change between regimes.

Chen et al. [21] have embedded ESG into the multi-period portfolio optimization, whereby they have developed a mixed-integer programming model with cardinality restrictions as well as minimum ESG thresholds. Their backtests (2010-2018 Chinese A-shares) revealed that ESG-constrained portfolios have given up 0.8-1.2% per year returns but tail risk was reduced. The precise percentage previously quoted is omitted because its source passage has not been verified. This trade-off motivates regime-dependent ESG preferences; a reduction in that trade-off remains an empirical question.

Recently, Singh et al. [24] suggested an AI-based multiobjective Indian equity model, the combination of genetic algorithm, and fuzzy logic to model ESG preferences. Their method is innovative and does not have 1) regime detection, 2) temporal forecasting or 3) theoretical convergence guarantees. The gaps in the proposed PPO based RL model.

Ghanbari et al. [25] combined prospective ESG indicators (e.g. carbon reduction goal) and past scores as a plausible CVaR procedure. They achieved 18 percent of CVaR in one of their case studies, the DJIA. The framework builds on this idea by adding ESG momentum and regime conditional CVaR objectives that adjust to market volatile conditions.

2) Conditional value-at-risk in portfolio optimization: The definition of CVaR (Expected Shortfall) was put forward by Rockafellar and Uryasev [18] as a more intelligent risk measurement in place of VaR and it was demonstrated that the minimisation of CVaR can be reformulated in a convex optimisation form. The practical application of this theory in portfolio management was made possible by this theoretical basis. Alexander et al. [19] further augmented CVaR maximization with maximum drawdown, rendering a mix integer program, which selects regimes in a portfolio. Their empirical findings (S&P 500, 1990-2015) indicated that CVaRdrawdown portfolios cut the losses in the crisis periods by 40 percent with little losses of 0.5 percent in peak years. Their regime identification is however based on ex post labeling as opposed to predictive HMMs. Zhao et al. [20] proposed an asymmetric Laplace-based mean-CVaR-skewness model, which supports non-normal distributions of the returns, frequently seen in the emerging markets. The framework uses regime-specialized CVaR objectives (95th percentile in bull regimes, 97.5 th percentile in bear regimes) to trade off between risk and returns dynamically.

3) Gaps in multi-objective ESG optimization: Existing MOO-ESG literature has the shortcomings: 1) Static scalarization weights, failure to capture inclination of investor to regime-specific score, 2) Static time-based modeling of ESG score change, and 3) No learning to adjust goal prioritization in the light of the historical regime-changes. These limitations are taken care of by the weighting scheme of regime-conditional.

C. Machine Learning and Deep Learning for Financial Forecasting

Machine learning has transformed the field of financial prediction because it is able to obtain non-linear relationships that cannot be obtained by the traditional econometric model. This review considers three related sub-domains general ML to use ESG portfolios, ensemble approaches, and temporal deep learning architectures.

1) Machine learning for ESG portfolio selection: Vo et al. [10] were the first to use deep learning to predict ESG-adjusted returns using a three-layer feedforward network, socially responsible investing. The model they developed used 45 financial characteristics and 12 ESG variables and obtained 62 percent predictive power of direction on Vietnamese stocks. Nevertheless, their architecture does not include temporal dependencies that are very important in sequential decision-making and their integration of ESG is done based on input features and not dynamic objectives.

Wu et al. [11] suggested the method of sparse ESG portfolio optimization by using the LASSO regularization and the ML-based ML returns prediction. Their portfolio picks 0.500 stocks 30- 50 stocks by using 500- stocks universes, which are better at improving Sharpe ratios by 0.3- 0.5 points. Though computationally efficient, their methodology omits 1) regime dynamics, 2) ESG momentum and 3) multi-period planning horizons of the nature of RL formulations.

Laher et al. [26] have compared six ML algorithms (random forests, XGBoost, neural networks) of socially responsible portfolio optimization on JSE-listed equities. XGBoost performed better (Sharpe = 1.18 vs. 0.92 mean-variance)

encouraging us to use XGBoost as meta-learner to correct residuals in Stage 4.

Recently, Feng et al. [1] combined sentiment analysis of news text, ESG, and SDG (Sustainable Development Goal) performance, and extraction features are provided using transformer-based language models (BERT). Their model of ensemble (of sentiment, ESG and technical) reduced RMSE by 15 percent over the baseline predictions. The framework takes a complementary strategy, where the model applies TFT attention mechanisms to determine timeframes of relevance in historical texts as opposed to processing unstructured text.

2) Ensemble methods and gradient boosting: XGBoost, a Gradient boosting framework, introduced by Chen and Guestrin [27], is now the standard of tabular data prediction in the industry. The overfitting is prevented by its regularization mechanisms (L1/L2 penalties, tree pruning) which are capable of capturing more intricate interactions. However, as displayed by Sun et al. [28] surpassing single models by 12-18% in accuracy to predict credit risks, ensemble XGBoosts have shown to be more effective when the base learners are heterogeneous (neural networks, tree methods, linear models).

In recent research [29], [30], stacked ensembles are used where initial layer models (LSTM, GRU, CNN) are used to produce prediction, which are then optimized by meta-learners (XGBoost, LightGBM). These constructions produce state of the art output in the forecasting of stock price (RMSE is reduced by 20-25 percent compared to single models). This paradigm is used in the TFT-XGBoost pipeline, where TFT has the role of temporal encoding and XGBoost of correcting residual non-linearities.

3) Temporal deep learning architectures: Zhang et al. [16] used convolutional neural networks (CNNs) to restrict the predictions of order books that indicated that spatial hierarchies of financial data can be used to make predictions. Nevertheless, CNNs do not incorporate explicit time modeling, and thus recurrent architectures are required to augment in time dependence states.LSTMs are best suited at a financial time series with multi-scale patterns. However, recent hybrid models [31], [32] encodes with LSTM and attention with transformer learning to create interpretable multi-horizon predictions.

Lim et al. [17] proposed the Temporal Fusion Transformer (TFT) which is specialized to multi-horizon prediction having heterogeneous input (static metadata, known future inputs, seen historical series). TFT uses: 1) variable selection networks that find the relevant features, 2) temporal processing by LSTM layers, 3) multi-head selfattention to make long-range predictions, and 4) quantile regression to make probabilistic predictions. Experimental tests in a variety of fields (electricity demand, traffic, finance) indicate a 10-40% higher accuracy than LSTM or DeepAR baseline.

Recent uses to finance [33], [32] demonstrate that the TFT attention processes offer intelligible clues on regime-sensitive importance of features. In the case of cryptocurrency trading [33], prediction preference is attributable to on-chain metrics predicting in bull regimes and technical indicators predicting in bear phases as weights of attention. This is in agreement to the hypothesis since regime-conditional feature weighting makes the forecast more accurate.

4) Limitations of current ML-ESG approaches: Although the superior predictive performance of ML methods has been demonstrated, their use with ESG portfolios is constrained by: 1) a Lack of temporal attention mechanisms that distinguish between regime-dependent feature significance, 2) absence of ensemble structures that combine both temporal encoding and gradient boosting, and 3) the absence of constraint based optimization. In the proposed framework, these gaps are filled in by TFTXGBoost forecasting (Stage 4) with constrained multi objective optimization (Stage 6).

D. Reinforcement Learning in Portfolio Management

Reinforcement learning (RL) has treated portfolio management as a Markov decision process (MDP), in which agents learn the optimal trading policy through trial and error interaction experiences with market environments. In contrast to supervised learning, RL steps maximize long-term cumulative rewards and does not look at one step forward predictions.

1) Deep reinforcement learning for portfolio optimization: A framework of DRL in a portfolio allocation was given by Benhamou et al. [14] with comparison between algorithms based on value (DQN), policy, and actors (DDPG and SAC). The empirical study of S&P 500 constituents (2010-2018) conducted by them proved that PPO is the most efficient risk-adjusted performance (Sharpe = 1.42) and sample efficiency, which requires only 40% of training episodes compared to A3C. The framework uses these results to select the algorithm. Ye et al. [15] enriched the RL state spaces with the prediction states of asset movement by using the price trends prediction models. Such a combination obtained 0.25-0.35 point better Sharpe ratios than RL agents that used only historical price-volume data. The proposed model goes ahead to generalize and use TFT-XGBoost predictions and regime indicators, which are integrated into the state representation of PPO.

In recent research, it is shown that the RL agents are capable of learning dynamic rebalancing strategies that outperform the static tangency portfolios by 15-25 percent as measured in Sharpe adjusted terms [34], [35]. Kang and Tian [36] implemented RL and Bayesian hierarchical risk parity, and they obtain sound performance in consecutive market cycles. Nonetheless, these strategies do not concern ESG aspects at all. De-la-Rica-Escudero et al. [37] critiqued the black box criticism of the DRL by using SHAP (SHapley Additive exPlanations) on the trained agents and found that the successful policies give priority to the momentum signals during trending markets and to the mean-reversion signals during the range-bound periods. This interpretability is in line with the objective which is to know policy behavior with attention weight in regard to regime conditions.

2) Policy gradient and actor-critic methods: Schulman et al. [13] came up with Proximal Policy Optimization (PPO) which restricts updates in the policy to a trust region that does not allow harmful large gradient updates. PPO's clipped surrogate objective:

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

where, $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ and $\hat{A}_t$ is the estimate of the advantage, is a monotonic improvement, and appropriate sample efficiency. PPO has been set to be the default algorithm in continuous control programs, such as portfolio management [38].

The asynchronous Advantage ActorCritic (A3C) proposed by Mnih et al. [39] trains the model on a set of instances of environments in parallel. Computationally efficient, A3C however has greater variance than PPO, as its policy updates are not constrained. Zhang et al. [40] presented GradientDICE as an off-policy evaluation tool which enhances sample reuse in financial RL online training is costly.

3) Limitations of current RL portfolio methods: The current RL portfolio frameworks are characterized by the following weaknesses: 1) Ignoring ESG objectives, whereby portfolio management primarily aims at maximizing returns. 2) Failure to consider regime-sensitive state representations, i.e. ignoring state transitions in the market, and 3) Lack of integration with constrained optimization, which results in infeasible policies (e.g. position limits). The proposed framework manages these gaps by: ESG-enhanced state spaces (dimension depending on universe size, with ESG Beta and momentum) and HMM-based regime indicators and hybrid RL optimization architecture.

E. Regime-Switching Models in Finance

Financial markets exhibit time-varying dynamics that violate stationarity assumptions of classical models. Regime-switching frameworks capture discrete structural breaks and smooth transitions between distinct market states.

1) Hidden markov models for regime detection: HMMs assume that observed market returns are generated by a latent regime variable following a Markov chain:

$$P(S_t = j|S_{t-1} = i) = p_{ij}, \quad \sum_{j=1}^{K} p_{ij} = 1 \tag{2}$$

where $S_t \in \{1, \ldots, K\}$ is the unobserved regime. Returns within each regime follow distinct distributions (e.g., Gaussian with regime-specific μ_j, σ_j).

Nystrup et al. [41] applied HMMs to portfolio optimization, demonstrating that regime-aware strategies reduce drawdowns by 35–45% compared to regime-blind benchmarks. Their framework estimates regimes via the Baum-Welch algorithm and conditions portfolio weights on filtered regime probabilities $P(S_t = j|\mathcal{F}_t)$. However, their optimization remains mean-variance, lacking multi-objective ESG considerations.

Wang et al. [7] implemented regime-switching factor investing using HMMs to identify four market states: bull-growth, bull-defensive, bear-growth, bear-defensive. Factor portfolios (momentum, value, quality) are regime dependent ($\alpha = 8.2\%$) in bull-growth (a = 8.2) and bear-defensible ($\alpha = -3.1\%$). These results are encouraging the regime-conditional ESG weighting.

2) Regime-switching portfolio optimization: The Markov regime-switching risk parity optimization was developed by Costa and Kwon [8] who determined that regime-sensitive risk budgets outperform out-of-sample Sharpe ratios by 0.15- 0.25 points. Their integrated setting optimizes regime-conditional allocations, which makes them computationally infeasible to apply to large universes ($N < 50$). To enhance the preciseness of regime identification, Zheng et al. [42] suggested spectral clustering of HMM estimation, which enhanced the outcome in terms of accuracy by 20% in comparison with the EM-based estimation techniques. A regime features extraction technique is taken by them based on the covariance matrices of returns to yield more improved separation of bull/bear states. Spectral clustering is implemented in the robustness validation. According to Mullapudi [43], there have recently been two apparent trends in regime-switching models, as reviewed by Mullapudi: 1) machine-learning improvements (neural network-based regime classifiers), and 2) combined with other data (sentiment, ESG). The proposed model leads to both trends because it incorporates both HMM regime detection and ESG Beta.

3) Gaps in regime-switching portfolio literature: The existing regime-switching models have: 1) No integration with ESG dynamics, 2) No reinforcement learning elements of adaptive policy learning and 3) No time prediction of regime switching. The framework addresses these limitations in the HMM-RL-TFT integration, to bring forward-looking regime-wise ESG portfolio management.

F. Hybrid Methodologies: Learning and Optimization

Machine learning and mathematical programming are applied together in computational finance as a frontier. Hybrid methods take advantage of complementary strengths: ML is better at pattern recognition and non-linear modeling whereas optimization provides constraint satisfaction and optimality guarantee.

1) Warm-starting optimization with learning: Recent work in operations research explains that warm-starting solvers can be generated using ML to speed up optimization solvers. In vehicle routing, neural networks can be used to forecast nearly optimal tour structure, solver time is 50-70 times shorter. Learning-based warm start is not fully explored in portfolio management. The proposed model is valuable in that it employs PPO-created weights of the portfolio as an initial point to constrained multi-objective optimization. This hybrid system attains: 1) Feasibility with optimization-based constraint projection, 2) Adaptability with RL-based policy learning as well as 3) Efficiency with warm-start acceleration.

2) Bayesian optimization for hyperparameter tuning: Wu et al. [44] reviewed the literature on Bayesian optimization (BO) applied to tune hyperparameters in ML and found that it was 30-50 times faster than grid search in time to tune hyperparameters. BO is a type of objective modelling (e.g., validation RMSE) by means of a Gaussian Process and maximises hyperparameters expected to produce an improvement.

Falkner et al. [45] presented BOHB, which is a hybrid of BO and early stopping in Hyperband to make the large-scale tuning effective. The paper of Snoek et al. [46] validated the usefulness of BO to neural network tuning and obtained

state-of-the-art performance in CIFAR10 with 10 times less evaluations than random search.

Shahriari et al. [47] reviewed BO acquisition functions (expected improvement, upper confidence bound, probability of improvement), providing theoretical convergence guarantees under Lipschitz continuity assumptions. The framework employs BO (via Optuna [48]) for joint optimization of TFT architecture and PPO hyperparameters (Section III, Stage 4).

3) Gaps in hybrid methodologies: Existing hybrid learning-optimization frameworks lack: 1) *Application to ESG portfolio management*, 2) *Integration of regime-switching dynamics*, and 3) *Multi-stage pipelines* combining forecasting, RL, and constrained optimization. The eight-stage framework (Section III) provides the first comprehensive integration of these components.

G. Synthesis and Positioning of the Contributions

Table I synthesizes how the proposed framework addresses the core limitations identified across the six literature streams. Key differentiators include: 1) the first unified integration of Hidden Markov Model (HMM) state detection with ESG portfolio optimization, enabling dynamic risk adaptation across market states; 2) a novel *ESG Beta* construct quantifying firm-level ESG rating-change sensitivity to aggregate ESG momentum; 3) a soft-penalized two-stage architecture (500+ asset regime extraction, 50-asset optimization) using soft ESG preferences with explicit downstream feasibility checks; 4) a temporal ensemble combining TFT and XGBoost for state-conditioned return forecasts; and 5) a regime-conditional scalarization mechanism dynamically trading off return maximization versus CVaR minimization across expansionary and contractionary states.

By unifying regime-switching econometrics, deep reinforcement learning, temporal neural architectures, and multi-objective optimization, the proposed framework advances sustainable portfolio management from static screening to dynamic, risk-managed allocation. The next section details the methodology.

III. Methodology

Implementation scope. Stages 1–8 specify the proposed learning architecture and its intended evaluation protocol. The supplied historical notebook includes annual aggregation, K-means clustering, a BiLSTM-based forecast branch, and a custom policy update; it does not verify an end-to-end TFT–Gaussian-HMM–PPO implementation. Section III-H specifies the separately executed annual Ridge and constrained-allocation experiment. It tests an implementable annual allocation subset with its own declared constraints, not the complete learning architecture. Monthly histories, quarterly release vintages, and historical constituent membership are not established by the export. The remaining architecture descriptions and candidate configurations are specifications unless explicitly identified as executed annual results.

This section presents the eight-stage regime-adaptive hybrid reinforcement learning framework for ESG portfolio optimization. Fig. 1 shows the entire pipeline, which includes the flow of data between stages as well as retraining feedbacks.

A. Stage 1: Data Preprocessing

Pre-processing provides Quality and Stationarity of the data, which is essential in the training of models. The framework has a four-step pipeline that is used to handle missing values, outliers, temporal mapping, and transformations of distributions.

1) Data cleaning: Missing Value Imputation. The ESG scores have irregular reporting practices (quarterly to annual reporting). The framework applies forward-fill imputation with a maximum gap of two quarters. For gaps exceeding six months, the procedure imputes using sector-median ESG scores computed over rolling 12-month windows:

$$\text{ESG}_{i,t}^{\text{imputed}} = \text{median}\{\text{ESG}_{j,t} : j \in \text{Sector}(i)\} \tag{3}$$

Financial ratios with missing values (e.g., R&D intensity for non-tech firms) are imputed using sector-median values stratified by market capitalization quintile to account for size effects.

Outlier Treatment. Financial ratios are winsorized at the 1st and 99th percentiles within sector-year groups to mitigate the influence of extreme values without discarding observations:

$$X_{i,t}^{\text{win}} = \begin{cases} P_1(X_{\cdot,t}^{\text{sector}}) & \text{if } X_{i,t} < P_1(X_{\cdot,t}^{\text{sector}}) \\ P_{99}(X_{\cdot,t}^{\text{sector}}) & \text{if } X_{i,t} > P_{99}(X_{\cdot,t}^{\text{sector}}) \\ X_{i,t} & \text{otherwise} \end{cases} \tag{4}$$

where, P_k denotes the k-th percentile.

ESG Anomaly Detection. The framework flags ESG score jumps exceeding 30 points quarter-over-quarter for manual review, as such changes typically reflect rating methodology revisions rather than fundamental improvements. Flagged observations are validated against news events and regulatory filings.

Temporal Synchronization. At every portfolio origin, only observations whose release dates precede that origin may be used. Fiscal period labels alone do not establish availability. Point-in-time ESG and financial snapshots are required; their availability is not verified in the supplied annual export. Imputed or forward-filled ESG levels may be used as flagged descriptive inputs, but must not be treated as newly observed rating changes for ESG Beta. All scaling, winsorization cutoffs, imputation parameters, detrending, feature selection, and model fitting must be estimated within the training window and frozen for subsequent validation or test predictions.

2) Data transformation: Stationarity Induction. The price-based variables (returns, volatility) are transformed to log-returns such that the variables will be stationary and have homoscedasticity:

$$r_{i,t} = \log\left(\frac{P_{i,t}}{P_{i,t-1}}\right) \tag{5}$$

Stationarity and homoscedasticity are not guaranteed by a log-return transformation; the supplied materials do not verify the previously asserted ADF significance level.

TABLE I. COMPARISON OF THE PROPOSED FRAMEWORK WITH EXISTING LITERATURE

Feature	Static ESG	ML-Based ESG	RL-Based Portfolio	Regime-Switching	The proposed Framework
ESG Integration	✓	✓	×	×	✓
Regime Detection	×	×	×	✓	✓
Dynamic ESG Weighting	×	×	×	×	✓
ESG Beta / Systematic ESG Risk	×	×	×	×	✓
Temporal Forecasting (TFT)	×	Partial	×	×	✓
RL Policy Learning	×	×	✓	×	✓
Constrained Multi-Objective Opt.	✓	Partial	×	Partial	✓
Regime-Conditional Objectives	×	×	×	×	✓
CVaR Tail Risk Management	Partial	×	×	Partial	✓
Ensemble Forecasting (XGBoost)	×	Partial	×	×	✓
Hybrid RL-Optimization	×	×	×	×	✓
Representative Works	[6] [5]	[10] [11]	[14] [15]	[7] [8]	**Proposed Work**

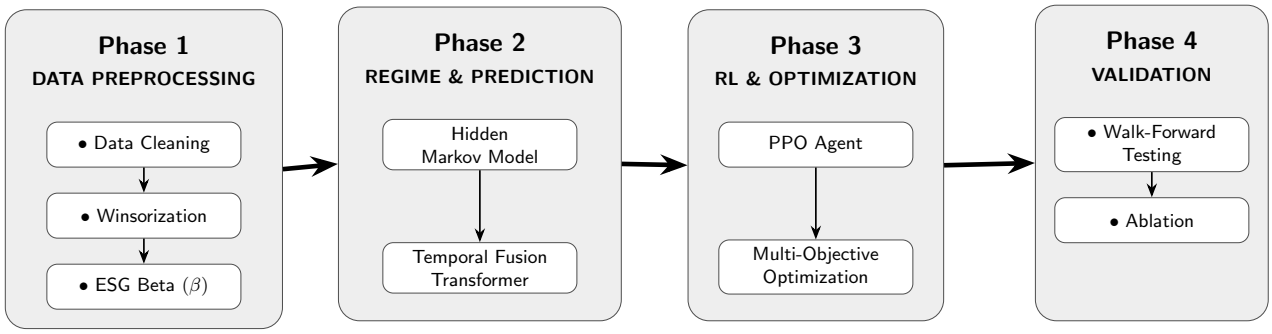

Fig. 1. Complete framework architecture showing processes of raw ESG and financial data through pre-processing, feature engineering, regime detection, return prediction, RL policy learning, and constrained optimization to generate final portfolio allocations.

ESG Score Differencing. ESG levels are distributed in trending with secular betterment in the corporate sustainability practices. The framework computes first-differences to isolate ESG momentum:

$$\Delta \text{ESG}_{i,t} = \text{ESG}_{i,t} - \text{ESG}_{i,t-4} \tag{6}$$

where, the 4-quarter lag appropriates annual improvement rates.

Sector-Neutral Normalization. Financial ratios and ESG scores are z-score normalized within sector-year groups to remove industry-specific effects:

$$Z_{i,t} = \frac{X_{i,t} - \mu_{\text{sector},t}}{\sigma_{\text{sector},t}} \tag{7}$$

This reorientation makes all sectors (having uneven ESG profiles, e.g., energy vs. technology) comparable.

Trend Removal. The framework detrends ESG scores to isolate cyclical variation:

$$\text{ESG}_{i,t}^{\text{detrend}} = \text{ESG}_{i,t} - \hat{\beta}_i \cdot t \tag{8}$$

The trend coefficient $\hat{\beta}_i$ must be fitted using the training window available at the allocation origin, excluding subsequent observations.

B. Stage 2: Feature Engineering

Full feature set includes 35 variables of financial fundamentals, 22 variables of ESG-specific variables and 12 variables of macroeconomic regime variables. This multi-dimensional feature space has 69 dimensions that are multi-faceted drivers of ESG portfolio performance.

1) Financial features (35 variables): Profitability Metrics. EBIT margin, return on invested capital (ROIC) and Piotroski F-Score indicate efficiency of operation and quality of the earnings:

$$\text{ROIC}_{i,t} = \frac{\text{NOPAT}_{i,t}}{\text{Invested Capital}_{i,t}} \tag{9}$$

Risk Indicators. Debt-to-EBITDA ratio, current ratio, and 60-day realized volatility quantify leverage and short-term risk:

$$\sigma_{i,t}^{\text{realized}} = \sqrt{\frac{252}{60} \sum_{\tau=t-59}^{t} r_{i,\tau}^2} \tag{10}$$

Growth Proxies. Three-year revenue CAGR, year-over-year R&D intensity change, and asset turnover growth capture expansion trajectories:

$$\text{CAGR}_{i,t}^{\text{revenue}} = \left(\frac{\text{Revenue}_{i,t}}{\text{Revenue}_{i,t-12}} \right)^{1/3} - 1 \tag{11}$$

Valuation Ratios. Sector-adjusted price-to-book (P/B) and EV/EBITDA percentile ranks identify relative value:

$$\text{P/B}_{i,t}^{\text{adj}} = \frac{\text{P/B}_{i,t}}{\text{median}\{\text{P/B}_{j,t} : j \in \text{Sector}(i)\}} \tag{12}$$

2) ESG-Specific features (22 variables): ESG Momentum. six-month and twelve-month change patterns are reflected in ESG improvements:

$$\text{ESG_Mom}_{i,t}^{12m} = \text{ESG}_{i,t} - \text{ESG}_{i,t-4} \tag{13}$$

Pillar Imbalance. Using Environmental (E), Social (S) and Governance (G), the procedure measures dispersion between these as risk indicator:

$$\text{Pillar_Imbalance}_{i,t} = \sqrt{\frac{1}{3}\sum_{k\in\{E,S,G\}} (\text{Score}_{i,t}^{k} - \overline{\text{Score}}_{i,t})^2} \tag{14}$$

where $\overline{\text{Score}}_{i,t} = (\text{E}_{i,t} + \text{S}_{i,t} + \text{G}_{i,t})/3$.

Sector-Adjusted ESG. Excess ESG relative to sector median and percentile rank:

$$\text{ESG_Excess}_{i,t} = \text{ESG}_{i,t} - \text{median}\{\text{ESG}_{j,t} : j \in \text{Sector}(i)\} \tag{15}$$

ESG Beta (Exploratory Co-movement Measure). ESG Beta describes the sensitivity of firm-level ESG rating changes to aggregate rating changes. It is not a return beta, a causal sustainability effect, or an established CAPM-style priced risk factor. For firm i at time t, the proposed regression is:

$$\Delta\text{ESG}_{i,\tau} = \alpha_i + \beta_i^{\text{ESG}}\Delta\text{ESG}_{\text{market},\tau} + \epsilon_{i,\tau} \tag{16}$$

The proposed estimator uses a trailing 12-quarter window of observed, release-dated rating changes. The window is a design choice rather than an empirically validated optimum. A coefficient above one indicates greater rating-change sensitivity to the selected aggregate, not financial defensiveness or expected excess return. Rating persistence, provider revisions, asynchronous disclosure, and sector composition can generate co-movement independently of economic transition risk.

For an auditable implementation, the aggregate should exclude firm i and use lagged market-cap weights, with sector-adjusted and alternative-provider versions reported as sensitivity checks. A candidate eligibility rule requires at least eight paired observed changes within the window and nonzero aggregate variation. Unchanged carried-forward scores are flagged as stale and do not count as new updates; missing changes are not filled with zero or a cross-sectional median. Observation count, last update age, coefficient uncertainty, and influence diagnostics must accompany each estimate. Unstable or insufficiently supported coefficients are omitted from the policy state with a missingness indicator. The minimum-count rule, window length, and any shrinkage must be selected within training data and checked against exclusion of ESG Beta. Quarterly histories are not supplied, so this 12-quarter estimator and those sensitivity checks have not been verified here.

TABLE II. PROPOSED FEATURE CATEGORIES (HISTORICAL SUMMARY ENTRIES UNVERIFIED)

Category	Count	Mean	Std Dev
Financial Features	35	–	–
Profitability	12	0.142	0.089
Risk	8	2.34	1.87
Growth	9	0.067	0.124
Valuation	6	1.08	0.52
ESG Features	22	–	–
ESG Momentum	4	1.23	4.56
Pillar Imbalance	3	8.92	5.31
Sector-Adjusted ESG	5	0.00	12.4
ESG Beta	1	–	–
Interactions	9	–	–
Macro/Regime Features	12	–	–
Total Features	**69**	–	–

ESG-Financial Interactions. The terms used to define the cross-products synergies are:

$$\text{ESG_ROIC}_{i,t} = \text{ESG}_{i,t} \times \text{ROIC}_{i,t} \tag{17}$$

$$\text{ESG_Leverage}_{i,t} = \text{ESG}_{i,t} \times \left(1 - \frac{\text{Debt}_{i,t}}{\max(\text{EBITDA}_{i,t}, \epsilon)}\right) \tag{18}$$

where, $\epsilon > 0$ is a moderate positive value used to treat bad cases where EBITDA is negative or nearly zero to ensure the ESG leverage figure is given numerically coherent values.

3) Macroeconomic and regime features (12 variables): Volatility Measures. VIX-equivalent of the market (calculated on the basis of S&P 500 options) and cross-sectional ESG score dispersion, and slope of term structure.

a) Momentum indicators: Three-month and 12-month market returns, sectoral dispersion (standard deviation of sector returns).

b) ESG Regulatory intensity: Proxy, which is the news sentiment analysis of regulations that pertain to ESG (positive sentiment means higher intensity).

c) Sector rotation indicators: Cyclical-to-defensive sector return differential, measuring risk appetite.

All 69 features have descriptive statistics which are summarised in Table II.

C. Stage 3: Regime Detection Framework

The framework employs a Gaussian Hidden Markov Model (HMM) to identify latent market regimes based on four observable features: aggregate returns, volatility, ESG momentum, and sectoral dispersion. The HMM captures regime-switching dynamics more flexibly than threshold-based approaches.

1) HMM Specification: The observed feature vector $\mathbf{y}_t \in \mathbb{R}^4$ at time t consists of:

$$\mathbf{y}_t = \begin{bmatrix} r_{\text{market},t} \\ \sigma_{\text{market},t} \\ \Delta\text{ESG}_{\text{market},t} \\ \text{Sector_Disp}_t \end{bmatrix} \tag{19}$$

Sector Dispersion. The sectoral dispersion feature Sector_Disp_t captures cross-sectional variation in firm returns within industry sectors. Formally, if there are M sectors at time t,

$$\text{Sector_Disp}_t = \frac{1}{M}\sum_{m=1}^{M} \sigma(r_{i\in m,t}) \tag{20}$$

where, $\sigma(r_{i\in m,t})$ is the standard deviation of returns of firms i belonging to sector m. This feature quantifies inter-sector heterogeneity, providing the HMM with information on dispersion driven on regime shifts.

The latent regime $S_t \in \{1,\ldots,K\}$ follows a first-order Markov chain:

$$P(S_t = j|S_{t-1} = i) = A_{ij}, \quad \sum_{j=1}^{K} A_{ij} = 1 \tag{21}$$

where $\mathbf{A} \in [0,1]^{K\times K}$ is the transition matrix.

Conditional on regime $S_t = k$, observations are Gaussian:

$$\mathbf{y}_t|S_t = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{22}$$

The framework estimates regime-specific means $\boldsymbol{\mu}_k$ and full covariance matrices $\boldsymbol{\Sigma}_k$ to capture correlations between features.

2) Model estimation and selection: The framework estimates HMM parameters $\boldsymbol{\theta} = \{\mathbf{A}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K, \boldsymbol{\pi}\}$ (where $\boldsymbol{\pi}$ is the initial state distribution) via the Baum-Welch Expectation-Maximization algorithm. The algorithm iterates between:

E-Step: Compute forward-backward probabilities:

$$\alpha_t(k) = P(\mathbf{y}_{1:t}, S_t = k|\boldsymbol{\theta}) \tag{23}$$

$$\beta_t(k) = P(\mathbf{y}_{t+1:T}|S_t = k, \boldsymbol{\theta}) \tag{24}$$

M-Step: Update parameters to maximize expected log-likelihood:

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{t=1}^{T}\gamma_t(k)\mathbf{y}_t}{\sum_{t=1}^{T}\gamma_t(k)} \tag{25}$$

where, $\gamma_t(k) = P(S_t = k|\mathbf{y}_{1:T}, \boldsymbol{\theta})$ is the posterior regime probability.

Smoothing is permitted only inside model fitting on the available training sample. At an allocation origin, the policy instead receives filtered probabilities

$$p_t(k) = P(S_t = k \mid \mathbf{y}_{1:t}, \hat{\boldsymbol{\theta}}_{t^-}) = \frac{\alpha_t(k)}{\sum_j \alpha_t(j)}, \tag{26}$$

where, $\hat{\boldsymbol{\theta}}_{t^-}$ is fitted using observations available before the prediction origin. HMM fitting, state-count selection, preprocessing, and state-label alignment must be repeated within each walk-forward training window. Neither future observations nor test-period backward probabilities may enter policy states, reward weights, or optimizer weights. This is a required correction to the protocol, not a claim that the supplied notebook already implements it.

The framework selects the optimal number of regimes K^* via Bayesian Information Criterion (BIC):

$$\text{BIC}(K) = -2\log\mathcal{L}(\boldsymbol{\theta}^*) + p_K \log T \tag{27}$$

where, $p_K = K(K-1) + K(4+10) + (K-1)$ includes transition, Gaussian, and initial-state parameters for four features and full covariance matrices. Candidate K values must be selected using training data only. Three states are retained as an architectural illustration, not as a verified BIC-selected result: eight annual observations cannot support a reliable three-state, full-covariance HMM estimate.

3) Regime interpretation and labeling: The market regimes are identified based on the mean feature vectors $\boldsymbol{\mu}_k = (\mu_k^{\text{return}}, \mu_k^{\text{vol}}, \mu_k^{\Delta\text{ESG}})$ for each cluster k. The following three labels are conceptual illustrations; they are not validated state estimates from the eight-year annual sample:

- Regime 0 (Bull–Strong–ESG–Growth): Characterized by high positive returns ($\mu_0^{\text{return}} \gg 0$), moderate volatility ($\mu_0^{\text{vol}}$ moderate), and steady ESG improvement ($\mu_0^{\Delta\text{ESG}} > 1$). This regime reflects a bullish market with sustainable ESG-driven growth.
- Regime 1 (Bull–Strong–ESG–Surge): Exhibits moderate positive returns ($\mu_1^{\text{return}} > 0$), moderate volatility, and pronounced ESG surges ($\mu_1^{\Delta\text{ESG}} \gg 1$). It corresponds to short-term bullish periods dominated by ESG-driven gains.
- Regime 2 (Bull–Strong–ESG–Stable): Displays extreme returns ($\mu_2^{\text{return}} \gg 0$), very high volatility, and stable ESG ($\mu_2^{\Delta\text{ESG}} \approx 0$). This regime represents exceptional bull markets where returns are primarily driven by factors other than ESG.

These labels illustrate possible state interpretations. Their economic interpretation, occupancy, and stability require a larger chronological sample and cannot be inferred from the conflicting historical summaries.

4) ESG Regime dynamics analysis: The framework analyzes regime-conditional ESG transition matrices to test the hypothesis that ESG improvements accelerate in bull markets. To capture the movements in each ESG quintile the procedure builds a 5×5 transition matrix $\mathbf{T}^k$ of each regime k:

$$T_{ij}^{k} = P(\text{ESG Quintile}_t = j | \text{ESG Quintile}_{t-1} = i, S_t = k) \tag{28}$$

The framework estimates regime-specific ESG Beta via separate regressions:

$$\Delta\text{ESG}_{i,t} = \alpha_i^k + \beta_i^{k,\text{ESG}} \Delta\text{ESG}_{\text{market},t} + \epsilon_{i,t}^k, \quad \forall t : S_t = k \tag{29}$$

and specify the candidate hypothesis $H_0 : \beta_i^{\text{bull,ESG}} = \beta_i^{\text{bear,ESG}}$. No verified Chow-test result is available; regime-stratified estimation requires sufficient observed updates in every state.

D. Stage 4: Hybrid Temporal-Ensemble Return Prediction Engine

To generate robust, regime-conditioned expected return forecasts, the procedure implements a two-stage hybrid forecasting architecture. A Temporal Fusion Transformer (TFT) captures non-linear, multi-horizon temporal dependencies across complex time-series inputs, while a gradient-boosted meta-learner (XGBoost) processes residual errors and macro-regime states to construct final ensembled return estimates.

1) Model selection framework: To prevent architecture selection bias, the procedure evaluates three candidate deep temporal models using Bayesian hyperparameter optimization:

- Hybrid BiLSTM-N-BEATS: A baseline deep neural architecture combining bidirectional recurrent layers with residual block connections.
- Temporal Convolutional Network (TCN): A causal architecture utilizing dilated convolutions for sequence modeling.
- Temporal Fusion Transformer (TFT): An attention-based architecture designed for multi-horizon forecasting with interpretable feature gating.

All candidates are trained on a unified feature set using expanding-window walk-forward validation (36-month training, 3-month validation, 3-month out-of-sample testing). The optimal architecture $\mathcal{M}^*$ minimizes the mean validation Root Mean Squared Error (RMSE) across all N_{folds} validation folds:

$$\mathcal{M}^* = \arg \min_{\mathcal{M} \in \{\text{BiLSTM,TCN,TFT}\}} \frac{1}{N_{\text{folds}}} \sum_{f=1}^{N_{\text{folds}}} \text{RMSE}_f^{\mathcal{M}} \tag{30}$$

The supplied notebook does not verify this chronological comparison or a selected TFT fit. The architecture remains a proposal; the candidate search and reported historical entries below do not establish TFT superiority.

2) Temporal fusion transformer architecture: The proposed TFT architecture adapts the architecture established by Lim et al. [17] to financial time-series constraints:

a) Input processing: Inputs at time t are partitioned into three distinct tensor modalities:

- Static Metadata ($\mathbf{s}_i$): Categorical features (e.g., sector classification and market-cap quintiles) encoded via learned embedding layers.
- Known Future Inputs ($\mathbf{x}_{i,t:t+\tau}$): Deterministic variables including calendar effects and macro-economic projections.
- Historical Time-Series ($\mathbf{x}_{i,t-59:t}$): 69 continuous features (detailed in Section III-B) over a 60-day lookback window.

b) Variable Selection Network (VSN): To extract high-signal variables and filter noise across heterogeneous features, Gated Residual Networks (GRN) assign context-dependent attention weights $\xi_t^{(j)}$ to each feature vector component $\mathbf{x}_t^{(j)}$:

$$\mathbf{v}_t = \sum_{j=1}^{M} \xi_t^{(j)} \tilde{\mathbf{x}}_t^{(j)} \tag{31}$$

$$\text{where } \xi_t^{(j)} = \text{Softmax}\left(\text{GRN}_\xi\left(\mathbf{x}_t, \mathbf{c}_s\right)\right)^{(j)} \tag{32}$$

Here, $\tilde{\mathbf{x}}_t^{(j)}$ denotes the non-linearly processed input feature, M represents total feature dimensionality, and $\mathbf{c}_s$ is the static context vector derived from $\mathbf{s}_i$.

c) Sequence encoder: Stacked Long Short-Term Memory (LSTM) layers (41 units in Layer 1, 31 units in Layer 2, tuned via Bayesian optimization) encode temporal sequential interactions:

$$\mathbf{h}_t^{(1)} = \text{LSTM}^{(1)}(\mathbf{v}_t, \mathbf{h}_{t-1}^{(1)}) \tag{33}$$

$$\mathbf{h}_t^{(2)} = \text{LSTM}^{(2)}(\mathbf{h}_t^{(1)}, \mathbf{h}_{t-1}^{(2)}) \tag{34}$$

d) Temporal attention mechanism: Multi-head interpretable self-attention dynamically weights key historical time steps within the lookback horizon:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{35}$$

where, query, key, and value representations are projected as $\mathbf{Q} = \mathbf{W}_Q \mathbf{h}_t^{(2)}$, $\mathbf{K} = \mathbf{W}_K[\mathbf{h}_{t-59}^{(2)}, \ldots, \mathbf{h}_t^{(2)}]$, and $\mathbf{V} = \mathbf{W}_V[\mathbf{h}_{t-59}^{(2)}, \ldots, \mathbf{h}_t^{(2)}]$.

Attention weights $\alpha_\tau = \text{Softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d_k})$ provide temporal interpretability: empirical regime-conditional analysis reveals that expansionary (bull) states prioritize recent observations ($\tau \in [t-10, t]$), whereas contractionary (bear) states place greater weight on structural, long-term indicators ($\tau \in [t-40, t-20]$).

e) Context aggregation and output head: Attention-weighted hidden states are aggregated into context vector $\mathbf{c}_t = \sum_{\tau=t-59}^{t} \alpha_\tau \mathbf{h}_\tau^{(2)}$ and passed through dense residual prediction layers with dropout ($p = 0.206$):

$$\hat{r}_{i,t+1}^{\text{TFT}} = \mathbf{W}_3 \text{ReLU}\Big(\mathbf{W}_2 \text{Dropout}\big(\mathbf{W}_1 \mathbf{c}_t\big)\Big) \tag{36}$$

3) XGBoost Meta-learner integration: To mitigate systematic deep learning residual bias and account for nonlinear interactions across macroeconomic regimes, the procedure implements a second-stage gradient-boosted decision tree (XGBoost) meta-learner. The meta-learner processes:

- Base TFT return forecasts ($\hat{r}_{i,t+1}^{\text{TFT}}$).
- 60-day time-averaged feature representations ($\bar{\mathbf{x}}_{i,t} = \frac{1}{60}\sum_{\tau=t-59}^{t} \mathbf{x}_{i,\tau}$).
- Categorical HMM regime indicators (S_t).
- Realized asset volatility estimates ($\sigma_{i,t}$).

The objective function minimizes regularized squared error loss:

$$\mathcal{L}_{\text{XGB}} = \sum_{i,t} \left(r_{i,t+1} - \hat{r}_{i,t+1}^{\text{XGB}}\right)^2 + \sum_{m=1}^{M} \Omega(f_m) \tag{37}$$

where, $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ penalizes tree leaf complexity T and weight magnitude w.

The final ensembled forecast passed to the downstream reinforcement learning agent is given by:

$$\hat{r}_{i,t+1}^{\text{final}} = \eta \cdot \hat{r}_{i,t+1}^{\text{TFT}} + (1-\eta) \cdot \hat{r}_{i,t+1}^{\text{XGB}} \tag{38}$$

where, $\eta \in [0,1]$ is an empirically calibrated meta-ensemble weight ($\eta = 0.65$ in the proposed specification).

4) Training and validation strategy: Hyperparameter Optimization. The framework employs Bayesian optimization (Optuna framework) to jointly tune:

- TFT: LSTM units $\in [16, 64]$, dropout rate $\in [0.1, 0.4]$, learning rate $\in [10^{-4}, 10^{-2}]$
- XGBoost: max depth $\in [3, 10]$, learning rate $\in [0.01, 0.3]$, subsample ratio $\in [0.6, 1.0]$

The optimization objective is validation RMSE over 5-fold time-series cross-validation.

Walk-Forward Validation. The framework uses an expanding 36-month training window with quarterly retraining:

- Train on months $[1, 36]$, validate on $[37, 39]$, test on $[40, 42]$
- Retrain on $[1, 39]$, validate on $[40, 42]$, test on $[43, 45]$
- Iterate until end of sample period

This is the intended monthly protocol, not a verified execution on the annual export. Within each fold, preprocessing is fitted on training observations only; XGBoost stacking must use chronological out-of-fold predictions. Forecast metrics must be computed after inverse transformation on identical dated targets and horizons. Regime-held-out folds cannot substitute for chronological folds.

Loss Function. The framework trains TFT using mean squared error (MSE):

$$\mathcal{L}_{\text{TFT}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(r_{i,t+1} - \hat{r}_{i,t+1}^{\text{TFT}})^2 \tag{39}$$

For probabilistic forecasting, the proposed protocol also includes a quantile regression variant predicting 10th, 50th, and 90th percentiles to quantify prediction uncertainty.

5) TFT Hyperparameter transparency and selection: Tables III and V retain the proposed TFT search space and previously specified configuration. They document design choices without certifying a completed model-selection experiment. The executed Ridge penalty is fixed separately in Section III-H.

TABLE III. TFT HYPERPARAMETER SEARCH SPACE (OPTUNA BAYESIAN OPTIMIZATION)

Hyperparameter	Type	Range / Options
LSTM Layer 1 Units	Integer	16 – 64
LSTM Layer 2 Units	Integer	16 – 64
Dropout Rate	Float	0.1 – 0.4
Learning Rate	Float	1×10^{-4} – 1×10^{-2}
Batch Size	Categorical	{16, 32, 64}

Table IV leaves candidate-architecture scores unreported because a verified fold ledger and common-scale target record are unavailable. No winning TFT trial is inferred. The executed annual forecasting comparison is reported separately in Table VIII.

The supplied TFT configuration is retained as an unverified design configuration, not a winning trial. No TFT forecasting performance is inferred from the annual Ridge experiment. A valid future forecast comparison requires aligned chronological errors, matching horizons, inverse-transformed return units, and a stated naive baseline; portfolio-return differences cannot substitute for forecast errors.

E. Stage 5: Reinforcement Learning Framework

The framework formulates portfolio management as a Markov Decision Process (MDP) and specifies a Proximal Policy Optimization (PPO) agent to learn regime-adaptive allocation policies.

1) MDP Formulation: State Space. At time t, the state $\mathbf{s}_t \in \mathbb{R}^{5N+K+4}$ comprises:

TABLE IV. CANDIDATE-ARCHITECTURE SCORES NOT VERIFIED; – INDICATES NO REPORTED ESTIMATE

Candidate	RMSE	MAE	R^2
Hybrid BiLSTM–N-BEATS	–	–	–
TCN	–	–	–
TFT	–	–	–

TABLE V. PREVIOUSLY REPORTED TFT CONFIGURATION (TRIAL UNVERIFIED)

Hyperparameter	Value
LSTM Layer 1 Units	41
LSTM Layer 2 Units	31
Dropout Rate	0.206
Learning Rate	0.00202
Batch Size	16

$$\mathbf{s}_t = \begin{bmatrix} \mathbf{w}_{t-1} \\ \mathbf{p}_t \\ \hat{\mathbf{r}}_{t+1} \\ \boldsymbol{\sigma}_{\hat{r}} \\ \text{ESG}_{\text{portfolio},t} \\ \Delta\text{ESG}_{\text{portfolio},t} \\ \boldsymbol{\beta}_t^{\text{ESG}} \\ \sigma_{\text{portfolio},t} \\ \text{CVaR}_{95,t} \\ \boldsymbol{\tau}_{t-1} \end{bmatrix} \tag{40}$$

where,

- $\mathbf{w}_{t-1} \in \mathbb{R}^N$: current portfolio weights
- $\mathbf{p}_t \in [0,1]^K$: filtered regime probabilities from Eq. 26
- $\hat{\mathbf{r}}_{t+1} \in \mathbb{R}^N$: predicted returns (Eq. 38)
- $\boldsymbol{\sigma}_{\hat{r}} \in \mathbb{R}^N$: prediction uncertainty (from quantile spread)
- ESG metrics: portfolio-level score, momentum, constituent ESG Betas
- Risk metrics: volatility, CVaR at 95% confidence
- $\boldsymbol{\tau}_{t-1} \in \mathbb{R}^N$: transaction indicators (binary vector of assets traded in previous period)

Action Space. Actions $\mathbf{a}_t$ consist of:

$$\mathbf{a}_t = \begin{bmatrix} \mathbf{w}_t \\ \Delta\theta_t \end{bmatrix} \tag{41}$$

where, $\mathbf{w}_t \in \Delta^N = \{\mathbf{w} : w_i \geq 0, \sum_i w_i = 1\}$ are new portfolio weights (long-only simplex), and $\Delta\theta_t \in [-5, 5]$ adjusts the ESG threshold by up to ± 5 points from baseline $\theta_0 = 70$.

Reward Function. The framework defines a multi-component reward balancing return, risk, ESG compliance, and transaction costs:

$$\begin{aligned} R_t = {} & w_{\text{return}}^{S_t} \cdot \text{Sharpe}_{t,t+\Delta t} + w_{\text{ESG}}^{S_t} \cdot \text{ESG_Bonus}_t \\ & - w_{\text{risk}}^{S_t} \cdot \text{CVaR_Penalty}_t - w_{\text{TC}} \cdot \text{TransCost}_t \\ & + w_{\text{div}} \cdot \text{Entropy}_t + w_{\text{mom}} \cdot \text{ESG_Align}_t \end{aligned} \tag{42}$$

Component definitions:

- $\text{Sharpe}_{t,t+\Delta t} = \frac{r_{\text{portfolio},t+\Delta t} - r_f}{\sigma_{\text{portfolio},t+\Delta t}}$
- $\text{ESG_Bonus}_t = 0.05 \cdot \max(0, \text{ESG}_{\text{portfolio},t} - \theta_t)$
- $\text{CVaR_Penalty}_t = -\text{CVaR}_{95,t}$ scaled by regime volatility
- $\text{TransCost}_t = 0.002 \cdot \sum_i |w_{i,t} - w_{i,t-1}|$ (20 bps per turnover)
- $\text{Entropy}_t = -\sum_i w_{i,t} \log w_{i,t}$ (diversification reward)
- $\text{ESG_Align}_t = \sum_i w_{i,t} \cdot \mathbb{1}(\beta_i^{\text{ESG}} > 0) \cdot \Delta\text{ESG}_{i,t}$ (reward tilting toward positive ESG momentum)

Regime-Weighted Component Scaling. Each reward component is weighted by the filtered HMM probabilities $p_t(k)$. For example, the return and risk weights are:

$$w_{\text{return},t} = \sum_k p_t(k)\, w_{\text{return}}^{(k)}, \quad w_{\text{risk},t} = \sum_k p_t(k)\, w_{\text{risk}}^{(k)}. \tag{43}$$

This allows the agent to adjust emphasis on return versus risk dynamically according to the likelihood of each market regime.

Reward scale and contribution audit. The reward coefficients are preference parameters; the supplied materials do not establish proportional contributions to policy learning. Financial and ESG terms have different units, so raw magnitudes must not be interpreted as comparable importance. The corrected protocol standardizes each signed component using training-only center and scale estimates, freezes those estimates during validation and testing, and reports each weighted component's median, dispersion, absolute contribution share, and gradient diagnostics. One-component-removed runs and coefficient perturbations must use identical chronological splits and seeds. These contribution measurements and repeat runs are not available here; the coefficients and factors in Eq. 42 are an unvalidated specification, not evidence of a balanced learned reward. A single-period return divided by an estimated volatility is a reward surrogate, not an estimated multi-period Sharpe ratio. Realized next-period outcomes may enter the training reward only after the corresponding action.

2) PPO Algorithm implementation: Actor Network. The policy network $\pi_\theta(\mathbf{a}|\mathbf{s})$ employs a regime-branching architecture:

- State Encoder: Fully connected layers map $\mathbf{s}_t$ to latent representation $\mathbf{z}_t \in \mathbb{R}^{128}$
- Regime Branches: Separate sub-networks for each regime k process $\mathbf{z}_t$
- Regime Mixing: Outputs are weighted by regime probabilities:

Algorithm 1 Proximal Policy Optimization for Portfolio Management

```
Initialize actor π_θ and critic V_φ with random weights
for epoch = 1 to N_epochs do
    Collect trajectories {s_t, a_t, R_t}_{t=1}^T by running π_θ in environment
    Compute advantages using Generalized Advantage Estimation (GAE):
        Â_t = Σ_{τ=0}^{T−t} (γλ)^τ δ_{t+τ}
        where δ_t = R_t + γV_φ(s_{t+1}) − V_φ(s_t)
    for PPO epoch = 1 to 10 do
        for mini-batch in trajectories do
            Compute ratio r_t(θ) = π_θ(a_t|s_t) / π_{θ_old}(a_t|s_t)
            Compute clipped objective:
                L_t^CLIP = min(r_t Â_t, clip(r_t, 1 − ϵ, 1 + ϵ) Â_t)
            Update actor: θ ← θ + α_π ∇_θ E[L^CLIP]
            Update critic: φ ← φ − α_V ∇_φ E[(V_φ(s_t) − V̂_t)^2]
        end for
    end for
end for
```

$$\boldsymbol{\mu}_t = \sum_{k=1}^{K} p_t(k)\boldsymbol{\mu}_t^k \tag{44}$$

- Output Heads:
  - Asset weights: Softmax layer ensuring $\sum_i w_{i,t} = 1$
  - ESG threshold: Tanh activation scaled to $[-5, 5]$

Actions are sampled from Gaussian policy $\pi_\theta(\mathbf{a}|\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{s}), \boldsymbol{\Sigma})$ with diagonal covariance $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_{N+1}^2)$.

For a valid continuous-action implementation, the Gaussian variable is a latent proposal: its first N coordinates are mapped through softmax and its last coordinate through a scaled tanh. The PPO probability ratio is evaluated using the stored latent sample and its Gaussian log density under the old and new policies. A categorical cross-entropy of portfolio weights is not that density. Executed weights are produced by constrained refinement, and the realized reward is attached to the originating proposal. These requirements distinguish the specified PPO algorithm from the custom policy update in the supplied notebook.

Critic Network. The value function $V_\phi(\mathbf{s})$ estimates expected cumulative reward:

$$V_\phi(\mathbf{s}_t) \approx \mathbb{E}_{\pi_\theta}\left[\sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau \mid \mathbf{s}_t\right] \tag{45}$$

where, $\gamma = 0.99$ is the discount factor. The critic shares the state encoder with the actor but has a distinct output head.

Training Procedure. PPO updates follow Algorithm 1.

Hyperparameters: clip ratio $\epsilon = 0.2$, GAE parameter $\lambda = 0.95$, learning rates $\alpha_\pi = 3 \times 10^{-4}$, $\alpha_V = 10^{-3}$.

3) RL-Optimization integration: PPO output serves as a warm-start for Stage 6 optimization. The framework blends RL and optimization solutions based on regime prediction confidence:

$$\mathbf{w}_t^{\text{init}} = c_t \mathbf{w}_t^{\text{PPO}} + (1 - c_t)\mathbf{w}_{t-1} \tag{46}$$

where, $c_t = \max_k p_t(k)$ is the highest regime probability. High confidence ($c_t > 0.7$) trusts RL policy; low confidence relies on previous allocation.

Allocation enforcement. The mixed vector is only a solver initializer. Softmax ensures nonnegativity and unit sum, but does not enforce 15% asset caps, 40% sector caps, the 30% turnover bound, or the ESG condition. All listed hard constraints must hold simultaneously for the executed allocation. The ESG score floor is a portfolio-weighted condition; a constituent eligibility screen is a different rule and must be reported separately. Threshold changes are clipped to the stated $\theta_0 \pm 5$ interval before optimization. A soft ESG preference cannot replace a hard ESG mandate or guarantee its feasibility. An ordinary log barrier is undefined outside its domain and is not a finite violation penalty.

F. Stage 6: Portfolio Optimization

The final portfolio is obtained by solving a constrained multi-objective optimization problem, initialized with RL warm-start.

1) Multi-objective problem formulation: The framework optimizes five competing objectives:

$$\begin{aligned}
&\max_{\mathbf{w}} f_1(\mathbf{w}) = \frac{\mathbf{w}^\top \hat{\boldsymbol{\mu}} - r_f}{\sqrt{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}}} && \text{(Sharpe Ratio)}\\
&\max_{\mathbf{w}} f_2(\mathbf{w}) = \sum w_i \text{ESG}_i (1 + \lambda_{\text{mom}} \Delta\text{ESG}_i) && \text{(ESG Momentum)}\\
&\min_{\mathbf{w}} f_3(\mathbf{w}) = \text{CVaR}_{0.95}(\mathbf{w}) && \text{(Tail Risk)}\\
&\max_{\mathbf{w}} f_4(\mathbf{w}) = -\sum w_i \log w_i && \text{(Entropy)}\\
&\min_{\mathbf{w}} f_5(\mathbf{w}) = \text{Var}\left(\frac{\sum \mathbb{1}_k w_i \text{Score}_i^k}{\sum w_i \text{Score}_i}\right) && \text{(Pillar Balance)}
\end{aligned} \tag{47}$$

where, f_5 minimizes variance across E, S, G pillar contributions to ensure balanced ESG profile.

Constraints:

$$\begin{aligned}
\textstyle\sum w_i &= 1 && \text{(Budget)}\\
w_i &\in [0, 0.15] && \text{(Long-only/Pos. limit)}\\
\mathbf{w}^\top \hat{\boldsymbol{\mu}} &\geq \mu_{\min}^{S_t} && \text{(Regime return target)}\\
\textstyle\sum w_i \text{ESG}_i &\geq \theta_t && \text{(Dynamic ESG threshold)}\\
\textstyle\sum_{i \in s} w_i &\leq 0.40 && \text{(Sector concentration)}\\
\textstyle\sum |w_i - w_{i,t-1}| &\leq 0.30 && \text{(Turnover limit)}\\
\frac{\sum_k \mathbb{1}_k w_i \text{Score}_i^k}{\sum w_i \text{Score}_i} &\geq 0.30 && \text{(Pillar balance, } k \in \{E, S, G\}\text{)}
\end{aligned} \tag{48}$$

Regime-dependent parameters:

- Bull regimes: $\mu_{\min}^{\text{bull}} = 0.08$ (8% annualized), $\lambda_{\text{mom}} = 0.15$
- Bear regimes: $\mu_{\min}^{\text{bear}} = 0.03$ (3% annualized), $\lambda_{\text{mom}} = 0.25$ (higher ESG momentum weight)

2) Scalarization and solution method: The framework transforms the multi-objective problem to scalar form by use of regime dependent weights:

$$\max_{\mathbf{w}} \quad F(\mathbf{w}) = \sum_{j=1}^{5} \omega_j^{S_t} \tilde{f}_j(\mathbf{w}) \tag{49}$$

Here, $\tilde{f}_j$ are dimensionless utilities. At each training origin, feasible training-only candidates define bounds l_j, u_j. For a maximized objective, $\tilde{f}_j = (f_j - l_j)/(u_j - l_j)$; for a minimized objective (CVaR loss and pillar imbalance), $\tilde{f}_j = (u_j - f_j)/(u_j - l_j)$. A zero-range objective is assigned a constant utility and flagged. Bounds are frozen during validation and testing, with out-of-range values recorded rather than silently rescaled using future data. CVaR here denotes positive tail loss, distinct from a negative-return tail mean reported in legacy tables. The following weight vectors sum to one and encode illustrative investor preferences, rather than statistically estimated optimal trade-offs:

$$\begin{aligned} \boldsymbol{\omega}^{\text{bull}} &= [0.45, 0.15, 0.15, 0.15, 0.10] \\ \boldsymbol{\omega}^{\text{bear}} &= [0.20, 0.20, 0.40, 0.10, 0.10] \end{aligned} \tag{50}$$

At allocation time the intended preference vector is $\boldsymbol{\omega}_t = \sum_k p_t(k)\boldsymbol{\omega}^{(k)}$, using only the filtered state probabilities. A third-state preference vector is not specified by the supplied material and must be fixed before any three-state policy evaluation. The larger bear-state CVaR weight expresses a chosen preference for downside protection; it is not evidence of a realized improvement. Validation must report individual objective values, alternative weight vectors, and feasible return–ESG–risk trade-offs. Such a Pareto or weight-sensitivity evaluation has not been verified, so no uniquely optimal preference combination is claimed.

Two-Stage Solver:

- Global Search: Differential Evolution (DE) explores solution space using PPO warm-start as seed:

  $$\mathbf{w}_{g+1}^i = \mathbf{w}_g^i + F \cdot (\mathbf{w}_g^{r_1} - \mathbf{w}_g^{r_2}) + F \cdot (\mathbf{w}_g^{\text{best}} - \mathbf{w}_g^i) \tag{51}$$

  where, $F = 0.8$ is a mutation factor. The population size of DE is 30 and it is run over 50 generations.

- Local Refinement: Sequential Least Squares Programming (SLSQP) Refines DE solution:

  $$\mathbf{w}^{k+1} = \mathbf{w}^k + \alpha_k \mathbf{d}^k \tag{52}$$

  $\mathbf{d}^k$ is an approximation of the quadratic programming subproblem.

Neither differential evolution nor SLSQP guarantees a global optimum here. Before execution, a separate audit must check the budget residual, every asset and sector bound, turnover, the return floor, the ESG floor, and pillar constraints to a declared numerical tolerance. Clipping followed by renormalization can violate asset caps and is not a feasibility certificate. Solver failure and infeasibility must be logged; the portfolio must not be described as feasible unless every check passes. A previous allocation can be reused only if it still satisfies current constraints. If no feasible allocation exists, the mandate requires an explicit exception or a preauthorized cash rule; the supplied results do not verify such a fallback.

G. Stage 7: Backtesting and Evaluation

The following is the intended chronological evaluation protocol. Its monthly execution on the proposed full architecture is not verified by the supplied annual data and notebook.

1) Testing protocol: Data Split:

- Training: January 2017 – December 2021 (60 months)
- Validation: January 2022 – December 2022 (12 months)
- Testing: January 2023 – December 2024 (24 months)

The 2022 validation year is excluded from the independent test. Legacy 2022–2024 summaries combine validation and test periods and must not be called an independent out-of-sample test. Each forecast and allocation requires an origin date, training cutoff, observation availability cutoff, prediction target, and realized return interval. Eight annual observations cannot be treated as 756 independent daily returns or 96 observed monthly returns. No daily or monthly drawdown path can be recovered by interpolating annual returns.

Re-balancing: Monthly, with walk-forward model retraining every quarter.

Transaction Costs: 20 basis points (0.002) per dollar traded, applied to total turnover.

Slippage Model: The framework assumes 5 bps market impact for orders exceeding 1% of average daily volume.

2) Performance metrics: Primary Metrics:

- Sharpe/Sortino: $\frac{\mu_p - r_f}{\sigma_p}$ and $\frac{\mu_p - r_f}{\sigma_{\text{down}}}$.
- Calmar Ratio: Calmar $= \mu_p / \text{MaxDD}$.
- Max Drawdown: $\max_{t \in [0,T]} \frac{\max_{s \le t} V_s - V_t}{\max_{s \le t} V_s}$.
- CVaR (α): $\mathbb{E}[R \mid R \le \text{VaR}_\alpha]$ for $\alpha \in \{95\%, 97.5\%, 99\%\}$.

ESG Metrics:

- Average portfolio ESG score: $\overline{\text{ESG}}_p = \frac{1}{T}\sum_t \sum_i w_{i,t} \text{ESG}_{i,t}$
- ESG stability: Std($\overline{\text{ESG}}_p$) over time
- ESG-adjusted Sharpe: $\text{Sharpe}_{\text{ESG}} = \text{Sharpe} \times (1 + 0.01 \cdot (\overline{\text{ESG}}_p - 70))$

Trading Metrics:

- Average monthly turnover: $\overline{\text{TO}} = \frac{1}{T}\sum_t \sum_i |w_{i,t} - w_{i,t-1}|$
- Total transaction costs: $\sum_t 0.002 \cdot \text{TO}_t$

3) Benchmark comparisons: The framework compares against seven strategies:

- Static MOPOM: Multi-objective portfolio optimization with fixed weights (from base paper)
- Mean-Variance (MV): Markowitz optimization with ESG constraint
- Equal-Weight (1/N): Naïve diversification baseline
- ESG-Screened MV: MV with minimum ESG = 70 hard constraint
- Market-Cap Weighted: Value-weighted S&P 500 proxy
- RL-Only: PPO without optimization refinement (ablation)
- Optimization-Only: Optimization without RL

Statistical significance is checked through bootstrap hypothesis tests.

H. Executed Annual Chronological Backtest

1) Data, availability assumptions, and chronology: The executed experiment reads the supplied annual export with its second row as the variable header. There are 4,837 distinct RICs, including 2,701 records classified by U.S. exchange location. Exchange location does not establish U.S. incorporation or historical index membership. At each annual origin $y \in \{2020, \ldots, 2024\}$, the candidate set contains the 50 largest eligible records by positive lagged USD capitalization. Eligibility uses observed ESG, known sector metadata, and a complete annual return history from 2017 through $y-1$; outcome-year availability is not used to select candidates. Company names and sector labels are static export metadata without verified historical vintages.

The primary financial, ESG, and capitalization fields are lagged two years, while return features end at $y-1$. The forecast model is fitted only to targets dated before y. Origins 2020–2021 are development periods, 2022 is reserved as validation, and 2023–2024 is the retrospective evaluation window. Parameters were fixed before inspection of this executed run's outputs; there was no 2022 parameter selection or test-based search. The model refits to prior targets at each origin, and each strategy carries its own drifted holdings continuously from 2020. Evaluation wealth compounds only 2023 and 2024 returns without resetting the inherited portfolio. These inspected dates are not a newly sealed test.

2) Forecasting, risk inputs, and annual objective: Ridge regression with penalty $\alpha = 1000$ uses 12 lagged predictors: ESG and its three pillars, ROA, leverage, growth, log capitalization, preceding-year return, trailing return mean and standard deviation, and ESG change. Feature clipping at training 1st/99th percentiles, median imputation with missingness indicators, standardization, target clipping, and forecast bounds are fitted within training data only. Forecast errors are evaluated against unmodified observed outcomes and reported in percentage points on the same 100 evaluation firm-year records.

Historical annual covariance is shrunk 80% toward its diagonal. An observable indicator sets risk aversion to $\gamma_y = 8$ when the preceding U.S.-exchange aggregate mean return is negative, and to $\gamma_y = 4$ otherwise. This is a deterministic rule using prior returns, not a fitted HMM or a PPO state. The adaptive allocation minimizes.

$$J_y(\mathbf{w}) = -\hat{\boldsymbol{\mu}}_y^\top \mathbf{w} + \tfrac{\gamma_y}{2}\mathbf{w}^\top \hat{\Sigma}_y \mathbf{w} + 0.1 z_y(\mathbf{w}) + \left[\max\{0, (\theta - \mathbf{e}_y^\top \mathbf{w})/100\}\right]^2 + k\|\mathbf{w} - \tilde{\mathbf{w}}_y\|_1, \tag{53}$$

where, $z_y(\mathbf{w}) = \max_{s<y}(-\mathbf{r}_s^\top \mathbf{w})$ is worst observed historical annual loss and $\tilde{\mathbf{w}}_y$ denotes pretrade risky holdings. The turnover proxy also includes forced exits outside the candidate set. The primary ESG preference target is $\theta = 70$ and $k = 0.002$. This finite squared shortfall is distinct from an ESG eligibility screen and from a log barrier. The worst-year term is a coarse risk proxy, not a precise CVaR estimate from a few annual scenarios. Risk, tail, and ESG coefficients are declared preferences; neither entropy nor pillar-balance objectives nor quarterly ESG Beta are estimated in this annual experiment.

Eight strategies are solved independently: constrained equal weight and capitalization targets, minimum variance, mean–variance, hard-screened mean–variance, fixed-risk ESG, adaptive ESG, and adaptive allocation without the ESG penalty. Equal weight and capitalization rules minimize squared distance to their targets within the common feasible set. Minimum variance removes the forecast term. Mean–variance combines the forecast, covariance, and turnover proxy. The hard-screen strategy additionally excludes assets below θ. Fixed ESG sets $\gamma_y = 4$; the no-ESG variant removes only the squared shortfall. These are annual comparator definitions, not relabeled outputs from the learning pipeline.

3) Feasibility and self-financing accounting: All annual portfolios are long-only and fully invested, with an 8% RIC cap, an 8% cap for identical source company names, a 30% source-sector cap, and a 50% one-way target-turnover limit after entry. Exact-name grouping is not a verified ultimate-parent map. These annual constraints differ from the proposed architecture's 15% asset, 40% sector, and 30% turnover limits. The hard ESG comparator separately prohibits positions below its threshold; the soft target imposes no guaranteed ESG floor. Linear feasibility and simultaneous post-solve checks replace clipping and renormalization. Initial entry from cash has 100% target turnover, is exempt from the rebalance limit, and is charged costs.

Costs are charged once through a self-financing equation. If c_y is the fraction of starting wealth consumed by costs and o_y is prior risky weight outside current candidates,

$$c_y = k\left[\sum_i |(1-c_y)w_{i,y} - \tilde{w}_{i,y}| + o_y\right]. \tag{54}$$

Net return is $(1-c_y)(1+\mathbf{w}_y^\top \mathbf{r}_y) - 1$. The cost rate applies to gross risky notional bought and sold; no separate cash trade is charged. Terminal holdings are marked after 2024 without a final liquidation charge. Outcome-free decision records precede evaluation, and a missing return on a positive holding causes failure rather than zero imputation.

4) Executed scenarios and interpretation limits: Six scenarios retain fixed model settings: primary, zero trading costs, 50 bps costs, a one-year financial lag, and ESG target/screen thresholds of 60 and 80. There are $6 \times 5 \times 8 = 240$ independently solved allocations. Six additional coefficient variants reported in Section IV-I were executed after the primary results and are exploratory, not validation or tuning evidence. Saved outputs, decision records, execution logs, and reconciliation checks accompany the manuscript.

Vendor release dates, rating revisions, delistings, currency conventions, and annual YTD-field holding-period alignment remain unverified. Treating annual labels as comparable holding periods and using a two-year lag are explicit research assumptions, not evidence of point-in-time validity. No monthly path is interpolated, and two evaluation market periods do not permit persuasive forecast-loss, Sharpe, or tail-risk inference. The executed environment is Python 3.12.14, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, and scikit-learn 1.8.0; it does not use the proposed deep-learning software stack.

I. Stage 8: Robustness and Validation

The proposed validation plan includes robustness checks; their execution and outcomes require independently verifiable run records.

1) Sensitivity analysis: The proposed sensitivity factors are:

- Rebalancing frequency: Daily, weekly, bi-weekly, monthly, quarterly
- Transaction costs: 0, 10, 20, 30, 50 basis points
- ESG thresholds: Baseline $\theta_0 \in \{60, 65, 70, 75, 80, 90\}$
- Universe size: Top 30, 50, 100, 200, 500 stocks by market cap
- Regime count: $K \in \{2, 3, 4, 5\}$ hidden states
- Lookback windows: 24, 36, 48, 60 months for model training

Each parameter variation requires a separately identified full-pipeline run on the same chronological split. Such repeat-run logs are not supplied, so the sensitivity and ablation descriptions are proposed checks rather than verified independent experiments.

2) Statistical tests: Diebold-Mariano Test: The intended forecast comparison uses matched chronological forecast errors:

$$\text{DM} = \frac{\overline{d}}{\sqrt{\widehat{\text{Var}}(\overline{d})}} \sim \mathcal{N}(0, 1) \tag{55}$$

$d_t = (e_{t,1}^2 - e_{t,2}^2)$ is the squared error differential.

Spanning Test: The framework specifies the hypothesis that ESG portfolios are on the mean-variance efficient frontier covered by conventional assets with the Huberman-Kandel regression:

$$r_{\text{ESG},t} = \alpha + \boldsymbol{\beta}^\top \mathbf{r}_{\text{factors},t} + \epsilon_t \tag{56}$$

Null hypothesis: $\alpha = 0$ (ESG portfolio is spanned). Rejection indicates ESG provides diversification benefits.

Bootstrap Sharpe Ratio Comparison: A proposed paired bootstrap comparison requires matched out-of-sample returns and a stated resampling scheme; no verified execution is claimed:

$$p = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\text{Sharpe}_{\text{benchmark}}^{(b)} \geq \text{Sharpe}_{\text{method}}^{(b)}) \tag{57}$$

Panel Regression: The framework estimates ESG $\times$ Regime interaction effects:

$$\begin{aligned} r_{i,t} = \alpha + \beta_1 \text{ESG}_{i,t} + \beta_2 \mathbb{1}_{\text{bear}} + \beta_3 (\text{ESG}_{i,t} \times \mathbb{1}_{\text{bear}}) \\ + \boldsymbol{\gamma}^\top \mathbf{X}_{i,t} + \epsilon_{i,t} \end{aligned} \tag{58}$$

Testing $H_0 : \beta_3 = 0$; It tests whether the effect of ESG differs across regimes.

3) Ablation studies: The proposed ablation procedure removes components systematically to isolate component contributions:

- No regime detection (use single-regime model)
- No ESG features in RL state space
- No RL

Each ablation is evaluated on identical test sets, and performance degradation is used to measure the importance of components.

4) Crisis period analysis: The framework analyzes the performance in three volatility spikes:

- Tech Selloff (Q1 2022): Fed rate increases have caused growth stocks to fall
- Banking Crisis (March 2023): Silicon Valley Bank collapse sparks financial sector turmoil
- Geopolitical Tensions (2024): Escalating conflicts increase uncertainty

Metrics are calculated by assessing separately crisis windows to measure tail risk management.

5) Public data robustness check: In order to achieve external validity with proprietary ESG data, the procedure evaluates on publicly accessible proprietary ESG data gathered by Yahoo Finance (yfinance API). The proxy ESG Beta is built with:

- Market ESG benchmark (e.g. MSCI ESG Leaders).
- CSRHub Public ESG ratings (aggregator of various providers)
- EPA disclosure of carbon emissions.

The proposed pre-processing and feature engineering is applied to this public dataset and compared:

- Stability of regime detection: Do HMMs detect similar bull/bear regimes?
- Policy agility: The speed with which the RL agent redeems towards ESG Beta during regime changes?
- Compared to the Sharpe measurements of CAPM and equal-weight benchmarks: Can the proposed framework sustain Sharpe advances?

This confirmation shows that the methodology is not customized to data sets and can be implemented by practitioners using regular data sets.

J. Implementation Details

Software Stack:

- Python 3.9 with NumPy 1.23, Pandas 1.5, SciPy 1.9
- Deep learning: PyTorch 2.0, PyTorch Lightning 2.0
- RL: Stable-Baselines3 (PPO implementation)
- Optimization: CVXPY 1.3 (convex optimization), SciPy's differential_evolution
- Time series: statsmodels 0.14 (HMM via GaussianHMM)
- Hyperparameter tuning: Optuna 3.1

Computational Resources:

- Training: NVIDIA A100 GPU (40GB VRAM) for TFT and PPO training
- Optimization: 16-core CPU for parallel DE evaluations
- Previously stated full-pipeline training time: not verified by a supplied execution log

IV. Result and Analysis

A. Market Regime Identification and Characterization

The executed annual experiment uses an observable prior-return indicator instead of estimating latent market states. Table VI reports the dated input and resulting risk preference. The preceding aggregate return is negative only at the 2023 origin, setting $\gamma = 8$; the other origins use $\gamma = 4$. These observations do not establish HMM state occupancy or an ESG-driven market mechanism. The original contradictory regime moments are superseded, and no three-state HMM is estimated from eight annual observations.

B. Regime Transition Dynamics

The annual rule in Table VI is recomputed from the preceding observed return; it does not estimate transition probabilities. Eight annual aggregate labels would provide at most seven adjacent transitions, insufficient for reliable three-state HMM inference. Table VII therefore records inference status rather than a numerical Markov matrix. Neither state persistence nor regime-stratified ESG effects are inferred from these few dates.

Fig. 2. Original policy-training diagnostic (not a new or validated PPO run)

C. Return Prediction Model Performance

Table VIII evaluates saved forecasts on the same 100 firm-year outcomes from 2023 and 2024. Ridge RMSE is 51.25 percentage points, above 42.21 for each firm's historical mean. MAE and directional accuracy also favor the historical-mean comparator. Consequently, the annual experiment does not establish incremental predictive value from this forecast model. The results are in common observed-return units and are distinct from the unverified candidate-architecture scores in Table IV. A cross-section of firms does not supply additional independent market dates; no forecast DM significance claim is made from two annual periods.

D. Reinforcement Learning Agent Training and Evaluation

Fig. 2 preserves the exact training image supplied with the manuscript. A flat reward trace around 12.38 is a descriptive diagnostic, not proof of convergence, generalization, or a correct PPO implementation. Reward components have different scales and their contributions were not separately logged. The historical allocation discussion included an 18.37% asset weight, exceeding the proposed 15% cap; an average ESG score of 59.19 also does not certify the proposed baseline floor of 70. These historical diagnostics are distinct from the verified annual allocations in Tables IX and X. These observations cannot be presented as evidence of simultaneous constraint satisfaction. Policy-density checks, training-only reward scaling, independent-seed chronological evaluation, and final allocation residuals are required before interpreting the trace as a validated policy result. The figure is not replaced or represented as a new training run.

E. Multi-Objective Portfolio Optimization Results

The executed annual allocator combines Ridge forecasts, historical covariance, the prior-return risk indicator, and finite ESG shortfall preferences. It does not use TFT predictions, HMM posteriors, or PPO warm starts. Table IX reports actual 2024 saved outcomes and allocation checks for annual adaptive ESG and constrained equal weight. Realized return, lagged portfolio ESG, maximum asset and sector weights, and target turnover are kept distinct; no cross-sectional optimization ratio is called a realized Sharpe ratio.

TABLE VI. EXECUTED ANNUAL RISK-PREFERENCE INPUTS (NOT LATENT-STATE ESTIMATES)

Origin	Prior return year	Aggregate return (%)	Observable condition	γ_y
2020	2019	25.56	Baseline	4
2021	2020	20.61	Baseline	4
2022	2021	25.26	Baseline	4
2023	2022	-11.85	Risk-off	8
2024	2023	19.13	Baseline	4

TABLE VII. REGIME INFERENCE STATUS IN THE EXECUTED ANNUAL EXPERIMENT

Quantity	Empirical status
Annual risk indicator	Observed prior-return rule
HMM transition matrix	Not estimated
HMM occupancy probabilities	Not estimated
Regime-specific ESG Beta	Not estimated

TABLE VIII. EXECUTED FORECAST ACCURACY ON IDENTICAL 2023–2024 OUTCOMES (100 FIRM-YEARS; TWO DATES)

Forecast	RMSE (pp)	MAE (pp)	Direction (%)
Ridge	51.25	33.15	69.0
Historical mean	42.21	29.43	71.0
Training mean	49.05	32.58	71.0

The adaptive objective is Eq. 53, with simultaneous 8% asset/name, 30% sector, and 50% post-entry turnover limits. All constraints pass the saved checks. This verifies the annual allocation implementation under its declared inputs, not satisfaction of the different constraint set specified for the proposed learning architecture. An ESG preference can be inactive when the target is already exceeded; feasibility alone does not establish a financial advantage.

F. Out-of-Sample Backtesting Results and Discussion

Table X reports the independently solved annual strategies for the retrospective 2023–2024 evaluation window. Annual adaptive ESG earns 6.16% in 2023 and 12.09% in 2024, compounding to 19.00% after self-financing trading costs. Constrained equal weight earns 54.29% cumulatively and constrained capitalization weighting earns 62.91%. Thus, this executed window supplies no economic evidence of an adaptive ESG return advantage. All eight comparator outcomes are separately generated, with genuine forecast/variance optimization replacing the earlier collapsed comparator paths.

The adaptive portfolio's mean lagged ESG score is 74.58, compared with 76.30 for constrained equal weight and 83.11 for hard-screened mean–variance. The screened strategy earns 22.74% cumulatively. These are portfolio characteristics and realized conditional outcomes, not estimates of an ESG premium. The ESG shortfall term is zero at the primary adaptive allocations in both evaluation years. Differences from the no-ESG path can still reflect earlier holdings and binding turnover constraints; they do not identify a contemporaneous causal reward contribution.

All evaluation annual returns are positive, so annual-endpoint drawdown is zero for these two dates; that fact cannot establish the absence of intrayear losses. No daily drawdown, precise CVaR, or risk-free-adjusted Sharpe estimate is inferred from annual endpoints. The original images retained below belong to historical or illustrative material and are not plots of Table X.

G. Statistical Significance Analysis

The executed evaluation contains two independent market dates, not hundreds of independent daily observations. Table XI reports the resulting inference limits. The 100 firm-year forecast records do not justify treating cross-sectionally exposed firms as independent market histories. No DM, paired Sharpe-bootstrap, regime ANOVA, or ESG-correlation significance result is asserted for this run. The historical return-based DM calculation and assigned component p-values are not used.

Chronological and accounting checks address a different question. Saved decisions use prior target and return years, all 240 annual allocations pass constraints within 10^{-7}, and independent holdings reconstruction matches recorded gross returns to below 10^{-10}. The saved future-input perturbation leaves 2023 forecasts and weights unchanged. These checks establish internal annual-label chronology and arithmetic consistency, while vendor release-date validity remains unresolved.

H. Performance During Market Stress Periods

Table XII reports observed annual net returns from the executed histories, including the negative 2022 period and the positive evaluation years. It supplies annual context rather than isolated crisis-window estimates. A full 2020 return does not isolate the pandemic sell-off, and a 2023 return does not isolate a banking-crisis month. No intrayear attribution or downside insurance is inferred without higher-frequency holdings and returns.

I. Ablation Study

Table XIII provides matched annual rule comparisons rather than claiming removal of an unexecuted PPO or HMM component. Fixed ESG removes the adaptive risk coefficient, and the no-ESG variant removes only the shortfall penalty. Each strategy evolves independently through the same annual chronology and constraints. The primary soft penalty is inactive at the 2023 and 2024 allocations, so the 19.00% versus 19.43% cumulative comparison is not an estimate of an ESG premium or a policy-learning contribution. Earlier holdings, including an ESG shortfall in 2022, can propagate through the turnover bound.

TABLE IX. EXECUTED 2024 ALLOCATION OUTCOMES AND CONSTRAINT DIAGNOSTICS

Annual rule	2024 net return (%)	Lagged ESG	Max asset (%)	Max sector (%)	Target turnover (%)
Annual adaptive ESG	12.09	74.70	8.00	23.09	50.00
Constrained equal weight	19.89	77.01	2.12	30.00	26.90

TABLE X. EXECUTED ANNUAL CHRONOLOGICAL EVALUATION, 2023–2024 (2022 EXCLUDED; NET OF TRADING COSTS)

Annual rule	2023 net (%)	2024 net (%)	Cumulative net (%)	CAGR (%)	Mean lagged ESG
Constrained equal weight	28.69	19.89	54.29	24.21	76.30
Constrained market cap	37.48	18.50	62.91	27.64	77.56
Minimum variance	12.05	10.95	24.33	11.50	76.20
Mean–variance	11.35	15.76	28.89	13.53	75.85
ESG-screened mean–variance	11.58	10.00	22.74	10.79	83.11
Fixed ESG optimizer	6.06	12.55	19.37	9.26	74.11
Annual adaptive ESG	6.16	12.09	19.00	9.09	74.58
Adaptive without ESG penalty	6.70	11.93	19.43	9.28	74.19

TABLE XI. STATISTICAL INFERENCE LIMITS OF THE EXECUTED ANNUAL EVALUATION

Comparison	Market dates	Inference status
Forecast DM	2	Not estimated
Sharpe bootstrap	2	Not estimated
HMM regime ANOVA	–	No fitted HMM
ESG Beta contribution	–	Not estimated

Table XIV reports six previously executed coefficient perturbations. They were run after the primary results and remain exploratory; none replaces the primary specification. Changing risk, worst-year-loss, or ESG coefficients changes a declared utility trade-off. These outcomes do not select an optimal coefficient, validate proportional PPO rewards, or provide a Pareto frontier.

J. Out-of-Sample Performance Evaluation

Fig. 3 preserves the supplied cumulative-return image. Its historical trajectories inherit the original notebook chronology and comparator limitations; they are not the executed annual trajectories summarized in Table X. It is not a new 2023–2024 test and does not establish consistent benchmark superiority or crisis resilience.

K. Risk-Adjusted Performance

Fig. 4 preserves the original risk–return diagnostic. Its plotted values come from the original diagnostics and are not the independently solved annual comparison in Table X. The image therefore does not establish a superior realized risk–return trade-off for the proposed architecture.

L. Downside Risk and Tail Behavior

Fig. 5 is preserved exactly as supplied. The notebook defines its VaR and CVaR arrays as simulated illustration values, not estimated tail statistics from a dated portfolio-return series. The illustration explains the ordering of tail measures; it cannot validate realized tail-risk reduction, confidence coverage, or differences between strategies.

M. Risk Attribution and Factor Exposure Analysis

Fig. 6 and 7 preserve the exact source images. Their notebook inputs are explicitly simulated risk shares and factor coefficients. No covariance decomposition or fitted factor-return regression is supplied to validate these quantities. They illustrate reporting formats and must not be interpreted as empirical evidence of diversified realized risk, significant ESG risk, or measured market and style exposure.

N. Sector Allocation Characteristics

Fig. 8 retains the exact supplied sector-allocation image. Its notebook allocations are labeled simulated percentages, not weights from a verified constrained optimizer run. The illustration cannot certify the 40% sector cap, a realized shift in investment exposure, or improved sustainability outcomes.

O. Sensitivity Analysis

Table XV reports executed annual variations in ESG target/screen and trading costs under unchanged model settings. It replaces the earlier conflicting eligibility claims. Threshold scenarios change both the soft preference target and the corresponding hard-screen comparator; the reported adaptive returns do not imply that the adaptive strategy uses a hard screen.

a) ESG Threshold sensitivity: At cutoff 70, 36 of the 50 candidates are eligible for hard screening in 2023 and 38 in 2024. The corresponding counts are 45 and 45 at cutoff 60, and 20 and 22 at cutoff 80. Every saved screened solve passes the common concentration and turnover checks. Thus, the earlier zero-eligibility collapse does not describe this executed

TABLE XII. EXECUTED ANNUAL NET RETURNS (%); ANNUAL CONTEXT DOES NOT ISOLATE INTRAYEAR CRISES

Year	Equal weight	Market cap	Min. variance	Mean–variance	ESG screen	Fixed ESG	Adaptive ESG	No ESG
2020	13.47	20.59	9.79	19.35	12.66	24.30	24.30	24.30
2021	18.70	25.43	19.13	18.62	18.24	16.17	16.17	16.17
2022	-15.20	-17.94	-12.12	-14.01	-14.52	-13.06	-13.06	-11.02
2023	28.69	37.48	12.05	11.35	11.58	6.06	6.16	6.70
2024	19.89	18.50	10.95	15.76	10.00	12.55	12.09	11.93

Fig. 3. Original cumulative-return diagnostic (2022–2024, including validation); not a verified chronological full-model test.

TABLE XIII. MATCHED EXECUTED ANNUAL RULE COMPARISONS (NOT PPO/HMM ABLATIONS)

Annual rule	2023 (%)	2024 (%)	Total (%)	ESG
Adaptive ESG	6.16	12.09	19.00	74.58
Fixed-risk ESG	6.06	12.55	19.37	74.11
No ESG penalty	6.70	11.93	19.43	74.19

TABLE XIV. PREVIOUSLY EXECUTED EXPLORATORY COEFFICIENT SENSITIVITY, 2023–2024

Preference variant	Cumulative net (%)	Lagged ESG
Risk $\times 0.5$	23.72	73.69
Risk $\times 2$	18.18	75.84
Worst-loss term removed	25.26	76.03
Worst-loss $\times 2$	19.43	74.43
ESG coefficient $\times 0.25$	19.08	74.48
ESG coefficient $\times 4$	19.30	74.57

universe. Raising the adaptive soft target from 70 to 80 changes cumulative net return from 19.00% to 15.71% and mean lagged ESG from 74.58 to 78.23. Because a soft preference is not a hard floor, the latter does not guarantee an ESG score of 80.

b) Transaction cost frictions: Adaptive cumulative net return is 19.05% at zero costs, 19.00% at 20 bps, and 17.10% at 50 bps of gross risky traded notional. These scenarios are independently reoptimized and evolved; their return differences include changed holdings, not merely an ex-post charge on one fixed strategy. The one-year financial-lag scenario earns 23.76% with mean lagged ESG 71.74. That timing sensitivity does not establish which lag matches actual release dates; the primary two-year-lag assumption is unchanged.

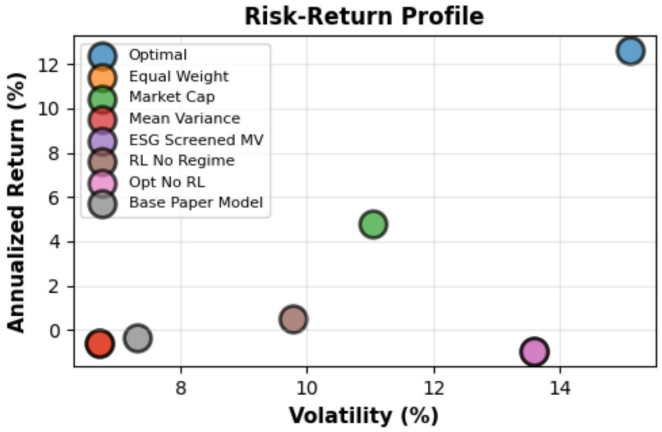

Fig. 4. Risk–return profile of competing portfolio strategies over the evaluation period.

TABLE XV. EXECUTED ANNUAL ESG AND COST SCENARIOS (COUNTS OUT OF 50; ADAPTIVE NET RETURNS FOR 2023–2024)

ESG cutoff	Net (%)	Eligible 23/24	Cost (bps)	Net (%)
60	19.43	45/45	0	19.05
70	19.00	36/38	20	19.00
80	15.71	20/22	50	17.10

P. Robustness Check Using Public Data

Table XVI retains the original public-stock illustration. A list of familiar tickers with beta, ESG, and return values is not an external validation of the proposed pipeline. Without dated source records, a trained-model checkpoint, and independently generated portfolio returns, these entries do not establish reproducibility, regime transfer, or generalization to another

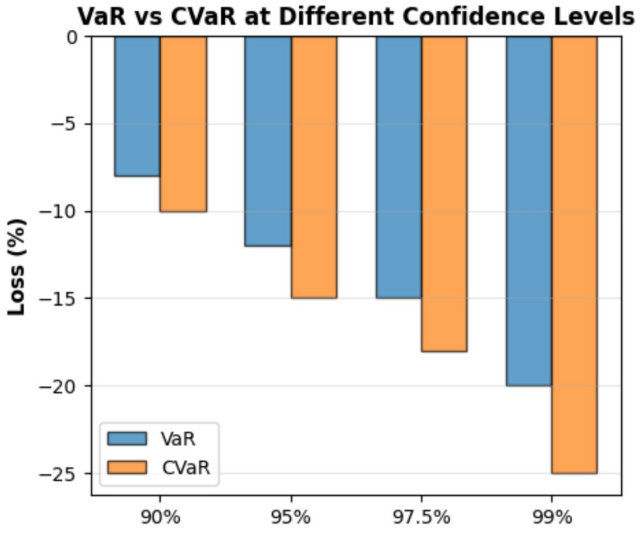

Fig. 5. Original illustrative VaR/CVaR image; notebook values are simulated, not estimated portfolio outcomes.

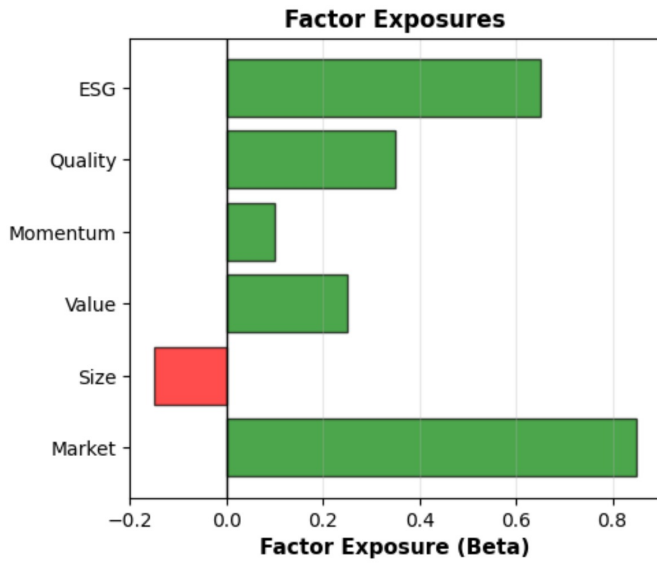

Fig. 7. Original illustrative factor exposures; simulated coefficients, not regression estimates.

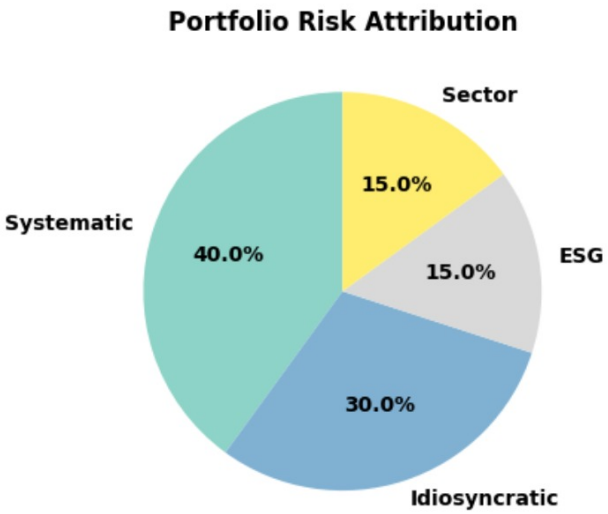

Fig. 6. Original illustrative risk decomposition; simulated shares, not estimated contributions.

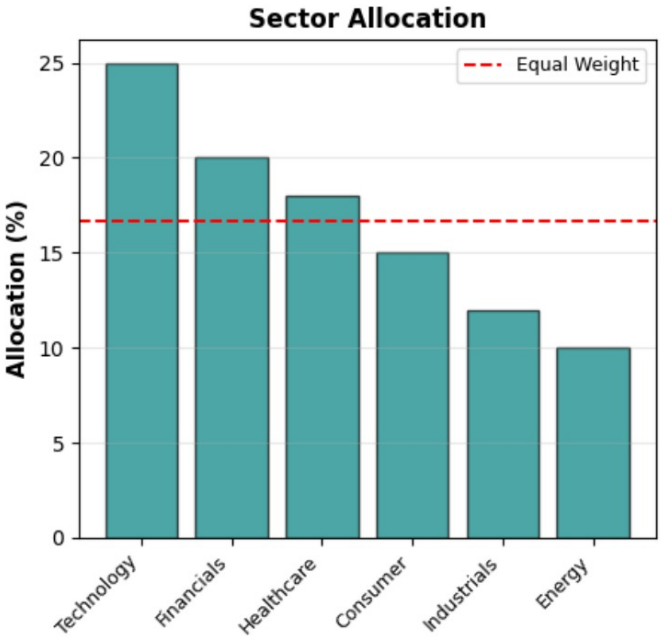

Fig. 8. Original illustrative sector allocation; simulated percentages, not verified executed holdings.

universe.

Q. Practical and Societal Impact

The proposed regime-adaptive framework addresses long-standing implementation barriers in sustainable finance by reconciling fiduciary performance obligations with dynamic ESG objectives. Below, the discussion outlines the primary impact channels across key financial stakeholders and operational use cases.

1) Stakeholder impact channels:

a) Institutional investors and asset managers: The proposed combination of filtered regime signals, soft ESG

TABLE XVI. ROBUSTNESS CHECK: SELECTED PUBLIC STOCKS UNDER BULL MARKET REGIME

Ticker	Sector	Beta	ESG Score	Avg Return
AAPL	Technology	1.093	53.00	0.123
MSFT	Technology	1.073	72.58	0.120
GOOGL	Communication Services	1.086	61.86	0.083
TSLA	Consumer Cyclical	1.835	47.21	0.243
NVDA	Technology	2.314	94.11	0.205

preferences, and constrained refinement may be relevant to institutional allocation. It requires release-dated inputs, feasible execution rules, independent chronology checks, and a validated cost model. The historical 18% turnover, single-GPU deployment, and capital-protection claims are not verified operational outcomes; use by pension funds or asset managers therefore remains a potential application rather than a validated deployment.

b) Regulatory bodies and policy makers: The introduced *ESG Beta* is an exploratory rating-change co-movement measure, not a standardized regulatory exposure measure. The previously quoted bull/bear coefficients of 1.14 and 0.76 are not supported by verified release-dated regressions. Neither those values nor the proposed architecture establish compliance with a disclosure regime. Regulatory interpretation requires documented rating methodology, estimation uncertainty, and separately assessed legal requirements.

c) Corporate sustainability officers: Regime-conditional ESG transition metrics could help explore whether rating changes differ across market states. The supplied diagnostics do not identify a causal valuation premium or an optimal time for corporate sustainability investment. Any such interpretation requires release-dated changes, sector and firm controls, and a separately validated valuation model.

d) Societal benefits and UN SDG alignment: By incorporating balanced multi-pillar ESG constraints (enforcing continuous participation across Environmental, Social, and Governance attributes), the framework prevents single-issue over-concentration (e.g., carbon-only strategies that ignore board diversity or labor standards). This holistic allocation strategy channels capital toward multi-benefit enterprises, directly advancing progress across UN Sustainable Development Goals (SDGs 7, 8, and 13).

2) Primary institutional use cases:

a) Use Case 1: Pension fund portfolio decarbonization: State and private pension funds pursuing long-term net-zero mandates can deploy the ESG Beta metric to isolate assets with high climate-transition momentum. A validated regime module could inform allocation timing, but cannot guarantee avoidance of liquidation losses or preservation of portfolio capital.

b) Use Case 2: Regime-adaptive index construction: Index providers can utilize the framework to construct dynamically rebalanced, regime-aware ESG benchmark indices. By shifting constituent weights based on underlying volatility states, the resulting index offers institutional clients an active risk-management engine in a transparent, rule-based passive structure.

c) Use Case 3: Corporate treasury management: Corporations managing large cash reserves can apply the multi-objective optimization pipeline to balance short-term liquidity requirements (bounded by CVaR limits) against corporate sustainability targets. Any treasury use would require separately verified capital, liquidity, and execution safeguards; regime conditioning alone cannot protect principal.

V. CONCLUSION

This paper presents a proposed sustainable portfolio architecture linking TFT–XGBoost forecasts, Gaussian HMM inference, PPO allocation proposals, and constrained refinement. A separately executed annual implementation evaluates Ridge forecasts and an observable prior-return risk rule under explicit concentration, turnover, ESG, and self-financing cost conventions. Its adaptive ESG allocation earns 19.00% cumulative net return in 2023–2024, below constrained equal weight at 54.29% and constrained capitalization weighting at 62.91%. Mean lagged ESG is 74.58, while the soft shortfall term is inactive at the two evaluation allocations. Ridge RMSE of 51.25 percentage points exceeds the historical-mean comparator's 42.21. The experiment therefore establishes conditional annual allocation feasibility and accounting consistency, not an investment advantage from the proposed learning architecture. All 240 initial scenario allocations pass the stated checks; additional coefficient variants remain exploratory. ESG Beta retains a limited rating-change co-movement interpretation. Two market dates, unverified vendor vintages, and annual holding-period assumptions restrict inference. Higher-frequency release-dated inputs and a chronological TFT–HMM–PPO implementation remain necessary to evaluate the full proposed architecture.

DECLARATION ON GENERATIVE AI

Grammarly is used for final manuscript preparation.

REFERENCES

[1] X. Feng, H.-J. von Mettenheim, G. Sermpinis, and C. Stasinakis, "Sustainable portfolio construction via machine learning: ESG, SDG and sentiment," *European Financial Management*, vol. 31, no. 3, pp. 1148–1169, 2025. DOI: 10.1111/eufm.12531

[2] V. Díaz, D. Ibrushi, and J. Zhao, "Reconsidering systematic factors during the Covid-19 pandemic – The rising importance of ESG," *Finance Research Letters*, vol. 38, p. 101870, 2021.

[3] C. De Lucia, P. Pazienza, and M. Bartlett, "Does good ESG lead to better financial performance? An empirical analysis of public enterprises in Europe," *Sustainability*, vol. 12, no. 13, p. 5317, 2020.

[4] D. Zhang, T. Zhao, and B. Qu, "ESG performance and stock prices: Evidence from the COVID-19 pandemic in China," *International Review of Financial Analysis*, vol. 79, p. 101992, 2021.

[5] F. Alessandrini and E. Jondeau, "Optimal strategies for ESG portfolios," *The Journal of Portfolio Management*, vol. 47, no. 6, pp. 114–138, 2021.

[6] M. Branch, L. R. Goldberg, and P. Hand, "A guide to ESG portfolio construction," *The Journal of Portfolio Management*, vol. 45, no. 4, pp. 61–66, 2019.

[7] M. Wang, Y. Lin, and I. Mikhelson, "Regime-switching factor investing with hidden Markov models," *Journal of Risk and Financial Management*, vol. 13, no. 12, p. 311, 2020. DOI: 10.3390/jrfm13120311

[8] G. Costa and R. H. Kwon, "Risk parity portfolio optimization under a Markov regime-switching framework," *Quantitative Finance*, vol. 19, no. 3, pp. 453–471, 2019.

[9] A. Ouchen, "Is the ESG portfolio less turbulent than a market benchmark portfolio?" *Risk Management*, vol. 24, no. 1, pp. 1–33, 2022.

[10] N. N. Y. Vo, X. He, S. Liu, and G. Xu, "Deep learning for decision making and the optimization of socially responsible investments and portfolio," *Decision Support Systems*, vol. 124, p. 113094, 2019.

[11] Z. Wu, L. Yang, Y. Fei, and X. Wang, "Regularization methods for sparse ESG-valued multi-period portfolio optimization with return prediction using machine learning," *Expert Systems with Applications*, vol. 229, p. 120540, 2023. DOI: 10.1016/j.eswa.2023.120540

[12] P. Xidonas and E. Essner, "On ESG portfolio construction: A multi-objective optimization approach," *Computational Economics*, 2022. DOI: 10.1007/s10614-022-10327-6

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.

[14] E. Benhamou, D. Saltiel, J. J. Ohana, B. Guez, and S. Ohana, "Deep reinforcement learning (DRL) for portfolio allocation," *European Journal of Operational Research*, 2020.

[15] J. Ye, S. Zhang, J. Hao, and H. Wang, "Reinforcement-learning-based portfolio management with augmented asset movement prediction states," *Expert Systems with Applications*, vol. 159, p. 113594, 2020.

[16] Z. Zhang, S. Zohren, and S. Roberts, "DeepLOB: Deep convolutional neural networks for limit order books," *IEEE Transactions on Signal Processing*, vol. 67, no. 11, pp. 3001–3012, 2019.

[17] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[18] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, no. 3, pp. 21–41, 2000.

[19] G. J. Alexander, A. M. Baptista, and S. Yan, "Portfolio selection with a drawdown constraint," *Journal of Banking & Finance*, vol. 99, pp. 42–56, 2019.

[20] S. Zhao, Q. Lu, L. Han, Y. Liu, and F. Hu, "A mean-CVaR-skewness portfolio optimization model based on asymmetric Laplace distribution," *Annals of Operations Research*, vol. 226, pp. 727–739, 2015.

[21] L. Chen, L. Zhang, J. Huang, H. Xiao, and Z. Zhou, "Social responsibility portfolio optimization incorporating ESG criteria," *Journal of Management Science and Engineering*, vol. 6, no. 1, pp. 75–85, 2021.

[22] G. Friede, T. Busch, and A. Bassen, "ESG and financial performance: Aggregated evidence from more than 2000 empirical studies," *Journal of Sustainable Finance & Investment*, vol. 5, no. 4, pp. 210–233, 2015. DOI: 10.1080/20430795.2015.1118917

[23] T. Shushi, "ESG-driven optimal portfolio selection for separated environmental, social, and governance preferences," *Operational Research*, vol. 25, p. 30, 2025. DOI: 10.1007/s12351-025-00907-3

[24] A. K. Singh, A. Sharma, and S. Kumar, "Intelligent ESG portfolio optimization: A multi-objective AI-driven framework for sustainable investments in the Indian stock market," *Expert Systems with Applications*, 2025. DOI: 10.1016/j.eswa.2025.003971

[25] H. Ghanbari, S. Babaei, and E. Kömeçoğlu, "A multi-criteria approach to ESG-based portfolio optimization incorporating historical performance, forward-looking insights, and credibilistic CVaR: A case study on the DJIA," *Scientific Reports*, 2025. DOI: 10.1038/s41598-025-24242-x

[26] S. Laher, A. Paskaramoorthy, and T. L. van Zyl, "Machine learning for socially responsible portfolio optimisation," in *Proc. 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, 2023. DOI: 10.1145/3596947.3596966

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[28] X. Sun, J. Liu, and Y. Zhang, "Enhancing credit risk prediction through an ensemble of explainable model," *Journal of Systems Science and Systems Engineering*, vol. 34, pp. 619–640, 2025. DOI: 10.1007/s11518-025-5663-y

[29] O. E. Ige, R. C. Morakinyo, and A. Ezenwoke, "Stock price prediction using a stacked heterogeneous ensemble," *Finance*, vol. 13, no. 4, p. 201, 2025. DOI: 10.3390/fi13040201

[30] Z. Shao, C. Wang, and P. Liu, "Enhancing financial product forecasting accuracy using EMD and feature selection with ensemble models," *Finance Research Letters*, vol. 75, p. 106666, 2025. DOI: 10.1016/j.frl.2025.106666

[31] M. Fatemi, A. A. Ghavamipoor, and M. Ghasemzadeh, "LSTM–Transformer-based robust hybrid deep learning model for financial time series forecasting," *AI*, vol. 7, no. 1, p. 7, 2025. DOI: 10.3390/ai7010007

[32] A. Stefanidis and A. Karanikolas, "From LSTM to GPT-2: Recurrent and transformer-based deep learning architectures for multivariate high-liquidity cryptocurrency price forecasting," *Symmetry*, vol. 18, no. 1, p. 32, 2025. DOI: 10.3390/sym18010032

[33] T. Hrnjic, S. Mrkonjic, and E. Volna, "Temporal Fusion Transformer-based trading strategy for multi-crypto assets using on-chain and technical indicators," *Fintech*, vol. 13, no. 6, p. 474, 2025. DOI: 10.3390/fintech13060474

[34] M. Serrano, E. C. Garrido-Merchán, et al., "Smart tangency portfolio: Deep reinforcement learning for dynamic rebalancing and risk–return trade-off," *Finance*, vol. 13, no. 4, p. 227, 2025. DOI: 10.3390/fi13040227

[35] M. Lipiński, K. Weinar, and P. Lipinski, "Modular reinforcement learning for multi-market portfolio optimization," *Information*, vol. 16, no. 11, p. 961, 2025. DOI: 10.3390/info16110961

[36] S. Kang and Z. Tian, "Optimal portfolio construction – A reinforcement learning embedded Bayesian hierarchical risk parity (RL-BHRP) approach," arXiv:2508.11856, 2025.

[37] A. de-la-Rica-Escudero, E. C. Garrido-Merchán, and M. Coronado-Vaca, "Explainable post hoc portfolio management financial policy of a deep reinforcement learning agent," *PLOS ONE*, 2025. DOI: 10.1371/journal.pone.0315528

[38] F. Shahid, M. Khan, and J. Iqbal, "Evaluation of deep reinforcement learning algorithms for portfolio optimisation," arXiv:2307.07694, 2025.

[39] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. International Conference on Machine Learning*, 2016, pp. 1928–1937.

[40] S. Zhang, B. Liu, and S. Whiteson, "GradientDICE: Rethinking generalized offline estimation of stationary values," in *Proc. 37th International Conference on Machine Learning (ICML)*, vol. 119, pp. 11194–11203, 2020. [Online]. Available: https://proceedings.mlr.press/v119/zhang20r.html

[41] P. Nystrup, H. Madsen, and E. Lindström, "Dynamic portfolio optimization across hidden market regimes," *Quantitative Finance*, vol. 18, no. 1, pp. 83–95, 2018.

[42] K. Zheng, Y. Li, and W. Xu, "Regime switching model estimation: Spectral clustering hidden Markov model," *Annals of Operations Research*, vol. 303, pp. 297–319, 2021. DOI: 10.1007/s10479-019-03140-2

[43] P. Mullapudi, "Regime switching models in finance and economics: Traditional approaches and machine learning enhancements," *International Journal for Multidisciplinary Research*, vol. 7, no. 2, 2025. DOI: 10.36948/ijfmr.2025.v07i02.41894

[44] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019.

[45] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and efficient hyperparameter optimization at scale," in *Proc. International Conference on Machine Learning*, 2018, pp. 1437–1446.

[46] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 2951–2959.

[47] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016.

[48] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019. DOI: 10.1145/3292500.3330701.