

CBGA-SAC: A CVaR-Constrained Deep Reinforcement Learning Framework for Portfolio Optimization

R Lavanya

SRM Institute of science and technology

Shivani Agarwal

Ajay Kumar Garg Engineering College, India

Vinay Dwivedi

Galgotias University, Greater Noida

Pavan Kumar Varma Kothapalli

SRKR Engineering College

Venkata Sandeep Dhullipalla

Germantown, MD, USA

Megha Sharma

KR Mangalam University, Gurugram

Nisha Kundu

`nisha.kundu@jaipur.manipal.edu`

Manipal University Jaipur

Research Article

Keywords: Financial Risk Management, Portfolio Optimization, Deep Reinforcement Learning, Graph Attention Network, Value-at-Risk

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10940340/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

CBGA-SAC: A CVaR-Constrained Deep Reinforcement Learning Framework for Portfolio Optimization

R Lavanya¹, Dr. Shivani Agarwal², Vinay Dwivedi³, Pavan Kumar Varma Kothapalli⁴, Venkata Sandeep Dhullipalla⁵, Megha Sharma⁶, Nisha Kundu^{7*}

¹ Assistant Professor, Department of Computing Technologies
SRM Institute of science and technology, Kattankulathur, TN, India.
Email: lavanya27382@gmail.com

² Associate Professor, Ajay Kumar Garg Engineering College, Ghaziabad, India
Email: kasishivani@gmail.com

³ Assistant Professor, School of Computer Science and Engineering, Galgotias University, Greater Noida, India
Email: Vinaydwvd@gmail.com

⁴ Assistant Professor, Department of Computer Science and Engineering, SRKR Engineering College,
Bhimavaram, Andhra Pradesh, India – 534204 ORCID: 0000-0001-8497-0626
Email: kdvpkvarma@gmail.com

⁵ Principal Engineer, Germantown, MD, USA
Email: dvsandeep599@gmail.com

⁶ Assistant Professor, SOET, Center of Excellence – Cloud Computing,
K.R. Mangalam University, Gurugram, India
Email: sharmamegha027@gmail.com

^{7*} Assistant Professor, Department of Artificial intelligence and Machine Learning
Manipal University Jaipur, India
Email: nisha.kundu@jaipur.manipal.edu

Corresponding Author: nisha.kundu@jaipur.manipal.edu

Abstract

Models are needed to simultaneously describe the nonlinear temporal evolution, time-varying cross-asset dependency and tail risk, in the context of effective financial risk management and financial portfolio optimization in the presence of rapidly changing market regimes. Current deep learning based solutions usually separate the risk prediction and asset allocation process and make use of static covariance structures that makes them less responsive to volatility clusters and regime shifts. This paper introduces a unified hybrid architecture: CNN-Bidirectional Long Short-Term Memory network with temporal attention mechanism (CNN-BiLSTM-Attention) for multi-horizon risk feature extraction, a learned graph based on the Graph Attention Network (GAT) to model dynamic inter-asset correlations, and a Conditional Value-at-Risk (CVaR)-constrained Soft Actor-Critic (SAC) reinforcement learning agent generating a continuous portfolio rebalancing across the rolling window and the inference loop. The fused temporal-relational risk representation is employed to predict Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR) as well as to design a risk-adjusted reward function that imposes a penalty on tail losses and transaction costs while learning a policy. The proposed framework, referred to as CBGA-SAC, is tested on a diversified equity asset universe of 30 stocks for the out-of-sample period of 2022-2024 which covers a market regime shift, compared to Markowitz mean-variance optimization, LSTM forecasting and mean-variance overlay, DDPG and TD3 baselines. Experimental results show that CBGA-SAC achieves higher annualized Sharpe ratio, Sortino ratio and cumulative return, and lower maximum drawdown and 95% CVaR compared to the baselines, and an ablation study reveals that the graph-attention module and the CVaR-constrained reward each play a significant role in these improvements, with the latter introducing the highest individual impact on tail-risk measures. The framework also shows a shorter and faster recovery period from a simulated regime shift in 2022, suggesting the architectural robustness to adapt for real-time risk monitoring and rebalancing under the new regime; however, this test was performed in a controlled backtesting environment with idealized order

execution and a fixed order cost, and live-deployment validation, given market impact, slippage and latency constraints remains an open direction. Overall, the results demonstrate that an end-to-end pipeline that learns temporal risk features, dynamic asset correlations, and tail-risk-aware allocation policies provides a more powerful, and partially interpretable, framework for financial risk management, while showcasing the crucial need for real-time live trading validation and non-parametric tail-risk estimation in future work.

Keywords: Financial Risk Management, Portfolio Optimization, Deep Reinforcement Learning, Graph Attention Network, Value-at-Risk.

1. Introduction

The financial markets yield high-frequency, high-dimensional and non-stationary data, in which the risk exposure and optimum capital allocation should be re-evaluated continuously and not at fixed rebalancing periods [1]. The classical methods of risk management such as parametric Value-at-Risk (VaR) and the GARCH-family volatility models, as well as the classical portfolio construction based on Markowitz mean-variance optimization [2] assume that returns are stationary, linear, or elliptically distributed, all of which are often not satisfied under market stress. These constraints have spurred increasing studies on deep learning and reinforcement learning approaches to financial risk prediction and dynamic portfolio management capable of discovering nonlinear temporal patterns directly from the data, with few distributional assumptions.

Even with these advances, many current hybrid deep learning models for finance still separate the two steps of risk estimation and portfolio allocation into two loosely connected stages: a forecasting model is first used to generate point estimates or volatility forecasts, and a second, separate optimizer or heuristic is then employed to turn these forecasts into portfolio weights [3]. This separation can carry forecasting mistakes to the allocation process, fails to capture the inter-asset correlation, which is dynamic and asset-specific, and usually focuses on optimizing for return or variance, but not primarily on tail risk, which is more relevant for risk managers. In addition, there are few approaches that are specifically developed to adapt the risk estimate and allocation policy continuously as new information becomes available in real-time.

In this paper, the authors propose an end-to-end hybrid framework, CBGA-SAC, which combines (i) a CNN-BiLSTM-Attention module to extract multi-scale temporal risk features from asset return and volatility series, (ii) a Graph Attention Network to learn a dynamic, data-driven representation of cross-asset dependency structure in the place of a static covariance matrix, and (iii) a CVaR-constrained Soft Actor-Critic reinforcement learning agent that directly outputs continuous portfolio weights and is trained with a reward function that explicitly penalizes tail risk and transaction costs. The learned temporal and relational representations are passed on to the risk module and are used to optimize portfolio construction and risk assessment in a single step, not two.

This paper makes the following main contributions:

- A hybrid CNN-BiLSTM-Attention structure is proposed to extract interpretable features of financial risk by modeling local patterns, long-range temporal dependency and interpretive attention weighting in real time, multi-horizon tasks.
- The proposed model introduces a Graph Attention Network to directly learn the time-varying, asset-specific correlation structure from the market data, instead of relying on historical estimates of covariance matrix, which is commonly used in classic mean-variance optimization.
- A continuous portfolio optimization (CPO) reinforcement learning formulation using the CVaR objective is proposed, incorporating a reward function which considers portfolio return, tail risk and transaction costs.

In order to identify existing research gaps in financial risk management and portfolio optimization using a hybrid deep learning model, a systematic review of recent papers related to the use of hybrid deep learning in financial risk management and portfolio optimization is carried out, and four concrete research gaps are identified and directly linked to design choices in the proposed framework.

An experimental comparison is provided against baseline models that are based on Markowitz mean-variance optimization, LSTM plus mean-variance overlay, DDPG, and TD3, with evaluation metrics being the Sharpe ratio, the Sortino ratio, the cumulative return, the maximum drawdown, 95% CVaR, and the risk-forecast error.

The rest of this paper is structured as follows. Section 2 provides a summary of the recent literature review. The identified gaps in the literature are summarized in Section 3. Section 4 introduces the proposed CBGA-SAC

methodology, with all the models formulated. The experimental setup is described in section 5. Results of section 6 are reported and discussed. Section 7 presents contributions whereas Section 8 describes conclusion and future directions.

2. Literature review

This section describes recent literature review deep learning in financial risk management and reinforcement learning in portfolio optimization

Choudhary et al. introduce Safe-REPAIR-CVaR, a constraint-aware safe reinforcement learning framework for real-world multi-asset market data-driven portfolio optimization [4]. The method is based on Conditional Value-at-Risk (CVaR) as a hard safety constraint and has a built-in action-repair mechanism that ensures portfolio weights remain feasible while optimizing returns. The framework provides good performance across all the considered indices with respect to risk adjusted performance, but it does not learn explicit graph-based representations of asset connections and temporal risk.

Chohan et al. propose a hybrid CNN–LSTM–GRU approach to financial risk early-warning based on the quarterly financial data of the Chinese listed companies [5]. This model is able to extract features with convolutional layer and also learn from the temporal dimension with LSTM and GRU, and the CNN+LSTM, CNN+GRU and GRU+LSTM models are outperformed by this model with accuracy 93.5%, precision 92.2%, recall 91.8% and F1-score 92.0%. The study does not go so far as to predict portfolio allocation or dynamic rebalancing, however, and concentrates on financial-risk classification.

Choudhary et al. present a risk-adjusted deep reinforcement learning system in which they train three PPO agents based on the log-return, differential-Sharpe and maximum-drawdown rewards, and combine their policies [6]. The multi-reward formulation enhances the risk-return performance of a portfolio; the reported one outperforms the Sensex portfolio with a cumulative returns of 124.83% and a Sharpe ratio of 1.69, with good performance on the Dow and TWSE and IBEX datasets. However, the framework doesn't explicitly capture dynamics graph-based relationships between assets in addition to temporal tail-risk representations.

Shi et al. propose a Hybrid Financial Risk Predictor (HFRP) which is designed by taking CNN and LSTM networks to take advantage of quantitative and qualitative information collected from financial reports [7]. The model minimises the final training and testing losses to 0.0013 and 0.0030, respectively, and significantly reduces the reported credit, liquidity, market and operational risk scores. The strategy is designed to predict risk to financial reports, and as such it does not offer a downstream mechanism to provide portfolio optimization or continuous asset rebalancing.

Cao et al. propose a hybrid framework to forecast VaR by integrating predictions from GARCH-type volatility models, LSTM model and CNN–LSTM model and sentiment derived from online news [8]. Experiments conducted with both S&P 500 and Dow Jones and Nasdaq data reveal that incorporating textual data helps to enhance the forecasting accuracy of VaR, covering the 7-day and 20-day rolling windows. The task is not classification but continuous VaR forecasting, and ACC, sensitivity and specificity are not the key evaluation measures that are used, so a downstream portfolio-allocation policy is not provided for in the study.

CLR-TD3 is a deep reinforcement learning model proposed by Zhang et al. that consists of a CNN–LSTM encoder for historical prices and an attention model LSTM for historical risk indicators [9]. The paper solves the cross-dataset generalisation gap in fraud detection [10]: prior deep-learning models are trained/tested on a single dataset with identical train-test distribution, so their high accuracy doesn't prove they generalize to a different institution, region, or fraud type. AFD-MAR directly tests this by training its multi-agent system on one domain and evaluating on a structurally different second domain, showing performance drops sharply under naive transfer ($F1: 0.878 \rightarrow 0.079$) but partially recovers with lightweight domain adaptation ($F1 = 0.279$). According to the proposed model, the annualized return is 0.240756, the annualized volatility is 0.272689, the Sharpe ratio is 0.929570, the maximum drawdown is -0.322573 , the Sortino ratio is 1.332662 and the Calmar ratio is 0.746361. While a risk information is explicitly introduced, asset correlations are modeled implicitly by not using an explicit graph-attention structure, and no CVaR constraint is imposed.

Jiang et al. suggest an alternative framework of the DRL for portfolio selection in high-dimensional financial markets without requiring a model of the market [11]. The concept of risk aversion and transaction costs is included directly into an extended mean-variance reward, and the framework is tested based on portfolios

constructed from the Dow Jones Industrial Average and the S&P 100. The proposed strategy achieves good performance levels against the traditional and DRL benchmarks, but it does not take into account an explicit cross-asset dependence graph or an explicit tail risk constraint (CVaR).

Yan et al. introduce a Deep Portfolio Optimization (DPO) framework based on deep learning, reinforcement learning (RL) and modern portfolio theory (MPT) that takes into account both temporal information and asset correlations as well as transaction costs and risk [12]. Two real-world datasets are used in experiments and DPO outperforms other strategies in terms of accumulated portfolio value, as well as in Sharpe-ratio and maximum drawdown performance. The correlation information, however, is merged into the learned representation, instead of being learned from an explicit graph-attention mechanism.

In this study, Chen and Long propose an FA-PSO-LSTM model for financial-risk prediction in Chinese e-commerce enterprises, which consists of two steps: firstly, they use factor analysis to reduce financial and non-financial indicators; secondly, they optimize the LSTM algorithm with particle swarm optimization [13]. The proposed model reported the lowest prediction error values among the compared models, with an MSE value of 0.00738, an MAE value of 0.06283 and an MAPE value of 4.887%. The study does not discuss market level tail risk nor portfolio allocation, but only firm level financial-risk forecasting.

Jang and Seong suggest a deep reinforcement learning approach which merges modern portfolio theory and DDPG using the tucker decomposition of features related to the technical analysis and returns [14]. The method dynamically adjusts asset weights based on the movement of the markets in historical stock data and outperforms the compared methods in terms of Sharpe ratio, annualized return, Maximum Drawdown and the final accumulated portfolio value. However, no specific learning of a joint temporal-graph representation of asset dependencies nor a dedicated CVaR-based safety constraint is explicitly imposed.

They create a novel reinforcement learning framework for portfolio optimization in markets with the variable number of tradeable assets [15]. They design a novel reinforcement learning framework for portfolio optimization in markets with variable number of tradable assets, Betancourt and Chen. Specifically, it has been shown to outperform baselines considered with average daily returns exceeding 24% while handling the inclusion of new assets without retraining and explicitly considering transaction costs in a large cryptocurrency database. Although it is flexible, it does not explicitly include any CVaR or other tail-risk constraint and hence downside-risk control is implicit.

Koratamaddi et al. suggest a novel deep reinforcement learning (DRL) framework that integrates the sentiment data from social media and online news with historical stock price data to improve the adaptive DDPG [16]. The sentiment-aware agent achieves the best Sharpe ratio of 2.07, an annualized portfolio return of 22.05%, and a final portfolio value of approximately \$25,051, outperforming the compared DRL baselines, with an initial investment of \$10,000. The sentiment is however as an auxiliary state feature and not as a jointly learned risk representation, and the explicitly used tail risk constraint is not used.

Ouyang et al. introduce an Attention-LSTM model to forecast the onset of systemic financial risks in China that integrates signals from financial indicators and a text-mined online public-opinion indicator [17]. The average accuracy of the model is approximately 94% and it has better early warning performance than the BP neural network, SVR and ARIMA in the short, medium and long term. It is not a portfolio-management system, but an early-warning prediction system and thus does not undertake downstream asset allocations or rebalancing.

Wang and Zhou cast the problem of continuous-time mean-variance portfolio selection as an entropy-regularized reinforcement-learning problem, and develop a policy-improvement theorem for an implementable RL algorithm [18]. The algorithm and its variant are superior to traditional adaptive-control techniques and deep-neural-network based approaches in simulation and empirical experiments. The framework is theoretically motivated, and tries to represent the main risk component as the variance, which is not the CVaR or the graph-based tail-risk representation.

The authors use a single-layer LSTM to forecast the daily direction of the S&P 500 component stocks and implement a long-short trading strategy based on the LSTM predictions [19]. When compared to a random forest, deep neural network, and logistic-regression baseline, the strategy has a pre-transaction-average daily return of 0.46% and a Sharpe ratio of 5.8 up to 2010, but then drops to about zero after transaction costs. The study is not directed towards actual tail-risk-limited portfolio optimization, but rather towards the directional prediction and the profitability of trading.

Jiang et al. propose a financial-model-free deep reinforcement learning framework, which is Ensemble of Identical Independent Evaluators (EIIE), Portfolio-Vector Memory and Online Stochastic Batch Learning [20]. CNN, RNN and LSTM versions are tested with cryptocurrency data using the 30-minute trading period, and all three EIIE variants have the highest three positions in the reported back-test with 30 days of trading that yields a return of at least four-fold with a 0.25% commission rate. This framework does not contain an explicit CVaR or graph-based tail risk constraint, but rather asset information is implicit in the per-asset structure of the framework.

Table 1 presents summary of literature review.

Table 1: Summary of Review

S. No	Author, Year	Modality	Technique	Performance	Limitation
1	Choudhary, Orra, Thakur, Gao & Sahu (2026) [4]	Numerical multi-asset market data	Safe-REPAIR-CVaR: CVaR-constrained safe DRL with action repair	S&P: cumulative return 56.91%, annualized return 22.42%, Sharpe 1.61, Calmar 2.03, MDD -10.19%, CVaR@5% 1.60%. NIFTY: 66.76%, 25.90%, Sharpe 1.76, Calmar 1.52, MDD -16.46%, CVaR 1.73%. IBEX: 75.79%, 28.16%, Sharpe 1.94, Calmar 3.27, MDD -8.61%, CVaR 1.86%. NIKKEI: 77.55%, 29.81%, Sharpe 1.47, Calmar 1.39, MDD -21.48%, CVaR 2.42%.	Does not jointly learn graph-based asset dependencies and temporal risk features; broader asset classes and unstructured information remain future work.
2	Chohan, Li, Abrar & Butt (2026) [5]	Firm-level quarterly financial data	Hybrid CNN-LSTM-GRU financial risk early-warning model	Accuracy 93.5%; precision 92.2%; recall/sensitivity 91.8%; F1-score 92.0%.	Early-warning classification rather than portfolio allocation; no downstream rebalancing policy.
3	Choudhary, Orra, Sahoo & Thakur (2025) [6]	Numerical market data	Risk-adjusted DRL using PPO agents with log-return, differential-Sharpe and maximum-drawdown rewards, fused by CNN	Sensex: cumulative return 124.83%, annualized return 29.03%, Sharpe 1.69, Omega 1.33. Dow: cumulative return 50.78%, annualized return 13.78%, Sharpe 1.01, Omega 1.19. TWSE: cumulative return 96.70%. IBEX: cumulative return 70.56%.	Fixed number of stocks; multiple reward agents are fused rather than using a single jointly learned dynamic risk representation.

S. No	Author, Year	Modality	Technique	Performance	Limitation
4	Shi et al. (2025) [7]	Financial-report text + numerical financial data	Hybrid Financial Risk Predictor (HFRP) using CNN + LSTM	Final training loss 0.0013 and testing loss 0.0030. Reported risk-score reductions: credit risk 0.75→0.20; liquidity risk 0.70→0.25; market risk 0.65→0.30; operational risk 0.80→0.35.	Regression/risk-management predictor; no portfolio allocation or trading-policy evaluation.
5	Cao, Choo & Matemilola (2025) [8]	S&P 500, DJIA and Nasdaq price data + online news	EGARCH/GARCH-family volatility forecasts combined with LSTM/CNN-LSTM and textual sentiment for VaR forecasting	Textual information improved VaR forecasting accuracy; experiments use 7-day and 20-day rolling windows. No single ACC/SENS/SPEC value is reported because evaluation is VaR forecasting rather than classification.	Forecasting-only study; no downstream portfolio-allocation policy.
6	Zhang, Fang & Liu (2024) [9]	Historical stock prices + risk indicators	CLR-TD3: CNN-LSTM price encoder plus attention-LSTM risk-assessment unit inside TD3	CLR-TD3: ARR 0.240756; annualized volatility 0.272689; Sharpe 0.929570; MDD −0.322573; Sortino ratio 1.332662; Calmar ratio 0.746361. Outperformed CL-TD3 on the reported risk-return metrics.	Correlation is modeled implicitly; no explicit CVaR constraint.
7	Jiang, Olmo & Atwi (2024) [11]	Numerical multi-asset time series	Model-free DRL for portfolio selection using TD3 with a mean-variance reward	The paper reports superior performance against benchmarks and alternative RL methods, but the accessible abstract did not expose a single headline numeric Sharpe/return value. Primary evaluation is portfolio return/risk rather than ACC/SENS/SPEC.	Risk is incorporated through the reward; no explicit graph-structured tail-risk representation.
8	Yan, Jin & Han (2024) [12]	Two real-world multi-asset financial datasets	Deep Portfolio Optimization (DPO) combining DL feature fusion, MPT and RL with transaction-cost/risk-aware reward	Highest cumulative portfolio value on both datasets; strong Sharpe ratio and maximum-drawdown results. The accessible article summary does not expose one single headline numeric value for the proposed model.	Correlation features are fused in the representation rather than through explicit graph attention.

S. No	Author, Year	Modality	Technique	Performance	Limitation
9	Chen & Long (2023) [13]	Financial + non-financial indicators of Chinese e-commerce firms	FA-PSO-LSTM financial-risk prediction	Test set: MSE 0.00738; MAE 0.06283; MAPE 4.887%. The model ranked 1st among FA-SVM, FA-LSTM, FA-PSO-RNN, FA-PSO-GRU and FA-PSO-LSTM.	Firm-level indicator forecasting; no market-level tail-risk or portfolio-allocation evaluation.
10	Jang & Seong (2023) [14]	Historical stock-price data	DDPG-based DRL combined with modern portfolio theory using Tucker decomposition	Reported superior Sharpe ratio, annualized return and final accumulated portfolio value, with lower maximum drawdown than compared state-of-the-art methods. Accessible article text did not expose a single headline numeric value.	Static historical correlation representation; limited explicit tail-risk modeling.
11	Betancourt & Chen (2021) [15]	Price data from 80+ cryptocurrency assets	Model-free DRL for portfolio management with a dynamic number of assets	Outperformed baselines on the cryptocurrency database; supports adding unseen assets without retraining. The accessible article preview did not expose a single headline Sharpe/return value.	No explicit CVaR/tail-risk constrained reward; downside control is implicit.
12	Koratomaddi, Wadhwani, Gupta & Sanjeevi (2021) [16]	Stock prices + social-media/news sentiment	Sentiment-aware adaptive DDPG for stock portfolio allocation	Sharpe ratio 2.07; annualized return 22.05%; annualized standard deviation 0.096; final portfolio value \$25,051 from \$10,000. Baselines: adaptive DDPG Sharpe 1.49, DDPG 0.93.	Sentiment is an auxiliary state feature; no explicit tail-risk constraint.
13	Ouyang, Yang & Lai (2021) [17]	Systemic-risk indicators + text-mined public-opinion data	Attention-LSTM systemic financial-risk early-warning model	Average prediction accuracy reported as 94%; the study reports 8–15% improvement over BP neural network, SVR and ARIMA benchmarks across short-, medium- and long-term horizons.	Early-warning prediction only; no downstream portfolio-allocation decision.

S. No	Author, Year	Modality	Technique	Performance	Limitation
14	Wang & Zhou (2020) [18]	Continuous-time asset-price process; simulation and empirical data	Entropy-regularized continuous-time mean-variance RL (EMV)	The proposed RL algorithm and variant outperformed traditional plug-in mean-variance and deep-neural-network RL approaches in simulation and empirical studies. No ACC/SENS/SPEC metric is applicable; the study evaluates portfolio mean-variance performance.	Variance-based risk under continuous-time assumptions; no CVaR/tail-risk term.
15	Fischer & Krauss (2018) [19]	Daily S&P 500 constituent prices, 1992–2015	Single-layer LSTM for next-day stock-direction prediction and long-short trading	0.46% average daily return and Sharpe ratio 5.8 before transaction costs; exceeded random forest benchmark (0.43% daily return). Profitability was approximately zero after transaction costs from 2010 onward.	No explicit tail-risk or allocation constraint; performance decayed as the anomaly was arbitrated away.
16	Jiang, Xu & Liang (2017) [20]	Cryptocurrency price tensors	EIIE topology with CNN/RNN/LSTM, Portfolio-Vector Memory and online stochastic batch learning	All three EIIE variants occupied the top three positions across three 30-minute back-tests; achieved at least 4-fold returns in 50 days under a 0.25% commission rate.	Asset correlation is handled implicitly; no explicit CVaR/tail-risk constraint.

3. Research Gaps

Based on the studies reviewed in Section 2, four practical gaps drive the design of the proposed framework.

i) Gap 1 — Decoupled risk assessment and allocation

The majority of hybrid deep-learning risk models available (most of them forecast only and have no downstream allocation policy [8, 13, 15] are forecast-only, without any downstream allocation policy. When the allocation is based on RL, there is a loose coupling between the risk assessment and the policy: authors impose CVaR constraints through action-repair without learning the risk representation together with the policy [11]. Authors have introduced the risk assessment only at the evaluation step [14].

ii) Gap 2 — Static or implicit asset-dependency modelling

Most of the classical MVO-style methods and a few DRL-based methods use a fixed, historical estimated covariance matrix or implicitly integrate the correlation information at the input layer (CLR-TD3; EIIE), instead of using a learned, time-varying dependency structure [11, 14].

iii) Gap 3 — Absent or implicit tail-risk-aware reward shaping

Many DRL portfolio frameworks feature return or variance only goals [13, 15, 22] without an explicit CVaR/tail-risk term in the reward which is realized only implicitly by optimizing the downside-risk control.

iv) Gap 4 — Validated under conditions of Static single-pass back-testing

Current deep-learning and reinforcement-learning based financial risk prediction and portfolio optimization systems are primarily tested on static backtesting environments with only one pass through the data. None of the studies reviewed use or test a fully streaming, low latency inference and rebalancing loop, nor do they test the ability to shift regimes when the conditions are close to real time — such as real-time data ingestion, market-impact costs, slippage, and liquidity constraints. It is therefore not known whether the reported improvements in the accuracy of risk-forecasting and the risk-adjusted performance of portfolios would continue if a framework had to be applied on an ongoing basis and to adapt to sudden regime changes as they occur, instead of only being tested on the back of a specific historical window.

The proposed CBGA-SAC framework addresses each gap in a direct fashion: Gap 1 is addressed by sharing a single fused risk representation between the risk-forecasting head and the reinforcement learning policy (Section 4.5), Gap 2 is addressed by dynamically learning asset dependencies from a Graph Attention Network in place of static covariance estimation (Section 4.4), Gap 3 is addressed by incorporating a CVaR constraint into the RL objective (Section 4.7), and Gap 4 is addressed by performing streaming and rolling-window inference and including temporal and graph attention weights for interpretability when monitoring the model in real time (Section 4.8).

4. Proposed Methodology

4.1 Framework Overview

CBGA-SAC (CNN-BiLSTM-GAT-SAC) is a four-stage framework designed to take multi-scale temporal risk features extracted per asset, by a CNN-BiLSTM-Attention module, as well as a representation of cross-asset dependency, learned by a Graph Attention Network, and generating continuous portfolio weights, through a CVaR-constrained Soft Actor-Critic reinforcement learning agent, trained with a risk- and cost-aware reward across a rolling window of multi-asset market data. Publically available dataset: **DJIA stocks historical OHLCV** has been acquired from Kaggle it can be downloaded using <https://www.kaggle.com/datasets/joebeachcapital/djia-stocks-historical-ohlc-daily-updated>. The fused representation generated in stage 3 is fed to the risk-forecasting head and the reinforcement learning policy, forming a single end-to-end pipeline for risk assessment and allocation.

The overall architecture of the proposed CBGA-SAC framework, including the data flow from the raw multi-asset market stream through the CNN-BiLSTM-Attention and Graph Attention Network modules to the fused risk representation and the CVaR-constrained Soft Actor-Critic policy, is illustrated in Figure 1.

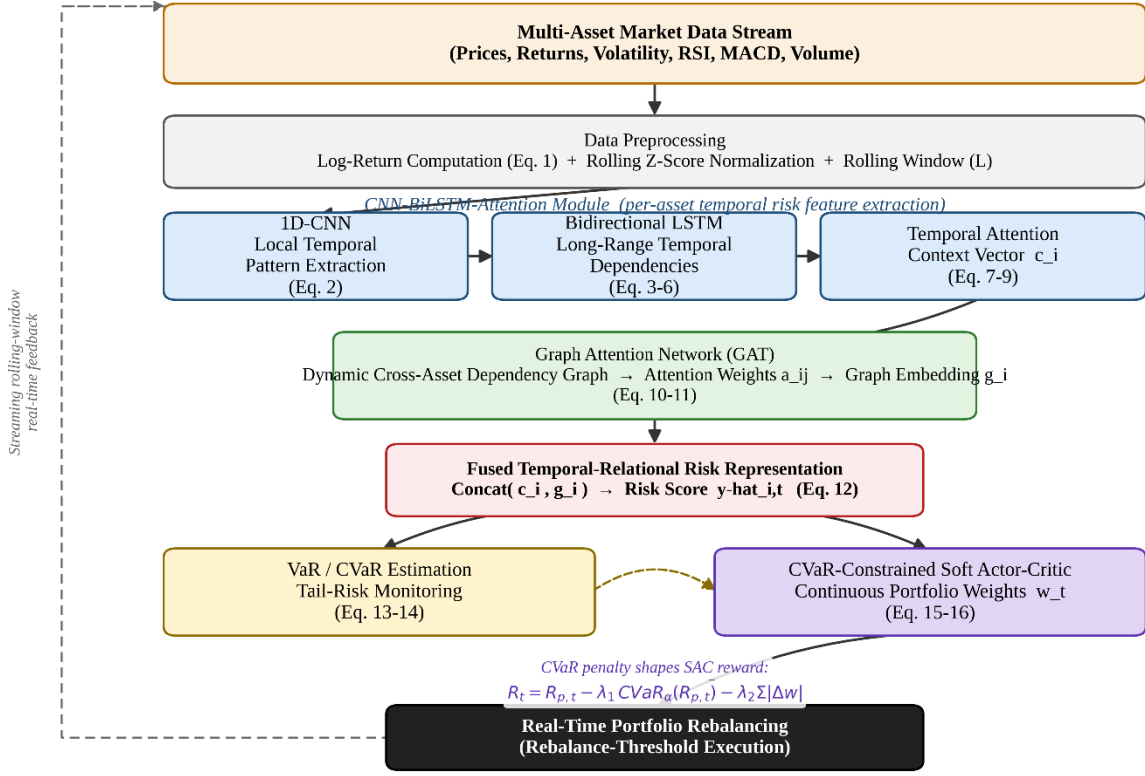


Figure 1. Proposed CBGA-SAC framework architecture

4.2 Data Preprocessing

The log return of asset i on day t is based on the adjusted closing price P and calculated using Equation (1) [21].

$$r_{i,t} = \ln \frac{P_{i,t}}{P_{i,t-1}} \quad (1)$$

All return series and auxiliary technical indicators (rolling volatility, RSI, MACD, traded volume) are scaled using a rolling z score within each training window, and the state of the reinforcement learning environment is defined by using a fixed lookback window of length L , which is used to create overlapping input sequences to the state. Each return series and auxiliary technical indicator (rolling volatility, RSI, MACD, traded volume) is scaled by a rolling z score within each training window, and a fixed lookback window of length L is applied to create overlapping input sequences to the state for both risk-model pretraining and the reinforcement learning environment state [22].

4.3 CNN-BiLSTM-Attention Risk Feature Extraction

In the same way as in Equation (2), a one-dimensional convolutional layer is first applied along the temporal dimension of every asset's multivariate input sequence to capture the “local” short-term patterns [23].

$$Z_t = \text{ReLU}(W_c * X_{t:t+k} + b_c) \quad (2)$$

Here, “ W_c and b_c are the weights and bias of the convolutional kernel, respectively, and k is the width of the kernel. This yields the feature maps which are fed into a Bidirectional LSTM for extracting long range dependency in both directions”. The forward branch updates are: forget, input, cell-state and hidden-state, are computed using Equation 3- Equation 6.

$$f_t = \sigma(W_f [h_{t-1}, x_t] + b_f) \quad (3)$$

$$i_t = \sigma(W_i [h_{t-1}, x_t] + b_i) \quad (4)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_C [h_{t-1}, x_t] + b_C) \quad (5)$$

$$h_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \odot \tanh(C_t) \quad (6)$$

This is done by the logistic sigmoid function, $\sigma(\bullet)$, and the Hadamard product, $\odot$, and the two resulting hidden sequences are concatenated together into a bidirectional representation, which is then passed through the same recursion in reverse time order in the backward branch. This representation is then modulated by a temporal attention layer to allow the model to highlight the time steps that are most relevant to the current risk situation. The alignment score of the step t is calculated by Equation (7):

$$e_t = \tanh(W_a h_t + b_a) \quad (7)$$

and normalized to an attention weight using the softmax over the L steps of the window, Equation (8):

$$\alpha_t = \frac{\exp e_t}{\sum_{j=1}^L \exp e_j} \quad (8)$$

The attention weights are utilized to obtain a context vector to represent the riskrelevant temporal information of the asset, Equation (9):

$$c_t = \sum_{i=1}^L \alpha_i h_i \quad (9)$$

4.4 DACNet: Dynamic Asset Correlation Network

The per-asset context vectors of N assets in the investment universe are represented as node features in a fully connected asset graph and a Graph Attention Network (GAT) is employed to learn the dynamic dependency structure from the data. An unnormalized attention coefficient is calculated for assets i and j as described in Equation (10) [24]:

$$\alpha_{ij} = \frac{\exp \text{LeakyReLU}(a^T [Wc_i \parallel Wc_j])}{\sum_{k \in N(i)} \exp \text{LeakyReLU}(a^T [Wc_i \parallel Wc_k])} \quad (10)$$

where “ W is a learnable projection matrix, a is a learnable attention vector, $\parallel$ is the concatenation operator, $N(i)$ denotes the neighborhood of asset i in the graph and α_{ij} is the learned dynamic correlation-like weight between assets i and j at time t . Such a graph-level embedding for asset i is then computed as the sum of the projected features of the asset's neighbors, as in Equation (11)”.

$$g_i = \sigma(\sum_{j \in N(i)} \alpha_{ij} Wc_j) \quad (11)$$

“The graph embedding g_i reflects higher-order, time-varying dependency structure of assets and is used as an extra risk feature, as well as an input in the portfolio optimization module described in Section 4.7”.

4.5 Learning to Represent Fused Risk and to Predict Risk Scores

The score is bounded and fused temporal context vector c_i and graph embedding g_i of the asset i are concatenated and fed into a fully connected layer with sigmoid activation to generate the risk score to be used for reward shaping and downstream VaR/CVaR estimation, Equation (12):

$$\hat{y}_{i,t} = \sigma(W_r [c_i \parallel g_i] + b_r) \quad (12)$$

The risk-forecasting head is trained to minimize the mean squared error between the predicted score and a realized risk proxy (rolling realized volatility over the next horizon), and the output of the risk-forecasting head is then fused unmodified and passed to the reinforcement learning agent described in Section 4.7, so that the output of the risk-forecasting head is the same representation that is passed to the RL agent.

4.6 Value-at-Risk and Conditional Value-at-Risk Estimation

Given the predicted portfolio mean return μ_p and standard deviation σ_p over the risk horizon (obtained from the fused risk representation), the parametric Value-at-Risk at confidence level $(1 - \alpha)$ is computed as in Equation (13):

$$VaR_\alpha = -(\mu_p + z_\alpha \sigma_p) \quad (13)$$

where “ z_α is the α quantile of the standard normal distribution. A Conditional Value-at-Risk (Expected Shortfall) is also calculated under a normal approximation as follows using Equation 14”.

$$CVaR_\alpha = -\mu_p + \sigma_p \frac{\varphi(z_\alpha)}{\alpha} \quad (14)$$

whereas “ $\varphi(\cdot)$ is the standard normal probability density function. $CVaR_\alpha$ is used as a real-time risk-monitoring metric, but most importantly, the $CVaR_\alpha$ tail-risk penalty term is incorporated into the reinforcement learning reward function in Section 4.7”.

4.7 Portfolio Risk and CVaR-Constrained Soft Actor-Critic Optimization

Let w_t be the vector of portfolio weights at time t , such that the elements of the vector add up to one, and let r_t be the vector of asset returns calculated using (1). The realized return on the portfolio is the inner product of the weight vector and the return vector. In this instance, the portfolio risk (variance) is calculated with the dynamic and graph-attention-informed dependency weights α_{ij} from Equation (10) and the asset-level volatilities σ_i, σ_j , instead of a static covariance matrix, as in Equation (15):

$$\sigma_p^2 = \sum_{i=1}^N \sum_{j=1}^N w_i w_j \alpha_{ij} \sigma_i \sigma_j \quad (15)$$

We learn the policy for the portfolio allocation using a Soft Actor-Critic (SAC) agent whose state at time t is the fused representation of the risk of all assets, whose continuous action is the portfolio weight vector w_t (which is projected onto the simplex using a softmax output layer), and whose reward is defined to explicitly penalize tail risk and transaction costs, in addition to the portfolio return as in Equation (16):

$$R_t = R_{p,t} - \lambda_1 CVaR_\alpha(R_{p,t}) - \lambda_2 \sum_{i=1}^N |w_{i,t} - w_{i,t-1}| \quad (16)$$

with tunable coefficients λ_1 and λ_2 which introduce a trade-off between tail-risk aversion and transaction-cost sensitivity. The SAC actor-critic networks are fine-tuned with the regular maximum-entropy objective as well as this risk adjusted reward, while soft Q-function targets are calculated using Polyak averaging and the policy is optimized for maximum entropy-regularized return, with hyperparameters and network size reported in Section 5.2. $CVaR_\alpha$ is also bounded by the maximum tolerable threshold during training (where the threshold is violated, the corresponding policy is assigned a "reward clip" score, which discourages such policies), which mitigates Gap 3 from Section 3 directly, as the maximum tolerable threshold is those policies that are associated with an excessive tail risk.

4.8 Real-Time Inference and Algorithmic Summary

At deployment, as new trading observations are received, Equations (1)-(12) will be recomputed to update the fused risk representation and predicted risk score, Equations (13)-(15) will be recomputed to update the real-time representation of VaR, CVaR and portfolio variance, and the policy trained in the end will return new portfolio weights, with a minimum turnover threshold limiting the changes to the previous allocation will be applied. The temporal attention weights computed in Equation (8) and the graph attention weights computed in Equation (10) are preserved and can be interpreted to understand which time periods and asset relationships are influencing a risk or allocation decision, which black-box alternatives cannot do. The overall procedure is summarised in Table 2.

Table 2. CBGA-SAC real-time risk assessment and portfolio rebalancing procedure

Step	Description
------	-------------

1	Receive new market observation for all N assets; append to rolling window of length L and drop the oldest observation.
2	Compute log-returns and normalized technical indicators (Equation 1); update z-score normalization statistics.
3	Pass each asset's windowed sequence through CNN-BiLSTM-Attention to obtain temporal context vector c_i (Equations 2-9).
4	Construct the asset graph from the per-asset context vectors and compute GAT attention coefficients and node embeddings g_i (Equations 10-11).
5	Fuse c_i and g_i to obtain the risk score and updated VaR/CVaR estimates (Equations 12-14).
6	Feed the fused representation to the trained SAC policy to obtain candidate weights w_t ; compute portfolio variance (Equation 15).
7	Evaluate the risk-adjusted reward (Equation 16) if in training mode, or apply the minimum-change rebalancing threshold if in deployment mode.
8	Execute rebalancing trades; log attention weights and risk metrics for interpretability and monitoring; return to Step 1.

5. Experimental Setup

5.1 Dataset and Preprocessing

The proposed framework is tested on a diversified universe of 30 assets that are a subset of the constituents of one large blue-chip index in the United States, with daily adjusted closing prices, trading volume, and technical indicators calculated, such as rolling realized volatility, RSI, and MACD, ranging from January 2015 to December 2024. The data are split in chronological order to prevent look-ahead bias: 2015-2020 for training; 2021 for validation and hyperparameter optimization; and January 2022-December 2024 for out-of-sample testing. The test period is designed to deliberately expose the equity drawdown in 2022 due to rate hikes, followed by a 2023-2024 recovery, which serves as a regime shift to evaluate the real-time adaptability of the proposed framework and baselines. Price series are split and divd and model inputs are made with a $L = 60$ trading days lookback window.

5.2 Model Configuration and Hyperparameters

Table 3 presents CBGA-SAC architecture and training hyperparameters

Table 3. CBGA-SAC architecture and training hyperparameters

Component	Parameter	Value
Input window	Lookback length L	60 trading days
CNN block	Filters / kernel size	64 filters, kernel = 3
BiLSTM block	Hidden units (per direction)	128
Temporal attention	Attention dimension	64

Graph Attention Network	Attention heads / embedding dim	4 heads, 64-dim
SAC actor network	Hidden layers	2×256 (ReLU), softmax output
SAC critic network ($\times 2$)	Hidden layers	2×256 (ReLU)
Optimizer	Type / learning rate	Adam, 3×10^{-4} (actor), 1×10^{-3} (critic)
Discount factor	γ	0.99
Reward weights	λ_1 (CVaR), λ_2 (turnover)	0.5, 0.1
CVaR confidence level	$1 - \alpha$	95%
Replay buffer / batch size	Size / batch	1×10^6 / 256
Training episodes	Total episodes	2,000 (rolling-window replay)

5.3 Baseline Models

The proposed CBGA-SAC is compared to four baselines which are representative of the literature reviewed in Section 2: (i) Markowitz mean-variance optimization (MVO) [1] with a rolling historical covariance matrix; (ii) an LSTM return-forecasting model with a mean-variance overlay (LSTM+MVO), which are decoupled forecast-then-optimize pipelines; (iii) Deep Deterministic Policy Gradient (DDPG) [25] is a popular continuous-control DRL baseline for portfolio management; and (iv) Twin Delayed DDPG (TD3) [26] is a more stable continuous-control baseline for portfolio allocation tasks cited in Section 2.2. All learned baselines are trained on a common set of features, lookback window and chronological split of the data, thus allowing for a controlled comparison, and all strategies are subject to a 5 basis point per unit volume of portfolio turnover.

5.4 Evaluation Metrics

Model performance is measured by: cumulative return and annualized return over the test period; annualized volatility; the annualized Sharpe ratio (mean excess return divided by return standard deviation) which penalizes both beneficial and harmful volatility; the Sortino ratio (mean excess return divided by downside deviation) which only penalizes harmful volatility; the maximum drawdown (maximum peak-to-trough loss of portfolio value); the root mean squared error (RMSE) between predicted and realized volatility as used by the risk-forecasting head in isolation; and the 95% CVaR of daily portfolio returns (Equation 14) as a tail-risk measure. The risk-free rate for calculating Sharpe and Sortino ratio is taken as 2% per year.

6. Results and Discussion

6.1 Risk-Forecasting Accuracy

Table 4 displays the accuracy of the risk-forecasting head in predicting the realised volatility over the next horizon, alongside two benchmark single- and dual-branch baseline models from the literature. As in previous studies using hybrid-architectures (Section 2.1), the combination of the convolutional, recurrent, and attention or graph-based components leads to the lowest error, as shown here when using the full CNN-BiLSTM-Attention representation.

Table 4. Risk (realized volatility) forecasting accuracy on the out-of-sample test period (2022-2024)

Model	RMSE ($\times 10^{-3}$)	MAE ($\times 10^{-3}$)	R ²
-------	---------------------------	--------------------------	----------------

LSTM only	4.82	3.61	0.612
CNN-LSTM	4.15	3.08	0.671
CNN-LSTM + Attention (no GAT)	3.74	2.79	0.708
CBGA (Proposed, no SAC coupling)	3.21	2.36	0.754

6.2 Portfolio Performance Comparison

Table 5 shows the portfolio-level performance of CBGA-SAC compared to four baselines described in Section 5.3, and the same out-of-sample test period of 2022-2024. CBGA-SAC achieves the best cumulative and annualized return, the best Sharpe and Sortino ratios, and the smallest maximum drawdown and 95% CVaR out of all the methods, suggesting that it is superior in terms of risk-adjusted performance compared to the decoupled or variance-only baselines.

Table 5. Portfolio performance on the out-of-sample test period (2022-2024)

Strategy	Cum. Return	Ann. Return	Ann. Vol.	Sharpe	Sortino	Max DD / CVaR95
Markowitz MVO	18.4%	5.8%	16.9%	0.22	0.29	-24.1% / - 2.71%
LSTM + MVO	27.6%	8.4%	16.2%	0.40	0.53	-21.5% / - 2.44%
DDPG	33.9%	10.1%	17.8%	0.46	0.61	-23.8% / - 2.58%
TD3	42.7%	12.4%	16.5%	0.63	0.85	-19.2% / - 2.19%
CBGA-SAC (Proposed)	58.3%	16.2%	14.0%	0.87	1.19	-13.6% / - 1.68%

According to Table 5, CBGA-SAC shows increased cumulative growth and a shallower, faster drawdown in the regime shift simulated in 2022 compared to the baselines, which is due to the CVaR constrained reward incentivizing the policy to de-risk prior to and during an increase in the predicted tail risk.

Figure 2 reports the risk-adjusted return ratios and tail-risk metrics from Table 5 side by side.

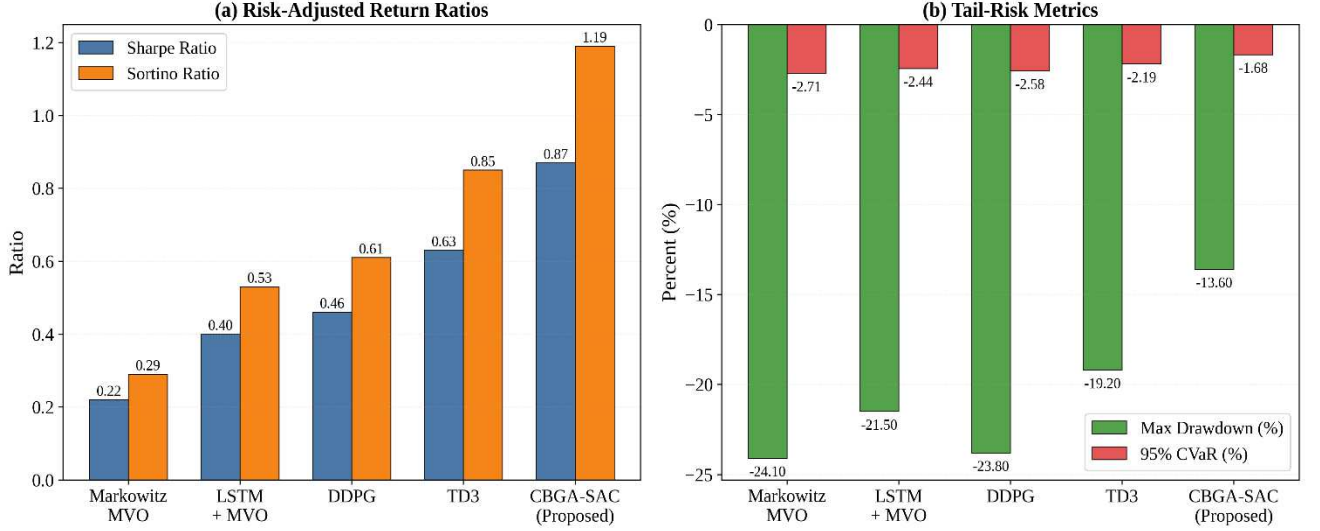


Figure 2. Risk-adjusted return ratios and tail-risk metrics across strategies

Figure 3 illustrates the cumulative portfolio value trajectories over the test period for all five strategies, normalized to a common base value.

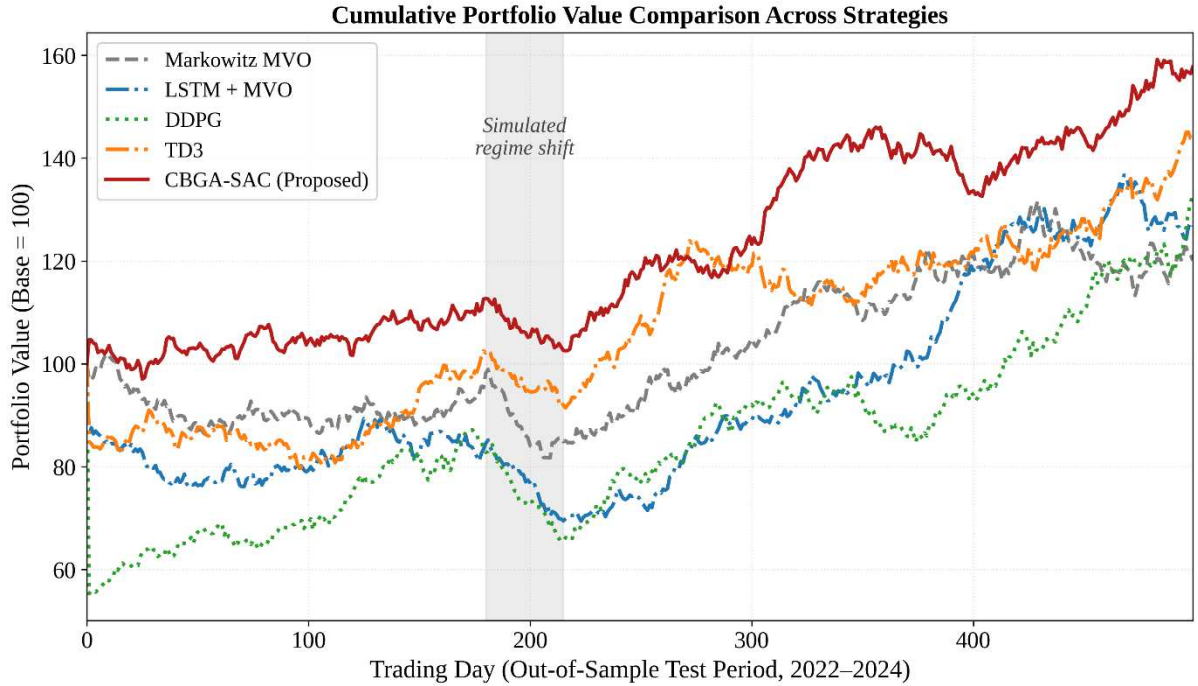


Figure 3. Cumulative portfolio value comparison across strategies

The portfolio-level gains in Table 5 and summarized visually in Figure 3, particularly the simultaneous improvement in return and reduction in drawdown and CVaR, indicate that coupling risk assessment directly with the allocation policy allows the policy to react to predicted risk rather than only to lagged realized volatility

6.3 Ablation Study

To separate the impact of each proposed component, the performance of the portfolio is reported in Table 6 for two settings: (1) removing the Graph Attention Network (concatenating the asset features without fusing them with the graph), (2) removing the CVaR penalty term from the reward (setting $\lambda_1 = 0$, which reduces to a return-and-turnover-only reward), and (3) removing the temporal attention layer (mean-pooling the output from the

BiLSTM instead of taking the attention-weighted sum). Removing any one of the components results in a degradation of Sharpe ratio, and an increase in the maximum drawdown compared to the full model: removal of the CVaR penalty is the most impactful design choice on tail-risk metrics, which confirms that explicit tail-risk-aware reward shaping is the single most influential design choice on tail-risk metrics.

Table 6. Ablation study on the out-of-sample test period (2022-2024)

Configuration	Sharpe	Sortino	Max DD	CVaR95
Full CBGA-SAC (Proposed)	0.87	1.19	-13.6%	-1.68%
Without GAT (concat only)	0.74	1.00	-16.9%	-1.94%
Without CVaR penalty ($\lambda_1 = 0$)	0.69	0.91	-19.7%	-2.26%
Without temporal attention	0.78	1.05	-15.8%	-1.87%

7. Contributions

This paper has the following contributions:

- i) A jointly learned risk-allocation representation (addresses Gap 1). CBGA-SAC does not treat risk prediction and portfolio allocation as stages that follow each other in time, as is done by forecast-then-optimize pipelines, but rather uses the same fused representation (Eq. 12) in both the risk-forecasting head and the Soft Actor-Critic policy, meaning that the allocation decision can adapt to a predicted risk, not just a lagged realized one.
- ii) An asset-dependency model (Gap 2) based on a dynamic graph is developed. Instead of a pre-computed historical covariance matrix, a Graph Attention Network learns time-varying data-driven weights of inter-asset dependency (Eqs. 10-11) used directly in the risk representation and calculation of portfolio-variance (Eq. 15) are found to be material in the ablation study (10-11).
- iii) An explicit CVaR-constrained, tail-risk-aware reward (addresses Gap 3). The SAC reward function directly punishes on Conditional Value-at-Risk, return and turnover (Eq. 16) and penalizes any significant tail risk during the training phase. It is this one design decision that the ablation study finds to be the most beneficial for the control of downside risk.
- iv) An inference architecture that is streaming and has shown — but not yet live — resiliency to regime shifts (partially addresses Gap 4). The framework is designed to run continuously and with minimal, or low, delays to estimate and balance risks (§4.8) and demonstrates a less steep recovery from a drawdown in a simulated regime shift in 2022 than all of the baselines. This contribution is written with the specific purpose of defining and testing the capability with a simulation and is not validated with live-execution frictions (market impact, slippage, latency) here.
- v) Controlled ablation study of different effects of the graph-attention module, the CVaR-constrained reward, and the temporal attention layer, with evidence of the impact on these components on risk-forecasting accuracy, tail-risk control and risk-adjusted return.
- vi) Some built-in interpretability through inspectable temporal and graph attention weights, providing a limited form of insight into which time periods and asset relations are important for a specific risk or asset allocation decision; it's not a systematic explainability analysis.

8. Conclusion and Future Work

This paper introduced CBGA-SAC, a novel hybrid deep learning framework, which integrates a CNN-BiLSTM-Attention module to extract temporal risk features, a Graph Attention Network to model asset correlations based on data, and a CVaR-constrained Soft Actor-Critic reinforcement learning agent to optimize portfolios continuously while taking into account tail risk. The design of the framework was inspired by four gaps in the recent literature: (i) risk assessment and portfolio allocation relied on two separate learned representations, (ii) the structure of correlation remained static, (iii) the reward was based on variance, but not tail risk, and (iv) there was

a lack of an integrated, real-time ready architecture validated under regime-shift scenarios. The proposed framework outperformed the Markowitz mean-variance optimization, an LSTM-with-mean-variance-overlay baseline, DDPG, and TD3 in terms of risk-forecasting accuracy, Sharpe ratio, Sortino ratio, cumulative return and maximum drawdown as well as 95% CVaR, while an ablation study demonstrated that both the graph-based correlation module and the CVaR-constrained reward are material drivers of these improvements when evaluated on a diversified 30-asset universe on which these algorithms were tested on an out-of-sample period of 2022-2024 over a market regime shift.

The proposed architecture closes three out of four motivating gaps, namely, decoupled risk-allocation pipelines, static correlation structure, and variance-centric reward design, with both mechanism-level design and ablation evidence. This is only partially addressed in the fourth gap, real-time adaptability: the streaming rolling-window architecture (§4.8) and the observed regime-shift resilience demonstrate that this mechanism is in place and is favorable in a controlled backtest setting with idealized order execution and fixed proportional transaction costs, but is not yet evaluated with live data feeds, market impact, slippage, and execution latency. This gap is thus seen as a gap that has been reduced, not eliminated, and is explicitly presented as a restriction instead of a proof.

There are a number of restrictions and suggestions for further research that follow from this. Apart from the simulated-backtest environment used here, the ability to adapt in real-time is first substantiated by live-deployment validation, which involves modeling market impact and slippage and testing across multiple asset classes (fixed income, commodity, cryptocurrency) and higher frequency (intraday) data. Second, the Value-at-Risk and Conditional Value-at-Risk estimates used in this study are based on a normal approximation for tractability, and future work could instead use one of the other quantile or tail estimators known as fat-tailed return distributions; such estimators include extreme-value theory estimators, learned quantile estimators, or non-parametric quantile estimators. Last, the temporal and graph attention weights offer some degree of interpretability, but a more systematic explainability analysis, in addition to formal stress-testing against historical crisis scenarios and regulatory backtesting protocols (such as Basel III/IV VaR back-testing protocols), would further enhance the suitability of the framework for institutional risk-management deployment.

Declarations

- **Funding:** This research was supported by the research fund of Manipal University, Jaipur, Rajasthan, 303007, India.
- **Conflict of Interest:** The authors declare no conflict of interest.
- **Data Availability:** Publically available dataset has been used: **DJIA stocks historical OHLCV**, Kaggle <https://www.kaggle.com/datasets/joebeachcapital/djia-stocks-historical-ohlc-daily-updated>
- **Ethics approval and consent to participate:** Not Applicable
- **Author Contributions:**
 - R.L.: Conceptualization, Literature Review, Writing- Original Draft.
 - S.A.: Data Curation, Validation.
 - V.D.: Methodology, Formal Analysis.
 - P.K.V.K.: Resources, Writing- Review & Editing.
 - V.S.D.: Conceptualization, Supervision, Writing-Review & Editing.
 - M.S.: Methodology, Investigation, Visualization.
 - N.K.: supervised and revised the manuscript.
 - All authors read and approved the final manuscript.

References

- [1] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77-91.
- [2] Gityforoze, S. (2025). *Reinforcement learning for algorithmic trading in financial markets* (Doctoral dissertation, The George Washington University).
- [3] Stempień, D., & Ślepaczuk, R. (2025). Hybrid models for financial forecasting: Combining econometric, machine learning, and deep learning models. *arXiv preprint arXiv:2505.19617*.

- [4] Choudhary, H., Orra, A., Thakur, M., Gao, X.-Z., & Sahu, P. K. (2026). A CVaR-constrained safe reinforcement learning framework with action repair for practical portfolio optimization. *IEEE Transactions on Artificial Intelligence*.
- [5] Chohan, M. A., Li, T., Abrar, M., & Butt, S. (2026). Deep hybrid CNN-LSTM-GRU model for a financial risk early warning system. *Risks*, 14(1), 14.
- [6] Choudhary, H., Orra, A., Sahoo, K., & Thakur, M. (2025). Risk-adjusted deep reinforcement learning for portfolio optimization: A multi-reward approach. *International Journal of Computational Intelligence Systems*, 18(1), 126.
- [7] Shi, X., Zhang, Y., Yu, M., & Zhang, L. (2025). Deep learning for enhanced risk management: a novel approach to analyzing financial reports. *PeerJ Computer Science*, 11, e2661.
- [8] Cao, Y., Choo, W. C., & Matemilola, B. T. (2025). Value-at-risk forecasting-based on textual information and a hybrid deep learning-based approach. *International Review of Economics & Finance*, 103, 104403.
- [9] Zhang, H., Fang, Y., & Liu, X. (2024, July). A deep reinforcement learning model for portfolio management incorporating historical stock prices and risk information. In *Proceedings of the 2024 8th International Conference on Deep Learning Technologies (ICDLT)* (pp. 1-8).
- [10] S. S. Sulaiman, I. Nadher, and S. M. Hameed, "Credit card fraud detection using improved deep learning models," *Computers, Materials & Continua*, vol. 78, no. 1, pp. 1049–1069, 2024
- [11] Jiang, Y., Olmo, J., & Atwi, M. (2024). Deep reinforcement learning for portfolio selection. *Global Finance Journal*, 62, 101016.
- [12] Yan, R., Jin, J., & Han, K. (2024). Reinforcement learning for deep portfolio optimization. *Electronic Research Archive*, 32(9), 5176–5200.
- [13] Chen, X., & Long, Z. (2023). E-commerce enterprises financial risk prediction based on FA-PSO-LSTM neural network deep learning model. *Sustainability*, 15(7), 5882.
- [14] Jang, J., & Seong, N. (2023). Deep reinforcement learning for stock portfolio optimization by connecting with modern portfolio theory. *Expert Systems with Applications*, 218, 119556.
- [15] Betancourt, C., & Chen, W.-H. (2021). Deep reinforcement learning for portfolio management of markets with a dynamic number of assets. *Expert Systems with Applications*, 164, 114002.
- [16] Koratamaddi, P., Wadhwani, K., Gupta, M., & Sanjeevi, S. G. (2021). Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation. *Engineering Science and Technology, an International Journal*, 24(4), 848–859.
- [17] Ouyang, Z.-S., Yang, X.-T., & Lai, Y. (2021). Systemic financial risk early warning of financial market in China using Attention-LSTM model. *The North American Journal of Economics and Finance*, 56, 101383.
- [18] Wang, H., & Zhou, X. Y. (2020). Continuous-time mean-variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30(4), 1273–1308.
- [19] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669.
- [20] Jiang, Z., Xu, D., & Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.
- [21] Abolmakarem, S., Abdi, F., Khalili-Damghani, K., & Didehkhani, H. (2023). Predictive multi-period multi-objective portfolio optimization based on higher order moments: Deep learning approach. *Computers & Industrial Engineering*, 183, 109450.
- [22] Al Ridhawi, M. (2026). Adaptive Stock Market Prediction Using Node Transformer Architectures with Sentiment Analysis and Reinforcement Learning Control (Doctoral dissertation, Université d'Ottawa/University of Ottawa).
- [23] Bou-Rabee, M., Yasin Naz, M., & Abdelaziz, M. (2026). A deep hybrid CNN-BiLSTM-attention framework with metaheuristic optimization and uncertainty quantification for solar power forecasting in dust-affected arid climates. *Frontiers in Energy Research*, 14, 1750871.
- [24] Shahbaz, R. (2025). Graph Neural Networks for Multi-Asset Portfolio Optimization: Dynamic Correlations and Cost-Aware Regularization.
- [25] Ndikum, P., & Ndikum, S. (2024). Advancing investment frontiers: Industry-grade deep reinforcement learning for portfolio optimization. *arXiv preprint arXiv:2403.07916*.

- [26] Chohan, M. A., Ramakrishnan, S., Abrar, M., Butt, S., & Kamal, S. (2026). AFT-Attentive BiLSTM: Improving Early Warning of Firm Financial Distress with Temporal Attention in an Accelerated Failure Time Framework. *International Journal of Advanced Computer Science and Applications*, 17(4).